SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0001 $01.00/0

# EXTREMAL VALUES OF THE INTERVAL NUMBER OF A GRAPH*

JERROLD R. GRIGGS† AND DOUGLAS B. WEST‡

**Abstract.** The interval number $i(G)$ of a simple graph $G$ is the smallest number $t$ such that to each vertex in $G$ there can be assigned a collection of at most $t$ finite closed intervals on the real line so that there is an edge between vertices $v$ and $w$ in $G$ if and only if some interval for $v$ intersects some interval for $w$. The well known interval graphs are precisely those graphs $G$ with $i(G) \leqq 1$. We prove here that for any graph $G$ with maximum degree $d$, $i(G) \leqq \lceil \frac{1}{2}(d+1) \rceil$. This bound is attained by every regular graph of degree $d$ with no triangles, so is best possible. The degree bound is applied to show that $i(G) \leqq \lceil \frac{1}{3}n \rceil$ for graphs on $n$ vertices and $i(G) \leqq \lfloor \sqrt{e} \rfloor$ for graphs with $e$ edges.

**1. Introduction to interval numbers.** We begin by discussing earlier work on interval graphs and boxicity in order to motivate the definition of interval numbers. A simple bound on the interval number given the numbers of edges and of vertices of a graph is presented along with results on the interval number of some basic graphs. In the next section we prove our main result, which gives the best-possible upper bound on the interval number of a graph given its maximum degree. This is applied to obtain an upper bound on the interval number given only the number of vertices. We conclude by listing several interesting open problems.

Interval graphs are simple undirected graphs $G$ with the property that there exists a collection of finite closed intervals on the real line such that an interval $[a_v, b_v]$ is assigned to each vertex $v$ in $G$ and such that the intervals assigned to two vertices $v$ and $w$ in $G$ intersect each other if and only if they are joined by an edge in $G$. Interval graphs have been studied extensively and can be nicely characterized [1], [3], [4], [5]. They have important applications to various problems of scheduling, allocation, and sequencing.

It is natural to try to extend this idea of representing graphs by intersections of intervals to all graphs. For even some simple graphs, such as the $n$-cycles $C_n (n > 3)$, are not interval graphs. One approach, taken by Roberts [7], [8], is to go to higher dimensional intervals: define the *boxicity* of a graph $G$ to be the smallest integer $t$ such that $G$ can be represented by the intersections of $t$-dimensional "boxes" which have their edges parallel to the coordinate axes. That is, to each vertex $v$ is assigned an ordered collection of $t$ finite closed intervals

$$([a_{v,1}, b_{v,1}], [a_{v,2}, b_{v,2}], \cdots, [a_{v,t}, b_{v,t}]),$$

and two vertices $v$ and $w$ are joined by an edge in $G$ if and only if $[a_{v,i}, b_{v,i}]$ intersects $[a_{w,i}, b_{w,i}]$ for all $i$. In these terms, interval graphs are precisely the graphs with boxicity at most 1.

Here we present a different approach to extending interval representations to all graphs. We expect that this approach will be useful in dealing with certain scheduling and allocation problems, such as traffic light assignments [9] and radio frequency assignments [2]. Rather than going to higher-dimensional intervals, we allow each vertex to be represented by a collection of several intervals. Define the *interval number*

* Received by the editors January 22, 1979.

† Department of Mathematics, California Institute of Technology, Pasadena, California 91125.

‡ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

of a graph $G$, denoted $i(G)$, to be the smallest integer $t \geqq 0$ such that there can be assigned to each vertex in $G$ a collection of at most $t$ finite closed intervals so that there is an edge in $G$ between vertices $v$ and $w$ if and only if some interval for $v$ intersects some interval for $w$. This definition is due to R. McGuigan [6].

$i(G)$ exists for any graph $G$: An interval representation of $G$ is obtained by taking a pair of overlapping intervals, one labelled $v$ and the other $w$, for each edge $\{v, w\}$ in $G$. These pairs of intervals are to be separated from each other. Of course, this construction will not achieve the value $i(G)$ in general.

Interval graphs are precisely those graphs with $i(G) \leqq 1$. Only for graphs with no edges does $i(G) = 0$. Complete graphs are interval graphs, so $i(K_n) = 1$. To represent $K_n$, just stack up $n$ intervals, one per vertex, so that their mutual intersection is nonempty. The cycles $C_n$, $n > 3$, are not interval graphs. $i(C_n) = 2$ as the representation in Fig. 1 shows for $C_4$.

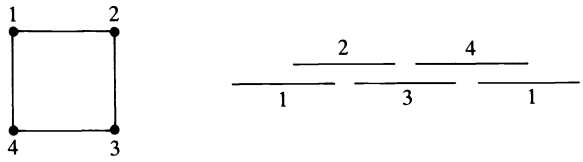



FIG. 1

Using the Lekkerkerker–Boland forbidden subgraph characterization of interval graphs [5], it is straightforward to show that for trees $T$, $i(T) = 1$ if and only if $T$ contains no induced subgraphs of the form shown in Fig. 2. Otherwise, $i(T) = 2$. See [10] for details.

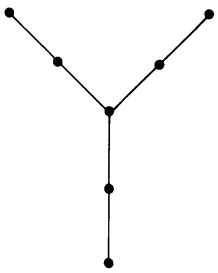


FIG. 2

Next we consider arbitrary graphs which contain no triangles. We present a simple proof of a useful lower bound on $i(G)$ given the number of edges and the number of vertices in $G$. Here $\lceil x \rceil$ denotes the least integer no smaller than $x$, and $\lfloor x \rfloor$ denotes the integer part of $x$.

THEOREM 1. *Let $G$ be a simple graph on $n$ vertices and $e > 0$ edges which contains no $K_3$. Then $i(G) \geqq \lceil (e+1)/n \rceil$.*

*Proof.* As $e > 0$, we have $i(G) \geqq 1$. Suppose we are given an interval representation $I$ for $G$ which attains the bound $i(G)$, i.e., uses no more than $i(G)$ intervals per vertex. As $G$ contains no $K_3$, no three intervals in $I$ may share a point. For each edge $\{v, w\}$ in $G$, there must be a stretch of points on the real line where a $v$-interval overlaps a $w$-interval. At the right end of such a stretch, one of the two intervals must end. It follows that there must be at least $e+1$ intervals in $I$. Thus some vertex must be represented by at least $\lceil (e+1)/n \rceil$ intervals, so that $i(G) \geqq \lceil (e+1)/n \rceil$. □

We thus derive a lower bound on the interval numbers of complete bipartite graphs:

COROLLARY.

$$i(K_{m,n}) \geqq \left\lceil \frac{mn+1}{m+n} \right\rceil.$$

We have constructed interval representations which show that this lower bound is actually the correct value of $i(K_{m,n})$ in various special cases. But Trotter and Harary [10], working independently of us, have proposed the same definition of $i(G)$ and have come up with a construction for all $m$ and $n$ of interval representations of $K_{m,n}$ using at most $\lceil (mn+1)/(m+n) \rceil$ intervals per vertex to show that this lower bound is always the actual value of $i(K_{m,n})$.

We have some results on $i(G)$ for complete $p$-partite graphs with $p > 2$ which we are currently trying to improve and plan to discuss in another paper.

**2. The degree bound.** Now we come to the main result of this paper, the best-possible upper bound on $i(G)$ for graphs $G$ which have degree at most $d$ at each vertex. Note that an *upper* bound is what is interesting; the lower bound is just 1 for all $d > 0$, because $i(K_{d+1}) = 1$. The construction following the definition of $i(G)$ established an upper bound of $d$ on $i(G)$. Here we shall lower this bound to $\lceil \frac{1}{2}(d+1) \rceil$ and show that it is actually attained by some graphs. The interval representation to attain this upper bound is simple to construct for a given graph, and it has some other nice properties. We apply this result in the next two sections to obtain upper bounds on $i(G)$ when $G$ has a given number of vertices or edges. For convenience let $d(v)$ denote the degree of vertex $v$.

THEOREM 2. *If $G$ is a graph with $d = \max_v d(v) > 0$, then $i(G) \leqq \lceil \frac{1}{2}(d+1) \rceil$.*

*Proof.* For any graph $G$ with $d$ as above we must give an interval representation for $G$ using at most $d$ intervals per vertex in order to prove the theorem. We do this by induction on the number $n$ of vertices of $G$ using this stronger induction hypothesis:

(*) For any graph $G$ on $n$ vertices and any vertex $v$ in $G$ there is an interval representation of $G$ in which the leftmost interval is a $v$-interval and in which, for each vertex $w$ in $G$, there are at most $\lceil \frac{1}{2}(d(w)+1) \rceil$ $w$-intervals.

An interval $[a, b]$ is *leftmost* (respectively, *rightmost*) if for any other interval $[c, d]$ in the representation, $a < c$ ($b > d$).

Hypothesis (*) holds trivially for $n = 1$. So assume that $G$ has $n > 1$ vertices and that (*) holds for all graphs on fewer than $n$ vertices. Let $v$ be any vertex in $G$. We now construct an interval representation satisfying (*).

Suppose first that there is a circuit $C_1$ passing through $v$ in $G$. By *circuit* we mean a path which begins and ends at $v$ without repeating edges and without passing through any vertex twice. Say $C_1 = v, w_1, w_2, \cdots, w_k, v$ lists the vertices in $C_1$ in order, where $k \geqq 2$. Figure 3 shows an interval representation of the edges in $C_1$. Now remove these edges from $G$ (but not the vertices). Suppose there remains another circuit $C_2$ through $v$. Then represent each edge in $C_2$ using the same idea as for $C_1$, except the $v$-interval on the right for $C_1$ is used as the $v$-interval on the left for representing $C_2$. Continue this

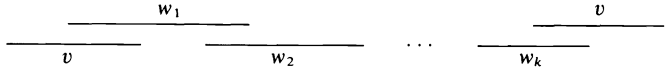



FIG. 3

procedure of representing and deleting the edges in circuits through $v$ until no more circuits pass through $v$.

In the case that there is no such circuit $C_1$ passing through $v$, just put down a single $v$-interval. So, in general, if we remove $m$ circuits through $v$, the number of $v$-intervals used is precisely $m+1$, and the left-most and rightmost intervals are $v$-intervals. In removing these circuits, the degree of each vertex $w \neq v$ is reduced by twice the number of $w$-intervals used in the representation.

If $v$ now belongs to no edges, apply (*) by induction to the rest of $G$ to obtain a representation which satisfies the degree bound in (*) at each vertex. Otherwise, suppose that there are $p>0$ vertices adjacent to $v$ in $G$, which we call $u_1, u_2, \cdots, u_p$. Since no circuits pass through $v$ now, the vertices $u_i$ lie in distinct components if $v$ is deleted from $G$. Let $G_i$ be the component containing $u_i$. By induction there is an interval representation $I_i$ of $G_i$ in which $u_i$ is leftmost and in which the number of intervals for each vertex is bounded according to (*).

Put $I_1$ to the right of the intervals used to represent the circuits of $G$ so that the leftmost $u_1$-interval in $I_1$ overlaps the rightmost $v$-interval. This represents the edge $\{v, u_1\}$ and all edges in $G_1$. For $p>1$, add $[p/2]$ *additional* $v$-intervals to the right of the intervals used thus far. Reverse the order of representations $I_2, I_4, I_6, \cdots$ so that there are intervals for $u_2, u_4, u_6, \cdots$ which are rightmost in their representations. Then insert the $I_i$ in the representation of $G$ so that $I_2$ is to the left of the leftmost new $v$-interval, $I_3$ is to its right, $I_4$ is to left of the second new $v$-interval, and so on. The extreme $u_i$-interval in $I_i$ should overlap the $v$-interval. Figure 4 shows the construction. To complete the construction, represent any remaining edges, by induction on (*), with intervals to the right of all the other intervals.

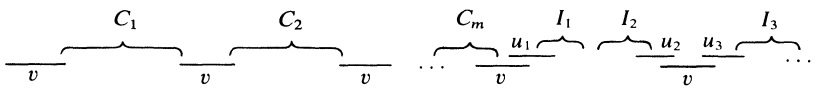



FIG. 4

We have now represented all the edges of $G$, and no others. A $v$-interval is leftmost. By counting the intervals used, it follows that not more than $\lceil \frac{1}{2}(d(w)+1) \rceil$ intervals are used for any vertex $w$ in $G$. Thus (*) is satisfied, and the theorem is proven. □

That this bound is best possible follows from this result:

COROLLARY. *For any regular graph $G$ of degree $d$ containing no $K_3$,*

$$i(G) = \lceil \tfrac{1}{2}(d+1) \rceil .$$

*Proof.* Suppose $G$ is a regular graph of degree $d$ containing no $K_3$, and let $n$ be the number of vertices of $G$. $G$ *has exactly* $\frac{1}{2}nd$ edges, so by Theorem 1,

$$i(G) \geqq \lceil (\tfrac{1}{2}nd+1)/n \rceil = \lceil \tfrac{1}{2}(d+1) \rceil ,$$

and this is just the upper bound on $i(G)$ in Theorem 2. □

Two important examples of such graphs are $K_{d,d}$ and $Q_d$, the $d$-dimensional cube.

A strong property of the representation in the proof of Theorem 2 that may be useful in some applications is that it has *depth two*: no three intervals overlap on the real line.

**3. The vertex bound.** Another extremal problem of interest is this: Among all graphs $G$ on $n$ vertices, how large can $i(G)$ be? Since $d \leqq n-1$, it follows from Theorem

2 that $i(G) \leqq \lceil \frac{1}{2}n \rceil$. As an application of Theorem 2 we present here a nice interval construction to improve this bound on $i(G)$ to $\lceil \frac{1}{3}n \rceil$.

THEOREM 3. *If $G$ has $n$ vertices, then $i(G) \leqq \lceil \frac{1}{3}n \rceil$.*

*Proof.* The proof is by induction on $n$. It is certainly true for $n \leqq 3$. Now suppose $G$ has $n > 3$ vertices. We prove $i(G) \leqq \lceil \frac{1}{3}n \rceil$ in two cases, depending on whether or not $G$ has a triangle.

First suppose $G$ contains some triangle, on vertices $T = \{u, v, w\}$. The intervals shown in Fig. 5 represent the edges in $T$ and have the additional property that for any subset of $T$ there is a stretch of the line in which intervals for precisely this subset

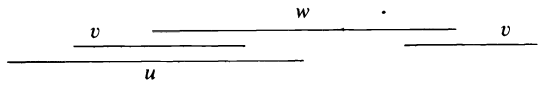



FIG. 5

overlap. Hence we can represent all edges between $T$ and $G - T$ (the vertices outside $T$) using at most one interval per vertex in $G - T$: For example, if a vertex $x$ in $G - T$ neighbors $u$ and $w$, put a small $x$-interval inside the interval where the $u$- and $w$-intervals overlap, but no $v$-interval does. At most two intervals are used for each vertex in $T$ in taking care of all edges involving $T$. Now, by induction, represent all edges between vertices in $G$-$T$, using at most $\lceil \frac{1}{3}n \rceil - 1$ intervals per vertex. Place these intervals away from those involving $T$ to complete the construction.

It remains to consider graphs $G$ with no triangle. By Theorem 2, $i(G) \leqq \lceil \frac{1}{3}n \rceil$ holds provided that $\lceil \frac{1}{2}(d+1) \rceil \leqq \lceil \frac{1}{3}n \rceil$, or, equivalently, $d \leqq 2\lceil \frac{1}{3}n \rceil - 1$. Thus in the remaining case it suffices to assume that $G$ contains some vertex $v$ of degree at least $\frac{2}{3}n$. Let $W$ be the set of vertices adjacent to $v$. There are no edges in $W$ because $G$ has no triangles. Let $X$ be the set of vertices outside $W \cup \{v\}$. To represent all edges incident on $W$, take a long interval for each of the vertices in $X \cup \{v\}$, no two intersecting, and for each edge $\{w, y\}$, with $w \in W$ and $y \in X \cup \{v\}$, put a small $w$-interval inside the $y$-interval. (See Fig. 6.) At most $\frac{1}{3}n$ intervals are used for each $w \in W$ because $|X \cup \{v\}| \leqq \frac{1}{3}n$. The only remaining edges involve pairs of vertices in $X$ and can be represented, by induction, using at most $\lceil \frac{1}{3}|X| \rceil$ intervals per vertex in $X$. This represents $G$ with at most $\lceil \frac{1}{3}n \rceil$ intervals per vertex. □

Trotter and Harary [10] independently discovered the same $\lceil \frac{1}{3}n \rceil$ bound on $i(G)$. The construction given here is simpler. How good is this bound? The balanced complete bipartite graphs, $K_{\lfloor n/2 \rfloor, \lceil n/2 \rceil}$, show that $i(G)$ can get at least as large as $\lceil \frac{1}{4}(n+1) \rceil$. This agrees with $\lceil \frac{1}{3}n \rceil$ for $n < 7$. At $n = 7$ it is not difficult to prove that $i(G)$ can be at most 2, so the $\lceil \frac{1}{3}n \rceil$-bound is not always best possible. It is natural to conjecture that these graphs $K_{\lfloor n/2 \rfloor, \lceil n/2 \rceil}$ are extremal among all graphs $G$ on $n$ vertices, just as they were for graphs of maximum degree $d$. That is, the best possible upper bound on $i(G)$ should be $\lceil \frac{1}{4}(n+1) \rceil$. One of us (Griggs) has recently succeeded in showing this, but owing to the length and complexity of the proof, it will appear elsewhere [13].

**4. The edge bound.** How large can the interval number of a graph with $e$ edges get? Theorem 2 can again be applied to give an upper bound.

THEOREM 4. *If $G$ has $e$ edges, then $i(G) \leqq \lceil \sqrt{e} \rceil$.*

*Proof.* The theorem holds trivially if $e \leqq 1$, so assume that $G$ has $e > 1$ edges and that the theorem holds for all graphs with fewer than $e$ edges. Let $k = \lceil \sqrt{e} \rceil$. If $d < 2k$, then $i(G) \leqq k$ by Theorem 2. So assume that $d \geqq 2k$ and let $v$ be a vertex of degree $d$. Represent all edges containing $v$ by $a$ long $v$-interval overlapped by a small $w$-interval

for each neighbor $w$ of $v$ (see the $v$-interval in Fig. 6). This requires at most one interval per vertex in $G$. The remaining $e-d<(k-1)^2$ edges in $G$ can be represented using at most $k-1$ intervals per vertex by induction. □

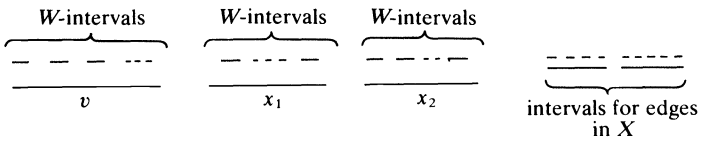



FIG. 6

This argument can be refined to obtain the slightly stronger result that $i(G) \leqq \lfloor \sqrt{e} \rfloor$. It can be shown that $i(G) \leqq 2$ for $e = 9$, so this upper bound $\lfloor \sqrt{e} \rfloor$ is not best possible. The graphs $K_{2m,2m}$, $m = 1, 2, 3, \cdots$, show that $i(G)$ can get at least as large as $1 + \lfloor \frac{1}{2}\sqrt{e} \rfloor$. We conjecture that this is also an upper bound, which would be best possible.

**5. Areas requiring further study.** Applications will motivate the study of other problems related to interval numbers. We propose the following:

1. Give a forbidden subgraph characterization of the graphs with interval number at most $k$, where $k \geqq 2$.
2. Interval numbers minimize the maximum number of intervals used for any vertex in representing $G$. One could instead seek to minimize the *total* number of intervals required in a representation.
3. Representations could be restricted to being of depth at most $r$ by not allowing any $r+1$ intervals to share a point. What can be said about the "depth $r$ interval number"?
4. Rather than intervals, one can use circular arcs to represent vertices and ask for a circular interval number $i_c(G)$. This means that we allow a single interval to go to $+\infty$ and come back from $-\infty$, so that $(-\infty, a] \cup [b, \infty)$, with $a < b$, counts as a single circular interval. $i_c(C_n) = 1 < i(C_n) = 2$ for $n > 3$. Graphs with $i_c(G) \leqq 1$ are known as circular-arc graphs [11], [12]. What is the behavior of $i_c(G)$? It should be similar to $i(G)$ since for all graphs, $i(G) \geqq i_c(G) \geqq i(G) - 1$.

**Acknowledgments.** We are indebted to Fred Roberts for introducing us to interval graphs and for bringing the work of Trotter and Harary to our attention; to Robert McGuigan for proposing the study of interval numbers; and to Daniel J. Kleitman for making some valuable suggestions. This work originated at the NSF-CBMS Regional Conference in Graph Theory at Colby College, June, 1977.

## REFERENCES

[1] D. R. FULKERSON AND O. A. GROSS, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835–855.

[2] E. N. GILBERT, *Mobile radio frequency assignments*, unpublished technical memorandum, Bell Telephone Laboratories, 1972.

[3] P. C. GILMORE AND A. J. HOFFMAN, *A characterization of comparability graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.

[4] G. HAJOS, *Uber eine Art von Graphen*, Internat. Math. Nachr., 47 (1957), p. 65.

[5] C. B. LEKKERKERKER AND J. CH. BOLAND, *Representation of a finite graph by a set of intervals on the real line*, Fund. Math., 51 (1962), pp. 45–64.

[6] R. MCGUIGAN, presentation at NSF-CBMS Conference at Colby College, June, 1977.

[7] F. S. ROBERTS, *On the boxicity and cubicity of a graph*, Recent Progress in Combinatorics, W. T. Tutte, ed., Academic Press, New York, 1969, pp. 301–310.

[8] F. S. ROBERTS, *Graph Theory and its Applications to Problems of Society*, lectures presented at NSF-CBMS Conference at Colby College, June 1977, Society for Industrial and Applied Mathematics, Philadelphia, 1978.

[9] K. E. STOFFERS, *Scheduling of traffic lights—A new approach*, Transportation Research, 2 (1968), pp. 199–234.

[10] W. T. TROTTER, JR., AND F. HARARY, *On double and multiple interval graphs*, J. Graph Th., to appear.

[11] A. C. TUCKER, *Characterizing circular-arc graphs*, Bull. Amer. Math. Soc., 75 (1970), pp. 1257–1260.

[12] ———, *Matrix characterizations of circular-arc graphs*, Pacific J. Math., 39 (1971), pp. 535–545.

[13] J. R. GRIGGS, *Extremal values of the interval number of a graph, II*, Discrete Math., to appear.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0002 $01.00/0

# ON THE STRUCTURE OF $t$-DESIGNS*

R. L. GRAHAM†, S.-Y. R. LI‡ AND W.-C. W. LI¶

**Abstract.** It is possible to view the combinatorial structures known as (integral) $t$-designs as $\mathbb{Z}$-modules in a natural way. In this note we introduce a polynomial associated to each such $\mathbb{Z}$-module. Using this association, we quickly derive explicit bases for the important class of submodules which correspond to the so-called null-designs.

**Introduction.** Among the most fundamental (and least understood) types of combinatorial configurations are the $t$-designs [2], [5], [6]. These can be defined as follows. Let $v$, $k$, $t$ and $\lambda$ be positive integers satisfying $t \leqq k \leqq v$. A *$t$-design* $S_\lambda(t, k, v)$ is a collection $\mathcal{B}$ of $k$-subsets $B$ (called *blocks*) of a $v$-set $V$ with the property that every $t$-subset of $V$ occurs as a subset of exactly $\lambda$ blocks $B \in \mathcal{B}$. (It is not required that blocks be distinct.) It follows from this definition that for any $i \leqq t$, the number of blocks of a $t$-design which contain a fixed $i$-subset $I$ of $V$ is exactly

$$\lambda\binom{v-i}{t-i}\Big/\binom{k-i}{t-i} \tag{1}$$

independent of $I$, which implies, in particular, that a *necessary* condition for existence of an $S_\lambda(t, k, v)$ is that the expressions in (1) are integers for $1 \leqq i \leqq t$. In fact, Wilson [6] has shown that for any $t \leqq k \leqq v$, this is also a *sufficient* condition for the existence of an $S_\lambda(t, k, v)$ provided only that $\lambda \geqq \lambda_0(t, k, v)$ is sufficiently large.

Let $M$ be the free $\mathbb{Z}$-module generated by all the subsets of $V$; the elements of $M$ are all sums $\bar{c} = \sum_{X \subseteq V} c_X X$, where $c_X \in \mathbb{Z}$. In this terminology, a $t$-design is just an element $\bar{c} = \sum_{|Y|=k} c_Y Y$ with all $c_Y \geqq 0$ such that for all $t$-subsets $X$,

$$\sum_{Y \supseteq X} c_Y = \lambda.$$

A submodule of $M$ of particular interest is the module $N_k$ defined by

$$N_k = \left\{\bar{c} \in M: \sum_{X \subseteq V} c_X = 0 \text{ and when } |X| \neq k,\ c_X = 0\right\}.$$

The elements of $N_k$ are usually called *null-designs* since they result when the (module) difference of two $t$-designs is formed. In principle, if the structure of null-designs can be sufficiently well understood, then light will be shed on $t$-designs since any $S_\lambda(t, k, v)$ differs from a given $S'_\lambda(t, k, v)$ by a null-design.

In [2], Graver and Jurkat obtain a generating system for the module $N_k$ from a special construction which they call a "$(t, k)$-pod". In this note we recast the concept of null-designs in terms of polynomials. From this formulation we reproduce the above generators in a much simpler way. In fact we show that there are basically only five kinds of linear dependence among these generators, and thereby produce in Theorem 4 an

* Received by the editors February 12, 1979, and in revised form April 19, 1979.

† Bell Laboratories, Murray Hill, New Jersey 07974.

‡ Department of Mathematics, University of Chicago, Chicago, Illinois 60637. The work of this author was supported in part by the National Science Foundation under Grant MCS77-03533.

¶ Department of Mathematics, University of Chicago, Chicago, Illinois 60837. The work of this author was undertaken during a visit at the Institute for Advanced Study, Princeton, NJ.

explicit basis for $N_k$ described in terms of simple polynomials. In proving this theorem we use the fact that the $\binom{v}{t}$ by $\binom{v}{k}$ "inclusion" matrix $H_{v,k,t} = (h_{X,Y})$, with $|X| = t$, $|Y| = k$ and

$$h_{X,Y} = \begin{cases} 1 & \text{if } X \subseteq Y, \\ 0 & \text{otherwise} \end{cases}$$

has full rank. Although this is well known (see [3] for a short proof), we give a new proof of it by exhibiting an explicit (generalized) inverse for the inclusion matrix.

**The polynomial ring.** Let $\mathbb{Z}[x_1, \cdots, x_v]$ denote the polynomial ring with $v$ variables over $\mathbb{Z}$. For $\sigma \in S_v$, the group of permutations on $\{1, 2, \cdots, v\}$, and $f$ in $\mathbb{Z}[x_1, \cdots, x_v]$, define the polynomial $f^\sigma \in \mathbb{Z}[x_1, \cdots, x_v]$ by

$$f^\sigma(x_1, \cdots, x_v) = f(x_{\sigma(1)}, \cdots, x_{\sigma(v)}).$$

We shall say that $f$ is $x^2$-*free* if every monomial appearing in it is squarefree. With each multiset $\mathcal{B}$ of subsets of $V = \{1, 2, \cdots, v\}$ we can associate a polynomial $f_{\mathcal{B}}$ by

$$f_{\mathcal{B}} = \sum_{B \in \mathcal{B}} \prod_{i \in B} x_i. \tag{2}$$

If $\mathcal{B}$ forms a $t$-design $S_\lambda(t, k, v)$, then the polynomial $f_{\mathcal{B}}$ is a positive integral linear combination of squarefree monomials of degree $k$ with the property (by (1)) that for all $\sigma \in S_v$,

$$f^\sigma(x_1, \cdots, x_t, 1, \cdots, 1) = \lambda \sum_{i=0}^{t} \frac{\binom{v-t}{t-i}}{\binom{k-i}{t-i}} a_i^\sigma(x_1, \cdots, x_t), \tag{3}$$

where $a_i^\sigma(x_1, \cdots, x_t)$ denotes the $i$th symmetric function of the $x_j$'s. Thus, a null-design, being the difference of two $t$-designs, is a homogeneous $x^2$-free polynomial $g$ of degree $k$ satisfying

$$g^\sigma(x_1, \cdots, x_t, x, \cdots, x) \equiv 0 \tag{4}$$

for all $\sigma \in S_v$. These $g$ form a $\mathbb{Z}$-module $N$ (in the obvious way) which is free since it is contained in the free $\mathbb{Z}$-module of rank $\binom{v}{k}$ generated (over $\mathbb{Z}$) by all the monomials $\{\prod_{i \in I} x_i : I \subseteq V, |I| = k\}$.

**Generators for null-designs.**

THEOREM 1. (*Graver–Jurkat*). *The module $N$ of null-designs is generated over $\mathbb{Z}$ by the collection $\{\phi^\sigma : \sigma \in S_v\}$, where*

$$\phi(x_1, \cdots, x_v) = (x_1 - x_2)(x_3 - x_4) \cdots (x_{2t+1} - x_{2t+2}) x_{2t+3} \cdots x_{k+t+1}.$$

*This collection is void when $v \leqq k + t$ or $k \leqq t$.*

*Proof.* Suppose $f$ is a nonzero null-design. Without loss of generality, we may assume that the monomial $x_1 \cdots x_k$ occurs in $f$ with a nonzero coefficient $c$. Thus

$$f(\overbrace{1, \cdots, 1}^{k}, 0, \cdots, 0) = c \neq 0.$$

It follows from (4) that $k < v - t$ and $v - k < v - t$, i.e.,

$$v \geqq k + t + 1 \quad \text{and} \quad k \geqq t + 1.$$

In particular, this proves the theorem for the case that $v \leqq k + t$ or $k \leqq t$. □

We now show that $f$ is generated by the $\phi^\sigma$, $\sigma \in S_v$. The proof is by induction on $t$ and, for a fixed $t$, by induction on $v$. Because $f$ is $x^2$-free, we can write

$$f(x_1, \cdots, x_v) = g(x_1, \cdots, x_{v-1}) + h(x_1, \cdots, x_{v-1})x_v.$$

For any permutation $\tau \in S_{v-1}$ and any values of $x_1, \cdots, x_{t-1}, x_v$ and $x$, we have

$$\begin{aligned} 0 &= f^\tau(x_1, \cdots, x_{t-1}, x, \cdots, x, x_v) \\ &= g^\tau(x_1, \cdots, x_{t-1}, x, \cdots, x) + h^\tau(x_1, \cdots, x_{t-1}, x, \cdots, x)x_v, \end{aligned}$$

and therefore

$$h^\tau(x_1, \cdots, x_{t-1}, x, \cdots, x) = 0.$$

This shows that $h$ is a null-design with parameters $(t-1, k-1, v-1)$ when $t \geqq 1$. Let

$$\theta(x_1, \cdots, x_{v-1}) = (x_1 - x_2) \cdots (x_{2t-1} - x_{2t})x_{2t+1} \cdots x_{k+t-1}.$$

When $t \geqq 1$, we may assume by the induction hypothesis on $t$ that $h$ is an integral linear combination of $\theta^\tau$, $\tau \in S_{v-1}$. Of course this is also true when $t = 0$. Thus we can write

$$h(x_1, \cdots, x_{v-1}) = \sum_{\tau \in S_{v-1}} c_\tau \theta^\tau$$

with $c_\tau \in \mathbb{Z}$. Since $v > k + t$, there exists, for each $\tau$, a variable $x(\tau) \neq x_v$ not appearing in $\theta^\tau$. Therefore $\theta^\tau(x_v - x(\tau))$ is equal to $\phi^{\sigma(\tau)}$ for some $\sigma(\tau) \in S_v$. Now the polynomial

$$\begin{aligned} f - \sum_{\tau \in S_{v-1}} c_\tau \phi^{\sigma(\tau)} &= g + hx_v - \sum_\tau c_\tau \theta^\tau (x_v - x(\tau)) \\ &= g + \sum_\tau c_\tau \theta^\tau x(\tau) \end{aligned}$$

is a null-design with parameters $(t, k, v-1)$, which by induction on $v$, is an integral linear combination of the $\phi^\sigma$, $\sigma \in S_{v-1}$. This proves the theorem. □

Note that it follows from Theorem 1 that when $v \leqq k + t$, the only null design is $f \equiv 0$, which in turn implies that the only $t$-designs are the trivial design (the set of all $k$-subsets of $V$) and its multiples. This has previously been pointed out by Wilson [6]. We also remark that a topological proof of the special case of the theorem with $k = 3$, $t = 2$ has appeared in [4].

**A basis for null-designs.** Our next task will be to remove the linear dependence from the set of generators $\{\phi^\sigma : \sigma \in S_v\}$. Note that this set actually contains $v!/(t+1)!(k-t-1)!(v-k-t-1)!$ elements, substantially more than the $\binom{v}{k} - \binom{v}{t}$ we eventually shall be left with.

There are 5 kinds of linear dependence which will be removed. They are indicated symbolically as follows: For $a < b < c < d$, replace

(i) $b - a$ by $-(a - b)$;
(ii) $(b - c)a$ by $(a - c)b - (a - b)c$;
(iii) $(b - c)\bar{a}$ by $(a - c) - (a - b)$;
(iv) $(a - d)\bar{b}c$ by $(a - b)c - (a - b)d + (a - c)d - (a - c)b + (a - d)b$;
(v) $(a - d)(b - c)$ by $(a - c)(b - d) - (a - b)(c - d)$.

The meaning of this notation is as follows. If $\phi^{\sigma}$ is of the form $(x_{\sigma(1)}-x_{\sigma(2)})\cdots(x_b-x_a)\cdots x_{\sigma(2t+3)}\cdots$ with $a<b$, for example, then using (i) we replace it by $-\phi^{\sigma'}$ where

$$\sigma'(j)=\begin{cases} a & \text{if } \sigma(j)=b,\\ b & \text{if } \sigma(j)=a,\\ \sigma(j) & \text{otherwise.}\end{cases}$$

In other words, replace $\phi^{\sigma}$ by

$$-(x_{\sigma(1)}-x_{\sigma(2)})\cdots(x_a-x_b)\cdots x_{\sigma(2t+3)}\cdots.$$

In (iii) and (iv) the bar over the variable indicates that the replacement may be made provided that variable does not already occur in $\phi^{\sigma}$. Thus, with (iii), for example,

$$\phi^{\sigma}=(x_{\sigma(1)}-x_{\sigma(2)})\cdots(x_b-x_c)\cdots$$

is replaced by the two terms

$$(x_{\sigma(1)}-x_{\sigma(2)})\cdots(x_a-x_c)\cdots$$

$$-(x_{\sigma(1)}-x_{\sigma(2)})\cdots(x_a-x_b)\cdots$$

provided $x_a$ does not occur in $\phi^{\sigma}$.

Let $S^*_{v,k,t}$ consist of those $\sigma\in S_v$ which satisfy:

(a) $\sigma(1)<\sigma(3)<\cdots<\sigma(2t+1)$;

(b) $\sigma(2)<\sigma(4)<\cdots<\sigma(2t+2)$;

(c) $\sigma(2i-1)<\sigma(2i)$, $1\leqq i\leqq t+1$;

(d) $\sigma(2t+1)<\sigma(2t+3)<\sigma(2t+4)<\cdots<\sigma(k+t+1)$;

(e) $\sigma(2t+1)<\sigma(k+t+2)<\sigma(k+t+3)<\cdots<\sigma(v)$;

(f) If $2t+3\leqq i\leqq k+t+1<j\leqq v$ and $\sigma(i)<\sigma(2t+2)$ then $\sigma(i)<\sigma(j)$.

By repeatedly applying transformations (i)–(v), we can reduce the set of generators stated in Theorem 1 to a much smaller collection.

LEMMA 2. *The module $N$ is generated by $\{\phi^{\tau}\colon \tau\in S^*_{v,k,t}\}$.*

*Proof.* Because of Theorem 1 and the transformation (i), we need only to consider the polynomials $\phi^{\sigma}$ with $\sigma\in S'_v$, where

$$S'_v=\{\sigma\in S_v\colon \sigma \text{ satisfies the condition (c)}\}.$$

To each $\sigma\in S'_v$, we attach three values:

$$A_{\sigma}=\sum_{i=1}^{2t+2}\sigma(i),\qquad B_{\sigma}=\sum_{i=1}^{k+t+1}\sigma(i)$$

and

$$C_{\sigma}=\max\{\sigma(2i)-\sigma(2i-1)\colon 1\leqq i\leqq t+1\}.$$

Given two elements $\sigma,\sigma'$ of $S'_v$, we say that $\sigma'<\sigma$ if $(A_{\sigma'},B_{\sigma'},C_{\sigma'})$ is smaller than $(A_{\sigma},B_{\sigma},C_{\sigma})$ according to lexicographic order.

Let $\sigma\in S'_v$. If none of the four transformations (ii)–(v) can be performed on $\phi^{\sigma}$, then reordering the factors $(x_{\sigma(1)}-x_{\sigma(2)}),\cdots,(x_{\sigma(2t+1)}-x_{\sigma(2t+2)})$ of $\phi^{\sigma}$, the factors $x_{\sigma(2t+3)},\cdots,x_{\sigma(k+t+1)}$ of $\phi^{\sigma}$, and the unused variables $x_{\sigma(k+t+2)},\cdots,x_{\sigma(v)}$, respectively, by increasing subscript, we see that $\phi^{\sigma}=\phi^{\tau}$ for some $\tau\in S^*_{v,k,t}$. If any of the transformations (ii)–(v) can be performed on $\phi^{\sigma}$, then it is easy to check that $\phi^{\sigma}$ is a linear combination of $\phi^{\sigma'}$ with $\sigma'\in S'_v$ and $\sigma'<\sigma$. Consequently, $\phi^{\sigma}$ is generated by $\phi^{\tau}$ with $\tau\in S^*_{v,k,t}$. □

A more combinatorial way to view $S'_{v,k,t}$ is to consider it as the set of linear extensions $\sigma$ of the partial order $\prec$ on the set $\{1, \cdots, v\}$ shown in Figure 1 which satisfy (f) (where a linear extension of $\prec$ means a permutation $\sigma \in S_v$ such that $p \prec p'$ implies $\sigma(p) < \sigma(p')$).

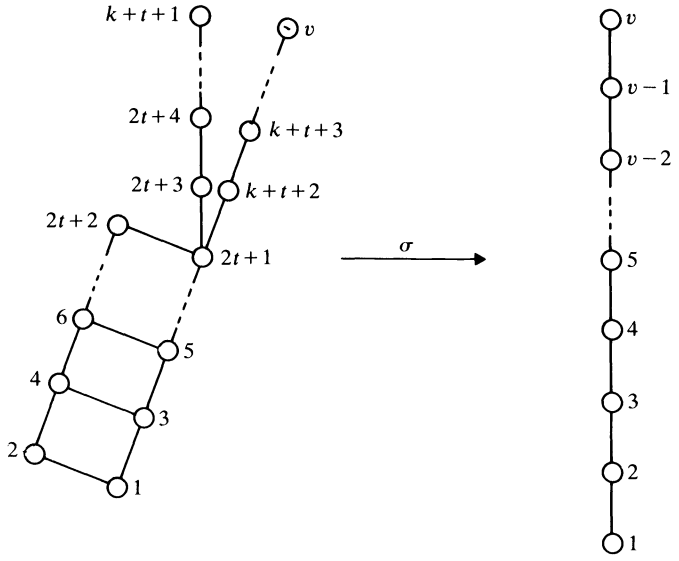



FIG. 1

Let $s_{v,k,t}$ denote $|S^*_{v,k,t}|$. The value of $s_{v,k,t}$ is unexpectedly simple.

THEOREM 3. *For* $v \geqq k+t+1$, $k \geqq t+1$,

$$s_{v,k,t} = \binom{v}{k} - \binom{v}{t}. \tag{5}$$

*Proof.* The proof will be by induction on $v$. First, assume $v = k+t+1$, i.e., $k = v-t-1$. In this case, the "tail" of $P$ beginning with $k+t+2$ is empty and conditions (e) and (f) are satisfied vacuously. We consider two cases. Since $\sigma$ is a linear extension of $P$, either $\sigma(v) = v$ or $\sigma(2t+2) = v$. If $\sigma(v) = v$, then by induction the number of $\sigma$ is $s_{v-1,v-t-2,t} = \binom{v-1}{t+1} - \binom{v-1}{t}$. If $\sigma(2t+2) = v$ then again by induction the number of $\sigma$ is $s_{v-1,v-t-1,t-1} = \binom{v-1}{t} - \binom{v-1}{t-1}$. Since the sum of these two expressions is $\binom{v}{t+1} - \binom{v}{t} = s_{v,v-t-1,t}$, the induction step is complete in this case.

Now, assume $v > k+t+1$. For a fixed $v$, we shall argue by induction on $k$. As before we distinguish cases according to the possible values of $\sigma^{-1}(v)$. In this case there are three possibilities: $v$, $k+t+1$ or $2t+2$. If $\sigma(v) = v$, then by induction on $v$ the number of these $\sigma$ is $s_{v-1,k,t} = \binom{v-1}{k} - \binom{v-1}{t}$. If $\sigma(v) = k+t+1$, then by induction on $k$ the number of these $\sigma$ is $s_{v-1,k-1,t} = \binom{v-1}{k-1} - \binom{v-1}{t}$. If $\sigma(v) = 2t+2$, then condition (f) and the induction hypothesis imply that the number of these $\sigma$ is $s_{v-1,v-t-1,t-1} = \binom{v-1}{t} - \binom{v-1}{t-1}$. Thus, the sum of these is $s_{v,k,t} = \binom{v}{k} - \binom{v}{t}$ which completes the induction step. Since (5) obviously holds for $v = 2$, the theorem is proved. □

Note that for $v = k+t+1$ or $k = t+1$, the mapping $\sigma: P \to V$ can be interpreted as a "voting sequence" for two candidates $A$ and $B$ [1] with the integers $\{2, 4, \cdots, 2t+2\}$ denoting votes for $A$, $\sigma(2i) = j$ indicating that the $j$th vote cast was the $i$th vote cast for $A$. The requirement that $\sigma$ is a linear extension implies that $A$ never leads $B$ during the voting. The number of such $\sigma$ is well known to be $\binom{v}{t+1} - \binom{v}{t}$ (see [1]).

Finally, we show that the elements of $S^*_{v,k,t}$ are linearly independent over $\mathbb{Z}$. Let $\mathbb{Z}_i[x_1, \cdots, x_v]$ denote the $\mathbb{Z}$-submodule of $\mathbb{Z}[x_1, \cdots, x_v]$ consisting of the homogeneous $x^2$-free polynomials of degree $i$. Consider the linear mapping $\mathscr{H}: \mathbb{Z}_k[x_1, \cdots, x_v] \to \mathbb{Z}_t[x_1, \cdots, x_v]$ given by defining

$$\mathscr{H}\Big(\prod_{\substack{j\in J\\ |J|=k}} x_j\Big) = \sum_{\substack{I\subseteq J\\ |I|=t}} \prod_{i\in I} x_i$$

on a basis of $\mathbb{Z}_k[x_1, \cdots, x_v]$ and extending $\mathscr{H}$ to $\mathbb{Z}_k[x_1, \cdots, x_v]$ by linearity. It is easy to see that $N = \operatorname{Ker}(\mathscr{H})$. Consider the matrix $H_{v,k,t}$ of $\mathscr{H}$ with respect to the basis of monomials of $\mathbb{Z}_k(x_1, \cdots, x_v)$ and $\mathbb{Z}_t(x_1, \cdots, x_j)$, respectively. $H_{v,k,t}$ is a $\binom{v}{t}$ by $\binom{v}{k}$ matrix with rows indexed by $t$-subsets $X$ of $V$, columns indexed by $k$-subsets $Y$ of $V$ and having as its $(X, Y)$ entry 1 if $X \subseteq Y$ and 0 otherwise. For our choice of parameters, $v \geqq k+t+1$ and $k \geqq t+1$. Thus, $H_{v,k,t}$ has at least as many columns as rows. Then as noted earlier, rank $(H_{v,k,t}) = \binom{v}{t}$. A direct way to verify this is as follows. Define the $\binom{v}{k}$ by $\binom{v}{t}$ matrix $H^* = (h^*_{Y,X})$ indexed by $k$-subsets $Y$ and $t$-subsets $X$ of $V$ by taking

$$h^*_{Y,X} = \frac{(-1)^{k-t}(k-t)}{(-1)^{|Y-X|}|Y-X|} \cdot \frac{1}{\binom{v-t}{|Y-X|}}.$$

Then the $(X, X')$ entry of $H_{v,k,t}H^*$ is

$$(-1)^{k-t}(k-t) \sum_{\substack{Y\supseteq X\\ |Y|=k}} \frac{(-1)^{|Y-X'|}}{|Y-X'|\binom{v-t}{|Y-X'|}}. \tag{6}$$

By partitioning the sum according to the values of $|Y-X'|$, standard binomial coefficient identities show that (6) is equal to 1 if $X = X'$ and 0 otherwise. Thus,

$$H_{v,k,t}H^* = I_{\binom{v}{t}}$$

where $I_x$ denotes the $x$ by $x$ identity matrix. Therefore, the rank of $\mathscr{H}$ is $\binom{v}{t}$ and $N$, being Ker $(\mathscr{H})$, has dimension $\binom{v}{k} - \binom{v}{t}$. As an immediate consequence we have:

THEOREM 4. $\{\phi^\sigma: \sigma \in S^*_{v,k,t}\}$ *forms a basis for* $N$.

**Concluding remarks.**

1. The form of the value of $s_{v,k,t}$, namely, $\binom{v}{k} - \binom{v}{t}$ suggests that there may be a more direct interpretation which would allow one to write this value down at once. If so, what is it?

2. In a similar spirit, one suspects that the inverse of $H_{v,k,t}$ given in (6) may be part of a much more general phenomenon, perhaps involving Möbius inversion. However, we have not pursued this here.

3. Is it feasible to search for new $t$-designs by starting from known (perhaps trivial) designs and augmenting them by null-designs? We have no computational evidence at present.

4. Consider the set of all polynomials $g \in \mathbb{Z}[x_1, \cdots, x_v]$ satisfying (4). These form an ideal which we denote by $I(v, t)$. Our null-designs are just the $x^2$-free homogeneous polynomials of degree $k$ in $I(v, t)$. If we were to allow repetitions of elements in the blocks of $\mathfrak{B}$, the corresponding null-designs would consist of *all* homogeneous polynomials of degree $k$ in $I(v, t)$. It is natural to ask for a set of ideal generators for $I(v, t)$ in general.

In view of Theorem 1, one would expect that $\{\psi^\sigma : \sigma \in S_v\}$ generates $I(v, t)$ when $v \geqq 2t+2$, where

$$\psi(x_1, \cdots, x_v) = (x_1 - x_2) \cdots (x_{2t+1} - x_{2t+2}).$$

For general $v$ and $t$ we do the following. Let $\pi$ be a partition of the set $\{1, \cdots, v\}$ into disjoint subsets $V_1, \cdots, V_{v-t-1}$ having as nearly equal cardinalities as possible. Define

$$\psi_\pi = \prod_{r=1}^{v-t-1} \prod_{\substack{i,j \in V_r \\ i<j}} (x_i - x_j).$$

One of us (W. Li) has conjectured that these $\psi_\pi$ generate the ideal $I(v, t)$. This is known to be true for $t = 2$.

**Note added in proof.** This conjecture has now been proved by W. Li and R. Li and will appear in a forthcoming paper.

## REFERENCES

[1] W. FELLER, *An Introduction to Probability Theory and its Applications*, vol. 1, John Wiley, New York, 1967.

[2] J. E. GRAVER AND W. B. JURKAT, *The module structure of integral designs*, J. Combinatorial Theory (A), 15 (1973), pp. 75–90.

[3] W. FOODY AND A. HEDAYAT, *On theory and applications of BIB designs with repeated blocks*, Annals Statist., 5 (1977), pp. 932–945.

[4] A. HEDAYAT AND S.-Y. R. LI, *Combinatorial topology and the trade off method in BIB designs*, Proc. Sym. on Combinatorial Mathematics and Optimal Design (Colorado State Univ.), 1978, to appear.

[5] D. R. RAY-CHAUDHURI AND R. M. WILSON, *On t-designs*, Osaka J. Math., 12 (1975), pp. 737–744.

[6] R. W. WILSON, *The necessary conditions for t-designs are sufficient for something*, Utilitas Math., 4 (1973), pp. 207–215.

SIAM J. ALG. DISC. METH.
Vol., No. 1, March 1980


0196-5212/80/0101-0003 $01.00/0

# ENSEMBLES AND LARGEST SOJOURNS OF RANDOM WALKS*

DANIEL J. KLEITMAN† AND KENNETH J. WINSTON‡

**Abstract.** We investigate the following problem: A particle starts at the origin and makes a random walk of $2n$ steps along a line. The steps are either one unit forward or one unit back, with equal probability. The particle returns to the origin for the $k$th time at step $2n$. What is the distribution in size of the largest of the $k$ "sojourns" between returns to the origin?

Using a technique similar to the introduction of the canonical ensemble in statistical mechanics, we show that if $k \leqq k_0 = A\sqrt{n \log n}$, then the probability that there is a sojourn with more than $n/(B \log n)$ steps in a walk returning to the origin for the $k$th time at step $2n$ is at least $1 - n^{-(A\sqrt{B/\pi} - A^2/4 - 2)} - n^4 4^{-n/k_0}$. Thus for large $n$ almost all walks have a "big" sojourn (one with at least $n/\log n$ steps).

**1. Introduction.** In the course of investigating the asymptotic behavior of the number of tournament score sequences, we encountered the following simple problem:

A particle starts at the origin and makes a random walk of $2n$ steps along a line. The steps are either one unit forward or one unit back, with equal probability. The particle returns to the origin for the $k$th time at step $2n$. What is the distribution in size of the largest of the $k$ "sojourns" between returns to the origin?

In this paper, we show that if $k \leqq A\sqrt{n \log n}$, then the probability that a walk returning to the origin for the $k$th time at step $2n$ has a sojourn with $n/(B \log n)$ or more steps approaches 1 as $n$ increases (Theorem 3 below is the precise formulation of the result). We also observe that the probability that a walk returning to the origin at step $2n$ has more than $A\sqrt{n \log n}$ sojourns is at most $n^{-A^2/4}$.

Feller investigated a similar problem, and observed that the results "play havoc with our intuition." We note below that the average size of a sojourn in a walk returning to the origin at step $2n$ is $2\sqrt{n/\pi}$ steps when $n$ is large. Thus we make the counter-intuitive assertion that almost all walks have a sojourn which is anomalous in that its size is far from average.

**2. Outline of arguments.** The number $W(n, k)$ of walks that return to the origin for the $k$th time at step $2n$ is known ([1, p. 76]). We review the formula in Proposition 1 below. We observe in passing that this indicates that the average sojourn in a walk returning to the origin at step $2n$ has $2\sqrt{n/\pi}$ steps when $n$ is large.

We then consider the quantity $S(n, k, z)$, which is defined as the number of walks returning to the origin for the $k$th time at step $2n$ and in which there is no sojourn with $2z$ or more steps. We introduce an ensemble which is the union over all integers $m$ of the walks counted by $S(m, k, z)$. We compute an upper bound on the size (after an appropriate normalization) of this ensemble. This upper bound also serves as a bound on $S(n, k, z)$ (after normalization) for any fixed value of $n$.

We use this bound to show (in Theorem 1) that when $k_0 = A\sqrt{n \log n}$ and $z_0 = n/(B \log n)$, the size of the ratio $S(n, k_0, z_0)/W(n, k_0)$ decreases exponentially as $B$ increases.

This establishes the desired result for $k = k_0 = A\sqrt{n \log n}$. We then prove (Theorem 2) that the ratio $S(n, k, z)/W(n, k)$ is a nondecreasing function of $k$ up to an

* Received by the editors December 19, 1978.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was supported in part by the Office of Naval Research Contract N00014-78-C-0366.

‡ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

exponentially small error term. This propagates the result of Theorem 1 down to smaller $k$, establishing our main result (Theorem 3).

We finally note that the formula for $W(n, k)$ implies that the probability that a walk returning to the origin at step $2n$ has more than $A\sqrt{n \log n}$ sojourns is less than $n^{-A^2/4}$.

**3. Proofs.** The random walks we study here are called in the literature *unrestricted one-dimensional symmetric random walks with a return to the origin at step* $2n$. By a *sojourn* of a random walk we mean a set of steps between successive returns to the origin. Let $W(n, k)$ be the number of walks returning to the origin at step $2n$ for the $k$th time. We will call walks counted by $W(n, k)$ *k-sojourn* walks.

PROPOSITION 1. *When n and k are integers with* $1 \leqq k \leqq n$,

$$W(n, k) = \frac{k \cdot 2^k}{2n-k}\binom{2n-k}{n}.$$

*Proof.* Let

$$P_k(x) = \sum_{n=k}^{\infty} W(n, k)x^n \tag{1}$$

be the generating function enumerating $k$-sojourn walks.

We can group $k$-sojourn walks by the size of the initial sojourn. This leads to a recursive sum:

$$W(n, k) = \sum_{i=1}^{n-k+1} W(i, 1)W(n-i, k-1) \tag{2}$$

which translates into a generating function identity

$$P_k(x) = P_1(x)P_{k-1}(x) = (P_1(x))^k. \tag{3}$$

There are clearly $\binom{2n}{n}$ walks returning to the origin at step $2n$. Thus we can use the binomial theorem and (3) to write

$$\sum_{n=0}^{\infty} \binom{2n}{n}x^n = (1-4x)^{-1/2} = 1 + P_1(x) + P_2(x) + \cdots = \frac{1}{1-P_1(x)}. \tag{4}$$

Solving (4) for $P_1(x)$ yields

$$P_1(x) = 1 - \sqrt{1-4x}. \tag{5}$$

An application of Lagrange's inversion formula yields

$$P_k(x) = (1-\sqrt{1-4x})^k = \sum_{n=k}^{\infty} \frac{k2^k}{2n-k}\binom{2n-k}{n}x^n. \tag{6}$$

([2, pp. 153–4]). This completes the proof of Proposition 1.

We note parenthetically that the average size of a sojourn of a walk returning to the origin at step $2n$ is $2\sqrt{n/\pi}$ when $n$ is large: differentiating (4) gives the generating function identity

$$2(1-4x)^{-3/2} = P_1'(x) + P_2'(x) + P_3'(x) + \cdots. \tag{7}$$

Using the umbral relation (3) with (7) yields

$$\frac{1-\sqrt{1-4x}}{1-4x} = P_1(x) + 2P_2(x) + 3P_3(x) + \cdots. \tag{8}$$

We equate coefficients of $x^n$ on the two sides of (8) to obtain

$$4^n - \binom{2n}{n} = \sum_{k=1}^{n} kW(n, k). \tag{9}$$

Dividing (9) by $\binom{2n}{n}$ and using Stirling's formula indicates that for large $n$ the average number of sojourns per $2n$-step walk is $\sqrt{\pi n}$. Dividing $\sqrt{\pi n}$ into $2n$ shows that when $n$ is large the average size of a sojourn in a walk returning to the origin at step $2n$ is on the order of $2\sqrt{n/\pi}$.

We now continue with the proof of our main result. Let $S^k$ be the set of all walks returning to the origin after exactly $k$ sojourns; $S^k$ is thus the union over all $m$ of the walks counted by $W(m, k)$. Let $S_z^k$ be the subset of $S^k$ consisting of walks all of whose sojourns have less than $2z$ steps. We let $S(n, k, z)$ be the number of $2n$-step walks in $S_z^k$.

We will consider the sets introduced in the previous paragraph as weighted ensembles in which walks with $2n$ steps are weighted by a factor of $4^{-n}$. We will change the $S$'s to $E$'s when the sets are considered as weighted ensembles; thus $E^k$ denotes $S^k$ considered in this way. For any weighted ensemble $E$ we let $|E|$ denote the sum of the weighted sizes of all members of $E$.

Note that the ensembles we have just introduced are similar to those of statistical mechanics. The ensemble $E_z^k$ is similar to the canonical ensemble; the walks counted by $S(n, k, z)$ are similar (after normalization) to the microcanonical ensemble. Here the size $|E|$ of an ensemble corresponds to the number of states satisfying the parameters of the ensemble.

THEOREM 1. *Let $k = A\sqrt{n \log n}$ and $z_0 = n/(B \log n)$, where $k_0, z_0$ and $n$ are integers with $1 \leqq k_0, z_0 \leqq n$. Then the probability that there is no sojourn with $2z_0$ or more steps in a random walk returning to the origin for the $k_0$th time at step $2n$ decreases exponentially with decreasing $z_0$ and increasing $n$. Specifically,*

$$\frac{S(n, k_0, z_0)}{W(n, k_0)} < n^{-(A\sqrt{B/\pi} - A^2/4 - 2)}.$$

*Proof.* If we set $x = \frac{1}{4}$ in (5), we see that $|E^1| = 1$. From Proposition 1 and Stirling's formula, we have

$$W(m, 1)4^{-m} = \frac{2 \cdot 4^{-m}}{2m-1}\binom{2m-1}{m} > \frac{1}{2\sqrt{\pi} m^{3/2}}. \tag{10}$$

Thus we have

$$|E_z^1| = |E_1| - \sum_{m=z}^{\infty} W(m, 1)4^{-m} < 1 - \frac{1}{2\sqrt{\pi}} \int_z^{\infty} m^{-3/2}\, dm < 1 - \frac{1}{\sqrt{\pi z}}. \tag{11}$$

We can bound $|E_z^k|$ by raising (11) to the $k$th power:

$$|E_z^k| = |E_z^1|^k < \left(1 - \frac{1}{\sqrt{\pi z}}\right)^k < \exp\left(\frac{-k}{\sqrt{\pi z}}\right). \tag{12}$$

But Proposition 1 and Stirling's formula yield the inequality

$$W(n, k)4^{-n} > \frac{k}{2\sqrt{\pi} n^{3/2}} \exp\left(\frac{-k^2}{4n}(1 + O(k/n))\right). \tag{13}$$

Note that $S(n, k, z)4^{-n}$ counts a component of $E_z^k$, so $|E_z^k|$ is larger than $S(n, k, z)4^{-n}$.

Hence, dividing (12) by (13) gives the inequality

$$\frac{S(n,k,z)}{W(n,k)} < \frac{2\sqrt{\pi}n^{3/2}}{k}\exp\left(\frac{k^2}{4n}-\frac{k}{\sqrt{\pi z}}\right) \tag{14}$$

when the $O(k/n)$ term is negligible in (13). Using the values of $k_0$ and $z_0$ given in the hypothesis of the theorem with inequality (14), we obtain the bound

$$\frac{S(n,k_0,z_0)}{W(n,k_0)} < \frac{2}{A}\sqrt{\frac{\pi}{\log n}}\,n^{-(A\sqrt{B/\pi}-A^2/4-1)}. \tag{15}$$

This is a lower bound on $A$ implied by $k_0 \geqq 1$; using this bound with (15) completes the proof of Theorem 1.

As $n$ increases (and the $A$ and $B$ of Theorem 1 stay fixed), the proportion $S(n,k_0,z_0)/W(n,k_0)$ of walks with small sojourns becomes smaller if $B$ is sufficiently large compared to $A$. Unfortunately, the bound given by (14) provides no information for $k$ and $z$ which are not near $k_0$ and $z_0$.

We now show that the ratio $S(n,k,z)/W(n,k)$ is a nondecreasing function of $k$ up to a small correction. This will propagate the result of Theorem 1 down to smaller $k$.

THEOREM 2. *The probability that a random walk returning to the origin for the kth time at step $2n$ has no sojourn of size $2z$ or more is (up to a correction that decays exponentially with decreasing $k$) nondecreasing in $k$. Specifically,*

$$\frac{S(n,k,z)}{W(n,k)}-\frac{S(n,k-1,z)}{W(n,k-1)} \geqq \frac{-k^2 z}{W(\lfloor n/k \rfloor,1)} > -n^3 4^{-n/k}.$$

*Proof.* Set $P(n,k,z)=S(n,k,z+1)-S(n,k,z)$. This means that $P(n,k,z)$ counts walks returning to the origin for the $k$th time at step $2n$ and in which the biggest sojourn has exactly $2z$ steps. We will show first that the difference $D(n,k,z)=P(n,k,z)/k-P(n,k-1,z)/(k-1)$ is small in magnitude.

Number the $W(z,1)$ sojourns with $2z$ steps from 1 to $W(z,1)$ in some way. We introduce a fourth argument to the functions $P$ and $S$ by using this numbering as follows: $S(n,k,z,i)$ is the number of walks counted by $S(n,k,z+1)$ in which the sojourns with $2z$ steps (if any) are numbered less than $i$. $P(n,k,z,i)$ is then $S(n,k,z,i+1)-S(n,k,z,i)$ (note that $S(n,k,z,1)=S(n,k,z)$). The introduction of this fourth argument has the effect of allowing us to graduate more finely the introduction of $2z$ step sojourns into the pool of allowable sojourns.

$P(n,k,z,i)$ satisfies the following identity:

$$\begin{aligned} P(n,k,z,i) = &\sum_{j=1}^{z-1} W(j,1)P(n-j,k-1,z,i-1) \\ &+(i-1)P(n-z,k-1,z,i)+S(n-z,k-1,z,i+1). \end{aligned} \tag{16}$$

The index $j$ is the size of the initial sojourn as in (2); the terms outside the sum correspond to walks beginning with a $2z$ step sojourn.

Consider the last term of (16); $S(n-z,k-1,z,i+1)$ counts those walks counted by $P(n,k,z,i)$ in which the initial sojourn is the $2z$ step sojourn numbered $i$. Cyclic rotation of the order of the $k$ sojourns in such walks will produce all walks enumerated by $P(n,k,z,i)$ at least once. In fact, the only duplication occurs when there are two or more instances of the $2z$ step sojourns numbered $i$ in the walk; we can thus write

$$0 \leqq S(n-z,k-1,z,i+1)-\frac{P(n,k,z,i)}{k} \leqq P(n-z,k-1,z,i). \tag{17}$$

Define $D(n, k, z, i) = P(n, k, z, i)/k - P(n, k-1, z, i)/(k-1)$. Then a calculation with (16) and (17) yields

$$\begin{aligned} D(n, k, z, i) \geqq & \sum_{j=1}^{z-1} W(j, 1) D(n-j, k-1, z, i) \\ & + (i-1) D(n-z, k-1, z, i) - \frac{P(n-z, k-2, z, i)}{k-2}. \end{aligned} \tag{18}$$

We now prove by induction on $n$ that $D(n, k, z) \geqq -kW(n, k)/W(z, 1)$. A verification shows that the induction hypothesis is true for all values of $k$ and $z$ when $n$ is small. Summing (18) over $i$ (from 1 to $W(z, 1)$) yields

$$\begin{aligned} D(n, k, z) \geqq & \sum_{j=1}^{z-1} W(j, 1) D(n-j, k-1, z) \\ & + \sum_{i=1}^{W(z,1)} (i-1) D(n-z, k-1, z, i) - \frac{P(n-z, k-2, z)}{k-2}. \end{aligned} \tag{19}$$

An application of the induction hypothesis and (2) produces the desired lower bound for $D(n, k, z)$:

$$\begin{aligned} D(n, k, z) & \geqq \sum_{j=1}^{z} W(j, 1) \frac{-(k-1)W(n-j, k-1)}{W(z, 1)} - \frac{W(n, k)}{W(z, 1)} \\ & \geqq \frac{-kW(n, k)}{W(z, 1)}. \end{aligned} \tag{20}$$

Now note that $S(n, k, z)$ is zero if $z$ is less than $(n+k)/k$. This is because each of the $k$ sojourns in walks counted by $S(n, k, z)$ can have at most $2(z-1)$ steps. From this observation and the definition of $D(n, k, z)$ it is clear that we can form a telescoping sum

$$D(n, k, z-1) + D(n, k, z-2) + \cdots + D(n, k, n/k) = \frac{S(n, k, z)}{k} - \frac{S(n, k-1, z)}{k-1}. \tag{21}$$

We can apply (20) to each term on the left-hand side of (21). This gives a lower bound for the right-hand side:

$$\frac{S(n, k, z)}{k} - \frac{S(n, k-1, z)}{k-1} \geqq -\frac{kzW(n, k)}{W(\lfloor n/k \rfloor, 1)}. \tag{22}$$

A calculation shows that

$$\begin{aligned} \frac{S(n, k, z)}{W(n, k)} & \geqq \frac{kW(n, k-1)}{(k-1)W(n, k)} \frac{S(n, k-1, z)}{W(n, k-1)} - \frac{k^2 z}{W(\lfloor n/k \rfloor, 1)} \\ & \geqq \frac{S(n, k-1, z)}{W(n, k-1)} - \frac{k^2 z}{W(\lfloor n/k \rfloor, 1)} \end{aligned} \tag{23}$$

This (along with (10)) completes the proof of the theorem, where we overestimate $z$ and $k$ by $n$ for use with the error term.

We conjecture that $S(n, k, z)/W(n, k)$ is actually nondecreasing in $k$. Intuitively, the truth of this conjecture would mean that the probability that a walk is composed of small sojourns increases monotonically as there are more sojourns. This seems intuitively clear, but finding a rigorous proof (without the error term we have used) seems to be difficult.

In any case, the correction that now appears on the right-hand side of Theorem 2 is small when $n$ is large and $k = k_0 = A\sqrt{n \log n}$. Note that this correction is decreasing (in magnitude) in $k$. If $m$ is less than $k_0$, we can form a telescoping sum of the differences on the left-hand side of the statement of Theorem 2 to obtain the difference $S(n, k_0, z)/W(n, k_0) - S(n, m, z)/W(n, m)$. Thus the following inequality holds:

$$\frac{S(n, k_0, z)}{W(n, k_0)} - \frac{S(n, m, z)}{W(n, m)} > -n^4 4^{-n/k_0}. \tag{24}$$

This allows us to state our main result:

THEOREM 3. *Let $k_0 = A\sqrt{n \log n}$ and $z_0 = n/(B \log n)$, where $k_0$, $z_0$ and $n$ are integers with $1 \leqq k_0, z_0 \leqq n$. Let $k$ be an integer, $1 \leqq k \leqq k_0$. Then the probability that there is a sojourn with $2z_0$ or more steps in a walk returning to the origin for the $k$th time at step $2n$ is at least*

$$1 - n^{-(A\sqrt{B/\pi} - A^2/4 - 2)} - n^4 4^{-n/k_0}.$$

We finally show that the proportion of the $\binom{2n}{n}$ walks returning to the origin at step $2n$ that have more than $A\sqrt{n \log n}$ sojourns is less than $n^{-A^2/4}$.

Note that if we define

$$f_n(m) = \sum_{k=m}^{n} W(n, k), \tag{25}$$

$f_n(m)$ is a decreasing function of $m$ with $f_n(1) = \binom{2n}{n}$. $f_n(m)$ is (by (4)) the coefficient of $x^n$ in

$$\frac{P_1^m(x)}{1 - P_1(x)} = P_1^m(x) + P_1^{m+1}(x) + \cdots = P_m(x) + P_{m+1}(x) + \cdots. \tag{26}$$

From ([2, p. 154]) and Stirling's formula we obtain

$$f_n(m) = 2^m \binom{2n - m}{n} < \binom{2n}{n} \sqrt{\frac{2n - m}{2n - 2m}} \exp\left(-\frac{m^2}{4n}\right). \tag{27}$$

Thus

$$f_n(A\sqrt{n \log n}) \Big/ \binom{2n}{n} < n^{-A^2/4}. \tag{28}$$

**Acknowledgment.** The authors would like to acknowledge the helpful comments of H. Niederhausen.

REFERENCES

[1] WILLIAM FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. I, 2nd ed., Wiley, New York, 1957.

[2] JOHN RIORDAN, *Combinatorial Identities*, Wiley, New York, 1968.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0004 $01.00/0

# A GROUP TESTING PROBLEM*

GERARD J. CHANG† AND F. K. HWANG‡

**Abstract.** Suppose we have two disjoint sets of items of cardinalities $m$ and $n$ where each set contains exactly one defective item. A group test is a simultaneous test on an arbitrary group of items with two possible outcomes: The group is identified as good if it contains no defective items; otherwise it is identified as defective. The problem is to find the two defective items with a worst-case minimum number of (group) tests. The first question is whether there is anything to be gained by considering the two sets together. The answer is a somewhat surprising "yes." The second question is this: Since there are $mn$ possible pairs of defective items, can we always solve the problem with $[\log_2 mn]$ tests—the information-theoretic lower bound where $[x]$ denotes the smallest integer not less than $x$. We conjecture that the answer is yes again and we provide partial evidence in favor of this conjecture. We also discuss several other related problems.

**1. Introduction.** Suppose we have two disjoint sets of items $M=\{M_1, \cdots, M_m\}$ and $N=\{N_1, \cdots, N_n\}$ where exactly one of the items in each set is defective and the others are good. A *group test* is a simultaneous test on an arbitrary group of items with two possible outcomes. The group is identified as *good* if it contains no defective item; otherwise the group is identified as *defective* but the test does not reveal how many or which ones are defective. The problem is to identify the two defective items with a worst-case minimum number of (group) tests. We will call such a problem with parameters $m$ and $n$ an $(m, n)$ problem.

By using the halving (binary search) method [2], we can identify the defective item in $M$ in $[\log_2 m]$ tests where $[x]$ is defined as the smallest integer not less than $x$. Similarly we can identify the defective item in $N$ in $[\log_2 n]$ tests. However, there are $mn$ possible pairs for the two defective items. Therefore, the information-theoretic lower bound for the number of tests is $[\log_2 mn]$ which never exceeds $[\log_2 m]+[\log_2 n]$ and is sometimes less. The question is: "Can the information-theoretic lower bound always be achieved?"

In other similar types of problems such as sorting, merging, searching etc., the general rule seems to be that the information-theoretic lower bound cannot be achieved except in special cases. In the present problem, one might also suspect that since $M$ and $N$ are disjoint and we have exact information on both sets, maybe nothing can be gained by pooling the two problems together. The latter illusion can be easily dispelled by the following example in which $M=\{M_1, M_2, M_3\}$ and $N=\{N_1, N_2, N_3, N_4, N_5\}$. We first test $\{M_1\}\cup\{N_1\}$. If that group is good, then we can find out one defective item in $M-\{M_1\}$ in one test, and the other defective item in $N-\{N_1\}$ in two tests. If $\{M_1\}\cup\{N_1\}$ is defective, we test $N_1$ next. If $N_1$ is good, we can decide $M_1$ is a defective item, and we find out the other defective in $N-\{N_1\}$ in two tests. If $N_1$ is defective, we find out the other defective in $M$ in at most two tests. In any case, four tests suffice although $[\log_2 3]+[\log_2 5]=5$.

Although the above special procedure cannot be readily generalized to deal with all values of $m$ and $n$, we conjecture that the information-theoretic lower bound can be achieved for all $m$ and $n$. We call a positive integer $m$ *favorable* if for any $n$, the number of tests required for the $(m, n)$ problem is $[\log_2 mn]$. In the paper, we show that an infinite number of $m$ are favorable. In particular, this set includes all $m \leqq 10$.

* Received by the editors March 8, 1979, and in revised form May 2, 1979.
† Institute of Mathematics, Academia Sinica, Nankang, Taiwan.
‡ Bell Laboratories, Murray Hill, New Jersey 07974.

**2. The main results.** A group testing algorithm for the $(m, n)$ problem can be represented by a rooted binary tree (see [2] for notation) where a test is associated with every internal node and the two links of that node represent the two outcomes "good" and "defective". Therefore a test sequence is represented by a path of the tree, and the two defective items identified by the test sequence are associated with the terminal node at the end of the path. The number of tests for that particular sequence is clearly just the length of the corresponding path. We will call the path on which all groups tested are defective the *all-defective path.*

Let $t(m, n)$ denote the worst-case minimum number of tests required for the $(m, n)$ problem. We first note an obvious monotonicity property of $t(m, n)$.

LEMMA 1. $t(m+1, n) \geqq t(m, n)$.

*Proof.* The proof is straightforward.

Let $m$ be an odd positive integer. Then we have

LEMMA 2. *There exists positive integers p and q such that* $mp = 2^q - 1$.

*Proof.* Since $m$ and 2 are relatively prime, there exists a positive integer $q$ such that

$$2^q \equiv 1 \pmod{m},$$

from which Lemma 2 follows immediately.

Define $s(m, n)$ to be 1 if $mn+1$ is not a power of 2, or if $mn+1$ is a power of 2, but the length of the all-defective path is exactly one unit shorter than every other path. Otherwise define $s(m, n)$ to be 0.

THEOREM 1. *Let m be an odd positive integer and let p and q be two positive integers satisfying* $mp = 2^q - 1$. *Then* $s(m, n) = 1$ *and* $t(m, n) = [\log_2 mn]$ *for* $n = 1, \cdots, p$, *imply* $s(m, n) = 1$ *and* $t(m, n) = [\log_2 mn]$ *for all n.*

*Proof.* Suppose $n > p$ and $2^{k-1} < mn < 2^k$ for some $k$. If $n \leqq p2^{k-q}$, then we can use the halving procedure on $N$ $k-q$ times and the remaining problem is $(m, n')$ where $n' \leqq p$. By our assumptions, $t(m, n') \leqq t(m, p) = q$. Therefore $t(m, n) \leqq k-q+q = k$. Furthermore, $s(m, n) = 1$ since

$$mn \leqq mp2^{k-q} = 2^k - 2^{k-q} < 2^k - 1.$$

Hence we assume that

$$n = p2^{k-q} + r \quad \text{where } r \geqq 1.$$

From

$$mn = mp2^{k-q} + mr = 2^k - 2^{k-q} + mr < 2^k,$$

we obtain

$$mr < 2^{k-q}.$$

Partition $N$ into $p$ groups of $2^{k-q}$ items $G_1, \cdots, G_p$ and one group of $r$ items $G_0$. Let $F$ be an algorithm for the $(m, p)$ problem achieving the information-theoretic lower bound. Let $F'$ be obtained from $F$ by replacing $N_j$ by $G_j$ for $j = 1, \cdots, p$, and adding $G_0$ to every group tested on the all-defective path. Associated with every terminal node of $A$, except the one in the all-defective path, is a set of pairs $M_i \times G_j$, where $M_i$ is the defective item in $M$ already identified, and $G_j$ is a subset of $N$ known to contain a defective item. Since the size of $G_j$ is $2^{k-q}$, $k-q$ more tests suffice to identify the defective item in $N$. Associated with the terminal node on the all-defective path is the union of two disjoint sets of pairs. The first set consists of pairs in $M_i \times G_j$ while the second set consists of pairs in $M \times G_0$. It takes one more test, for example, testing $G_0$, to separate these two sets. However, since the all-defective path is the short path by the

assumption $s(m, p) = 1$, only $q$ tests will have been used after the separation. We have already shown that the $M_i \times G_j$ case can be solved in $k - q$ more tests. Now the $M \times G_0$ case, actually an $(m, r)$ problem, can also be solved in $k - q$ more tests by the induction assumptions since $mr < 2^{k-q}$. Furthermore, if $mr + 1$ is a power of 2, then again by the induction assumptions there exists an algorithm for the $(m, r)$ problem such that $s(m, r) = 1$ and $t(m, r) = [\log_2 mr]$. However, $mn + 1$ is a power of 2 only if $mr + 1$ is. Therefore $s(m, n) = 1$. The proof is complete.

THEOREM 2. *If $m$ is odd and favorable and $s(m, n) = 1$ for all $n$, then $2^r m$ is favorable and $s(2^r m, n) = 1$ for all $n$ for $r = 1, 2, \cdots$.*

*Proof.* Apply the halving procedure on $Mr$ times and then use Theorem 1.

THEOREM 3. *$m$ is favorable and $s(m, n) = 1$ for all $n$ if $m = 2^k - 1$ for $k = 1, 2, \cdots$.*

*Proof.* Consider the $(m, 1)$ problem. By applying the halving procedure in the special manner that whenever the size of the current group is $2^i - 1$, we test a group of size $2^{i-1} - 1$, it can be easily verified that $s(m, 1) = 1$ and $t(m, 1) = k$. Now apply Theorem 1 with $p = 1$ to obtain Theorem 3.

THEOREM 4. *$m$ is favorable and $s(m, n) = 1$ for all $n$ if*

(i) $m = (2^{2k} - 1)/3$, $k = 1, 2, \cdots$.

(ii) $m = (2^{4k} - 1)/5$, $k = 1, 2, \cdots$.

(iii) $m = (2^{3k} - 1)/7$, $k = 1, 2, \cdots$.

(iv) $m = (2^{6k} - 1)/9$, $k = 1, 2, \cdots$.

*Proof.* (i) From Theorems 2 and 3, for $1 \leqq m \leqq 3$, $m$ is favorable and $s(m, n) = 1$ for all $m$. Now apply Theorem 1 with $p = 3$.

(ii) From Theorems 2, 3 and 4(i), for $1 \leqq m \leqq 5$, $m$ is favorable and $s(m, n) = 1$ for all $n$. Now apply Theorem 1 with $p = 5$.

(iii) and (iv) are proved along similar lines.

COROLLARY. *For $1 \leqq m \leqq 10$, $m$ is favorable and $s(m, n) = 1$ for all $n$.*

**3. Some other related problems.** One way to look at the $(m, n)$ problem is to consider an $m \times n$ matrix where row $i$ represents the item $M_i$, column $j$ represents the item $N_j$, and cell $C_{ij}$ represents the answer that $M_i$ and $N_j$ are the two defective items. A test on the group of items $M_{i_1}, \cdots, M_{i_x}, N_{j_1}, \cdots, N_{j_y}$ corresponds to a selection of rows $i_1, \cdots, i_x$ and columns $j_1, \cdots, j_y$. A selection partitions the cells into two groups, those in the selection and those not, just as a test partitions the possible answers into two groups. Therefore, the existence of an algorithm achieving the information-theoretic lower bound implies that every selection which corresponds to a test in the algorithm must partition the cells in the current group into two groups whose sizes are not separated by a power of 2. Since the current group is not necessarily in the shape of a matrix after the first selection, one must consider a more general framework to permit an induction proof.

Consider an $r \times c$ matrix with exactly $2^k \leqq rc$ entries of "1" and the rest "0". If for any distribution of the "1" entries in the matrix, we could always find a selection such that the number of "1" entries in the selection is exactly $2^{k-1}$, then we would have proved the group testing conjecture. This is because that we can imbed the original $m \times n$ matrix (filled with "1" entries) in the $r \times c$ matrix and add enough "1" entries elsewhere to make the total number of "1" entries $2^k$ (assuming $2^{k-1} < mn \leqq 2^k$). Whether such a selection always exists is unknown at present. Note that if we start with $2k$ entries of 1 and ask whether there always exists a selection partitioning it into $k$ and $k$, then a counterexample (provided by T. H. Foregger) is readily available. The counterexample has parameters $m = 5$, $n = 9$ and $2k = 44$, namely, there is a single cell with entry "0". Since the cells not in the first selection can always form a rectangle by

rearranging rows and columns, and there exists no rectangle containing either 22 or 23 cells, the first selection for an even partition is impossible. The attempt to find a counterexample for the $2^k$ case along similar lines has not been successful. In fact, Foregger and Odlyzko [1] have shown that no such counterexample can exist if the number of "0" entries is less than 6. (J. Spencer suggests that this matrix problem can be formulated in the terminology of graph theory. Namely, the problem is to decide whether a bipartite graph with $2^k$ edges can always be decomposed into two induced subgraphs with $2^{k-1}$ edges each.)

Another question raised is this: Suppose we have $k$ sets $N_1, \cdots, N_k$ with cardinalities $n_1, \cdots, n_k$, and each set contains exactly one defective item. Does there always exist an algorithm achieving the information-theoretic lower bound? (i.e. $t(n_1, \cdots, n_k) = [\log_2 \prod_{i=1}^{k} n_i]$). The answer is easily seen to be in the negative by the counterexample $k = 3$, $n_1 = n_2 = 3$ and $n_3 = 7$. It can be readily verified that there exists no test which partitions $3 \times 3 \times 7$ into 31 and 32. The fact that the group testing conjecture can only possibly be true for the two-set case makes it even more fascinating.

## REFERENCES

[1] T. H. FOREGGER AND A. D. ODLYZKO, Private communication.

[2] D. E. KNUTH, *The Art of Computing, vol. 3, Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0005 $01.00/0

# ON THE ORDER OF RANDOM CHANNEL NETWORKS*

A. MEIR†, J. W. MOON† AND J. R. POUNDER†

**Abstract.** The order of a stream with no tributaries is defined to be 1. In general, when two streams of orders $\alpha$ and $\beta$ flow together, the larger stream thus produced has order $\max\{\alpha, \beta\}$ or $\alpha+1$, according as $\alpha \neq \beta$ or $\alpha = \beta$. The order $\Omega$ of a river network $\mathcal{N}$ is the order of the highest ordered stream in $\mathcal{N}$. Our object is to investigate the distribution of $\Omega$ for random networks with $n$ sources. It follows from our results that the distribution of $\Omega$ is very highly concentrated about $1+\frac{1}{2}\log_2 n$.

**1. Introduction.** Consider an idealized river network that contains no lakes, no islands, and no junctions of more than two streams at the same place. The structure of such a river network lying upstream from a given nonjunction point may be represented by a trivalent planted tree (see, e.g., [3, p. 161] or [5, p. 67]). It is customary to call such a tree a *channel network* in this context (see [8] or [9]). The root of the tree, or the *outlet* of the network, corresponds to the point furthest downstream in the portion of the river being considered, and the other nodes of degree one, or *sources*, correspond to the points furthest upstream.

More formally, a channel network may be defined recursively as follows. The trivial network consists of a single edge joining a source node to an outlet node. Any nontrivial network may be constructed by identifying the outlets of an ordered pair of smaller networks, $\mathcal{L}$ and $\mathcal{R}$, with one end of a new edge whose other end serves as the outlet of the network $\mathcal{N}$ thus formed; the subnetworks $\mathcal{L}$ and $\mathcal{R}$ are the *main branches* of $\mathcal{N}$. Two channel networks are considered the same if and only if they have the same ordered pair of main branches.

Horton [6] (see also [7, p. 12]) introduced the *order* $\Omega = \Omega(\mathcal{N})$ of a channel network $\mathcal{N}$ as a parameter that provides a measure of the complexity of $\mathcal{N}$. If $\mathcal{N}$ is the trivial network then $\Omega(\mathcal{N}) = 1$, and if $\mathcal{N}$ is a nontrivial network with main branches $\mathcal{L}$ and $\mathcal{R}$, then

$$\Omega(\mathcal{N}) = \begin{cases} \max\{\Omega(\mathcal{L}), \Omega(\mathcal{R})\}, & \text{if } \Omega(\mathcal{L}) \neq \Omega(\mathcal{R}), \\ \Omega(\mathcal{L}) + 1, & \text{if } \Omega(\mathcal{L}) = \Omega(\mathcal{R}). \end{cases}$$

For example, the networks in Fig. 1 have order two and three (the outlets are the nodes at the bottom).

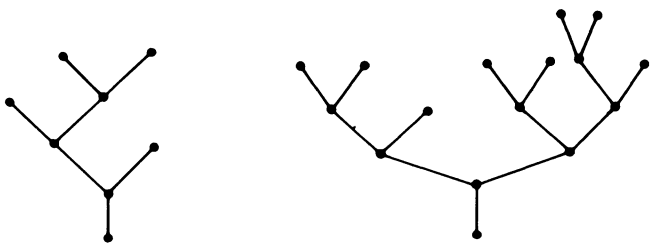



FIG. 1

Several authors have compared various statistical properties of river networks arising in nature with the corresponding properties of random channel networks and concluded that certain properties of river networks could be deduced from the

* Received by the editors April 9, 1979, and in revised form June 1, 1979. This work was supported by the National Research Council of Canada.

† Mathematics Department, University of Alberta, Edmonton, Canada, T6G 2G1.

hypothesis that the networks developed randomly; see, e.g., [9] and [10] and the references contained therein. In particular, Shreve [8] generated the numerical values of the probabilities $p(n, k)$ that a random channel network with $n$ sources has order $k$ for selected values of $n$ up to 100, among other things. He concluded, on the basis of this numerical data, that the maximum value of $p(n, k)$ for a given value of $n$ occurs for that value of $k$ for which $n^{1/(k-1)}$ is closest to 4.

Our object here is to investigate the distribution of $\Omega$ for channel networks $\mathcal{N}$ with $n$ sources. In § 2 we derive an explicit formula for $p(n, k)$ which we use in § 3 to determine the limiting behavior of $p(n, k)$ for large $k$ and $n$. The distribution of $\Omega$ is very highly concentrated; in fact, the probability that $\Omega$ is one of three particular consecutive values is at least .999 when $n$ is large. If $E\{\Omega\}$ denotes the expected values of $\Omega$ for networks with $n$ sources, then it follows from our results in § 4 that $E\{\Omega\} = \frac{1}{2}\log_2 n + O(1)$ as $n \to \infty$.

**2. Formulae for $p(n, k)$.** Let $y_n$ denote the number of channel networks $\mathcal{N}$ with $n$ sources. The generating function

$$y = y(x) = \sum_1^\infty y_n x^n$$

satisfies the relation

$$y(x) = x + y^2(x), \tag{2.1}$$

since $y_1 = 1$ and the number of sources in any nontrivial network $\mathcal{N}$ equals the sum of the number of sources in its two main branches. It follows from this relation that

$$y(x) = \tfrac{1}{2}(1-(1-4x)^{1/2}) = \sum_1^\infty \binom{2n-1}{n}\frac{x^n}{2n-1} = x + x^2 + 2x^3 + 5x^4 + \cdots, \tag{2.2}$$

a familiar result going back to Cayley [1]. We shall also need the fact that

$$y^m(x) = m \sum_{n=m}^\infty \binom{2n-m}{n}\frac{x^n}{2n-m} \tag{2.3}$$

for $m = 1, 2, \cdots$; this may readily be deduced from (2.1) with the aid of Lagrange's inversion formula.

If $f(n, k) = p(n, k)y_n$ denotes the number of channel networks $\mathcal{N}$ of order $k$ with $n$ sources, let

$$f_k = f_k(x) = \sum_{n=1}^\infty f(n, k)x^n$$

for $k = 1, 2, \cdots$. The following result is the generating-function version of a recurrence relation Shreve [8, p. 29] gave for the numbers $f(n, k)$.

LEMMA 1. *$f_1 = x$ and*

$$f_k = f_{k-1}^2 + 2f_k(f_1 + f_2 + \cdots + f_{k-1}) \quad \text{for } k = 2, 3, \cdots. \tag{2.4}$$

*Proof.* The trivial network is the only network of order 1 so $f_1 = x$. When $k \geqq 2$ it follows from the definition of the order of a network that the networks of order $k$ may be partitioned into two disjoint classes, namely, those in which both main branches have order $k-1$, and those in which one of the main branches has order $k$ and the other has order at most $k-1$. The two expressions in the right hand side of (2.4) are the generating functions for these two classes, respectively.

We now solve relation (2.4) for $f_k(x)$. It will be convenient to let $f_0(x) = x^{1/2}$ so that relation (2.4) will hold for $k \geqq 1$.

THEOREM 1. *If $k \geqq 0$, then for $x \neq 0$*

$$f_k(x) = x^{1/2} \sin \theta / \sin 2^k \theta, \tag{2.5}$$

*where* $\cos \theta = \frac{1}{2} x^{-1/2}$.

*Proof.* It follows from relation (2.4) that $f_k(x) \neq 0$ for $x \neq 0$. Hence for $x \neq 0$

$$1 = f_{k-1}^2 / f_k + 2(f_1 + f_2 + \cdots + f_{k-1})$$

if $k \geqq 1$. When we subtract this equation from the equation obtained from it by replacing $k$ with $k+1$, we find that

$$f_k^2 / f_{k+1} + 2f_k = f_{k-1}^2 / f_k,$$

or that

$$f_k / f_{k+1} + 2 = (f_{k-1} / f_k)^2.$$

Thus the functions $r_k = f_k / 2f_{k+1}$ satisfy the relations $r_0 = \frac{1}{2} x^{-1/2}$ and

$$r_k = 2r_{k-1}^2 - 1 \tag{2.6}$$

for $k \geqq 1$ and $x \neq 0$.

Recall that $\cos 2\phi = 2 \cos^2 \phi - 1$ for any $\phi$, real or complex. If we define the complex variable $\theta = \theta(x)$ for $x \neq 0$ by

$$\theta = -i \cdot \log (x^{-1/2} y),$$

where, for definiteness, $i\theta(x) < 0$ for $x > 0$, then $e^{i\theta} = x^{-1/2} y$ and by (2.1)

$$\cos \theta = \tfrac{1}{2}(x^{-1/2} y + x^{1/2} y^{-1}) = \tfrac{1}{2} x^{-1/2} = r_0.$$

Hence it follows from (2.6) that

$$r_k = \cos 2^k \theta$$

for $k \geqq 0$. We further recall that $\sin 2\phi = 2 \sin \phi \cos \phi$. Hence, if $s_k = \sin 2^k \theta$, then

$$s_k = 2 s_{k-1} r_{k-1} = s_{k-1} f_{k-1} / f_k$$

for $k \geqq 1$, and

$$s_k f_k = s_{k-1} f_{k-1} = \cdots = s_0 f_0 = x^{1/2} \sin \theta$$

for $k \geqq 0$. Therefore,

$$f_k = x^{1/2} \sin \theta / \sin 2^k \theta,$$

as required.

We adopt the notational convention that $(z)_0 = 1$ and that $(z)_j = z(z-1), \cdots, (z-j+1)$ for $j = 1, 2, \cdots$.

COROLLARY 1. *If $k \geqq 1$, then*

$$\mathrm{p}(n, k) = \frac{2}{n} \sum (2j^2 - n)(n)_j / (n+j)_j, \tag{2.7}$$

*where $j = (2\tau + 1) 2^{k-1}$ and the summation is over $\tau = 0, 1, \cdots$.*

*Proof.* Since

$$e^{i\theta} = x^{-1/2} y,$$

we have

$$\sin\theta = \tfrac{1}{2}i(e^{-i\theta} - e^{i\theta}) = \tfrac{1}{2}ix^{-1/2}(x/y - y) = \tfrac{1}{2}ix^{-1/2}(1-2y),$$

where we have appealed again to (2.1). Hence,

$$\begin{aligned} f_k &= x^{1/2}\sin\theta/\sin 2^k\theta \\ &= (1-2y)\exp(i2^k\theta)\{1-\exp(i2^{k+1}\theta)\}^{-1} \\ &= (1-2y)\sum_{\tau=0}^{\infty}(x^{-1/2}y)^{(2\tau+1)2^k}. \end{aligned}$$

We may use (2.3) to collect the coefficient of $x^n$ in this last expression; if we then divide by $y_n$ we obtain formula (2.7) after some cancellations.

COROLLARY 2. *If* $k \geqq 2$, *then*

$$f(n,k) = 4^n \cdot 2^{-k}\sum_{j=1}^{2^{k-1}-1}(-1)^{j-1}\cos^{2n-2}(j\pi/2^k)\sin^2(j\pi/2^k). \tag{2.8}$$

*Proof.* We recall that the Chebyshev polynomials $U_n(z)$ of the second kind may be defined (see [11, p. 3]) by the relation

$$U_n(z) = \sin(n+1)\theta/\sin\theta,$$

where $z = \cos\theta$. It follows from (2.5), therefore, that

$$f_k = \{2zU_{2^k-1}(z)\}^{-1}, \tag{2.9}$$

where $z = \cos\theta = \tfrac{1}{2}x^{-1/2}$.

The zeros of the function $U(z)/z = U_{2^k-1}(z)/z$ are located at the points $z_j$ and $-z_j$ where

$$z_j = \cos(j\pi/2^k) \tag{2.10}$$

for $1 \leqq j \leqq 2^{k-1}-1$. Thus when we expand in partial fractions we find that

$$\begin{aligned} z/U(z) &= \sum_{j=1}^{2^{k-1}-1}\frac{2c_jz_j^2}{z^2-z_j^2} \\ &= \sum_{j=1}^{2^{k-1}-1}2c_j\sum_{m=1}^{\infty}(z_j/z)^{2m}, \end{aligned}$$

where

$$c_j = \{U'(z_j)\}^{-1} = (-1)^{j-1}2^{-k}\sin^2(j\pi/2^k). \tag{2.11}$$

Consequently, (2.9) may be rewritten as

$$\begin{aligned} f_k &= \sum_{n=2}^{\infty}z^{-2n}\sum_{j=1}^{2^{k-1}-1}c_jz_j^{2n-2} \\ &= \sum_{n=2}^{\infty}(4x)^n\sum_{j=1}^{2^{k-1}-1}c_jz_j^{2n-2}, \end{aligned}$$

since $z^{-2} = 4x$. This proves formula (2.8), in view of relations (2.10) and (2.11).

We remark that it follows from formula (2.8) that

$$f(n,k) \sim 4^n2^{-k}\cos^{2n-2}(\pi/2^k)\sin^2(\pi/2^k)$$

if $4^k/n \to 0$ as $n \to \infty$. Consequently,

$$p(n, k) = f(n, k) y_n^{-1} \sim 4\pi^{1/2} 2^{-k} n^{3/2} \cos^{2n-2}(\pi/2^k) \sin^2(\pi/2^k)$$

if $4^k/n \to 0$, since $y_n \sim \pi^{-1/2} n^{-3/2} 4^{n-1}$. For example,

$$f(n, 3) \sim \tfrac{1}{4}\pi^{1/2} n^{3/2} \{\tfrac{1}{2}(1 + 2^{-1/2})\}^{n-1}$$

as $n \to \infty$. Furthermore, if $4^k n^{-1} \to 0$ and $4^k n^{-1/2} \to \infty$, then

$$p(n, k) \sim 4\pi^{5/2} 8^{-k} n^{3/2} e^{-\pi^2 n/4^k}$$

as $n \to \infty$.

**3. The behavior of $p(n, k)$ for large $n$ and $k$.** In what follows we shall write $\sum'$ instead of $\sum$ when the index of summation takes on only odd values.

THEOREM 2. *If $R = 4^{k-1}/n \to t$ as $n, k \to \infty$, where $t$ is any positive number, then*

$$\lim_{n,k\to\infty} p(n, k) = 2 \sum_{\nu}' (2\nu^2 t - 1) e^{-\nu^2 t}.$$

*Proof.* It follows from Corollary 1 that

$$p(n, k) = 2 \sum_{\nu}' a_\nu(n, k),$$

where

$$a_\nu(n, k) = (2\nu^2 R - 1) \frac{(n)_j}{(n + j)_j}$$

and $j = \nu 2^{k-1}$. Now,

$$e^{-j^2/(n-j)} \leqq \left(1 - \frac{j}{n}\right)^j < \frac{(n)_j}{(n+j)_j} \leqq \left(1 - \frac{j}{n+j}\right)^j \leqq e^{-j^2/(n+j)} \tag{3.1}$$

when $1 \leqq j \leqq n$. Consequently

$$\lim_{n,k\to\infty} a_\nu(n, k) = (2\nu^2 t - 1) e^{-\nu^2 t}$$

for each fixed $\nu$ if $R \to t$.

It follows from the upper bound in (3.1) that

$$|a_\nu(n, k)| \leqq (2\nu^2 R + 1) e^{-\frac{1}{2}\nu^2 R}$$

for all $\nu$, $n$, and $k$. If, as we may suppose, $n$ and $k$ are sufficiently large, then $R > \frac{1}{2}t$. The function $(2z + 1) e^{-(1/2)z}$ is decreasing when $z > \frac{3}{2}$, so

$$|a_\nu(n, k)| \leqq (\nu^2 t + 1) e^{-(1/4)\nu^2 t}$$

when $\nu^2 > 3/t$. The series $\sum' (\nu^2 t + 1) e^{-(1/4)\nu^2 t}$ converges for any positive $t$; hence, we may appeal to Tannery's theorem [4, p. 136] and conclude that

$$\lim_{n,k\to\infty} p(n, k) = 2 \sum_{\nu}' \lim_{n,k\to\infty} a_\nu(n, k) = 2 \sum_{\nu}' (2\nu^2 t - 1) e^{-\nu^2 t}.$$

Moreover, the convergence of $p(n, k)$ to its limit is uniform for $t \geqq \delta$, where $\delta$ is any fixed positive number.

Similarly, if $P(n, k) = p(n, k) + p(n, k+1) + \cdots$, then

$$\lim_{n,k\to\infty} P(n, k) = 2 \sum_{\nu=1}^{\infty} (2\nu^2 t - 1) e^{-\nu^2 t} = 1 - 4\pi^{5/2} t^{-3/2} \sum_{\nu=1}^{\infty} \nu^2 e^{-\pi^2 \nu^2/t}$$

if $R \to t$ for any positive $t$. The second expression for the limit follows from the first by Jacobi's identity for the theta-function [2, p. 11]. Some values of the function $g(t) = 2\sum'(2\nu^2 t - 1)\exp(-\nu^2 t)$ are given in Table 1.

TABLE 1

| $t$ | $g(\frac{1}{4}t)$ | $g(t)$ | $g(4t)$ |
|---|---|---|---|
| .75 | .00020 | .50163 | .49787 |
| 1.00 | .00361 | .73995 | .25641 |
| 1.25 | .01864 | .86007 | .12128 |
| 1.50 | .05287 | .89259 | .05453 |
| 1.75 | .10741 | .86887 | .02370 |
| 2.00 | .17792 | .81201 | .01006 |
| 2.25 | .25800 | .73779 | .00419 |
| 2.50 | .34157 | .65668 | .00172 |
| 2.75 | .42387 | .57535 | .00070 |
| 3.00 | .50163 | .49787 | .00028 |

The maximum value of $\mathrm{p}(n, k)$ for a given value of $n$, and the value of $k$ at which the maximum occurs, depend not so much on the magnitude of $n$, as such, as on the relative position of $n$ with respect to the powers of 4. For each integer $n$ there is a unique integer $K = K(n)$ such that

$$\frac{3}{4} \leqq \frac{4^{K-1}}{n} < 3,$$

i.e., $K = 1 + [\frac{1}{2}\log_2(3n)]$. We see from Table 1 that if $n$ is large and $4^{K-1}/n$ is close to $\frac{3}{2}$, then $\mathrm{p}(n, k)$ is close to $g(\frac{3}{2})$, or .89 approximately. If, however, $n$ is large and $4^{K-1}/n$ is close to $\frac{3}{4}$, then $\mathrm{p}(n, K)$ and $\mathrm{p}(n, K+1)$ are close to $g(\frac{3}{4})$ and $g(3)$, respectively, and hence both probabilities are close to $\frac{1}{2}$. Similarly, if $4^{K-1}/n$ is close to 3, then $\mathrm{p}(n, K-1)$ and $\mathrm{p}(n, K)$ are both close to $\frac{1}{2}$. For example, if $n = 43$ then $K = 4$ and $4^{K-1}/n = 1.488$; the actual value of $\mathrm{p}(43, 4)$ is .900, by formula (2.7). If $n = 85$, then $K$ is still 4 but now $4^{K-1}/n = .7529$; the values of $\mathrm{p}(85, 4)$ and $\mathrm{p}(85, 5)$ are .505 and .494, respectively.

In general, $K$ will be the most likely value of $\Omega$ if $4^{K-1}/n$ is not too close to $\frac{3}{4}$ or 3; in these extreme cases $K+1$ or $K-1$ (but not both) could be at least as likely. (One could attempt to remove this slight ambiguity by defining $K$ to be the integer such that $\alpha \leqq 4^{K-1}/n < 4\alpha$, where $\alpha = .748\cdots$ is the solution of the equation $g(\alpha) = g(4\alpha)$; but there seems to be little gained by this in view of the difference between $\mathrm{p}(n, k)$ and $g(4^{k-1}/n)$ for particular values of $n$ and $k$.)

We see that for most values of $n$ at least one-half the networks $\mathcal{N}$ with $n$ sources will have the same value of $\Omega$, namely $K$, and that the actual proportion may be considerably larger. As further indications of how concentrated the distribution of $\Omega$ is, we observe that it follows from Theorem 3 and some numerical calculations that

$$\min\{\mathrm{p}(n, K-1)+\mathrm{p}(n, K), \mathrm{p}(n, K)+\mathrm{p}(n, K+1)\} > .945$$

and

$$\mathrm{p}(n, K-1)+\mathrm{p}(n, K)+\mathrm{p}(n, K+1) > .999$$

when $n \leqq 100$ and for all sufficiently large values of $n$.

Shreve [8, p. 31] has suggested that the maximum value of $\mathrm{p}(n, k)$ for fixed $n$ occurs at that value of $k$ for which $n^{1/(k-1)}$ is closest to 4. This is frequently the case, but not

always; for example, $100^{1/3} = 4.64$ is closer to 4 than is $100^{1/4} = 3.16$, but

$$p(100, 4) = .36056 < p(100, 5) = .63819.$$

It is true, however, that if $E\{B\}$ denotes the expected value of the *geometric mean bifurcation ratio* $B = n^{1/(\Omega-1)}$ of a network $\mathcal{N}$ with $n$ sources and order $\Omega$ (see [8, p. 21]), then $E\{B\} \to 4$ as $n \to \infty$. We shall omit the proof of this.

**4. The moments of $\Omega$.** We have seen that the distribution of the order $\Omega$ of a random network $\mathcal{N}$ with $n$ sources is highly concentrated about $K = 1 + [\frac{1}{2}\log_2(3n)]$. It is not unreasonable, therefore, to suspect that the expected value $E\{\Omega\}$ of $\Omega$ is close to $1 + \frac{1}{2}\log_2 n$. We shall show that the absolute moments of $\Omega$ about $1 + \frac{1}{2}\log_2 n$ are in fact bounded.

THEOREM 3. *If $s$ is any fixed positive integer, then*

$$E\{|\Omega - \tfrac{1}{2}\log_2 n - 1|^s\} = O(1)$$

*as $n \to \infty$.*

Since

$$E\{a^X\} = \sum_s \frac{(\log a)^s}{s!} E\{X^s\}$$

for any nonnegative random variable $X$, it will suffice to show that

$$E\{4^{|\Omega - (1/2)\log_2 n - 1|}\} = O(1) \tag{4.1}$$

as $n \to \infty$. Before proving relation (4.1) we establish two lemmas.

LEMMA 2. *If $4^{k-1} < n$, then*

$$p(n, k) = O((n/4^{k-1})^{3/2} e^{-n/4^{k-1}})$$

*as $n \to \infty$.*

*Proof.* It is not difficult to see that the function $\cos^{2n-2} u \cdot \sin^2 u$ is decreasing in the interval $(n-1)^{-1/2} \leqq u \leqq \frac{1}{2}\pi$. The numbers $j\pi/2^k$, where $j = 1, 2, \cdots, 2^{k-1} - 1$, all lie in this interval since $4^{k-1} < n$. Consequently, formula (2.8) for $f(n, k)$ is an alternating sum of decreasing terms. When we bound $f(n, k)$ by the first term in formula (2.8) and use the inequalities $\sin u \leqq u$ and $\cos u \leqq e^{-u^2/2}$ for $0 \leqq u \leqq \frac{1}{2}\pi$, we find that

$$\begin{aligned} f(n, k) &\leqq 4^n 2^{-k} \cos^{2n-2}(\pi/2^k) \sin^2(\pi/2^k) \\ &< \pi^2 4^n 8^{-k} e^{-\pi^2(n-1)/4^k} \\ &= O(4^n 8^{-k} e^{-n/4^{k-1}}). \end{aligned}$$

The required result now follows from the fact that

$$p(n, k) = f(n, k) y_n^{-1} = O(n^{3/2} 4^{-n} f(n, k)).$$

LEMMA 3. *If $4^{k-1} \geqq n$, then*

$$p(n, k) = O((4^{k-1}/n) e^{-4^{k-1}/2n})$$

*as $n \to \infty$.*

*Proof.* Let $R = 4^{k-1}/n$. It follows from Corollary 1 and inequality (3.1) that

$$\begin{aligned} p(n, k) &\leqq 4 \sum_\nu \nu^2 \operatorname{Re}^{-(1/2)\nu^2 R} = 4 \operatorname{Re}^{-(1/2)R} \sum_\nu \nu^2 e^{-(1/2)R(\nu^2-1)} \\ &\leqq 4 \operatorname{Re}^{-(1/2)R} \sum_\nu \nu^2 e^{-(1/2)(\nu^2-1)} = O(\operatorname{Re}^{-(1/2)R}) \end{aligned}$$

as required.

We can now prove relation (4.1). Let $k_0 = [\frac{1}{2}\log_2 n] + 1$. It follows from Lemma 2 that

$$\begin{aligned}\sum_1 &= \sum_{k<k_0} (n/4^{k-1})\mathrm{p}(n,k)\\ &= O\Big(\sum_{k<k_0} (n/4^{k-1})^{5/2}\, e^{-n/4^{k-1}}\Big)\\ &= O\Big(\sum_{\nu=1}^{\infty} 4^{5/2\nu}\, e^{-4^{\nu-1}}\Big) = O(1),\end{aligned}$$

since for every $k$ such that $k<k_0$ there exists a unique integer $\nu$ such that $4^{\nu-1} \leqq n/4^{k-1} < 4^{\nu}$. It follows from Lemma 3 that

$$\begin{aligned}\sum_2 &= \sum_{k\geqq k_0} (4^{k-1}/n)\mathrm{p}(n,k)\\ &= O\Big(\sum_{k\geqq k_0} (4^{k-1}/n)^2\, e^{-4^{k-1}/2n}\Big)\\ &= O\Big(\sum_{\nu=1}^{\infty} 4^{2\nu}\, e^{-(1/2)4^{\nu-1}}\Big) = O(1),\end{aligned}$$

since for every $k$ such that $k \geqq k_0$ there exists a unique integer $\nu$ such that $4^{\nu-1} \leqq 4^{k-1}/n < 4^{\nu}$. This suffices to complete the proof of relation (4.1) since

$$\mathrm{E}\{4^{|\Omega-(1/2)\log_2 n-1|}\} = \sum_1 + \sum_2.$$

We remark that the values of $\mathrm{E}\{4^{\Omega-1}/n\}$ and $\mathrm{E}\{n/4^{\Omega-1}\}$ oscillate and do not tend to a limit as $n\to\infty$. Numerical evidence suggests, however, that

$$1.6 < \mathrm{E}\{4^{\Omega-1}/n\} < 1.9$$

and

$$.7 < \mathrm{E}\{n/4^{\Omega-1}\} < .9$$

for $n \geqq 5$.

**Acknowledgments.** We are indebted to Mr. W. Aiello for performing numerical calculations for us. The preparation of this paper was assisted by grants from the National Research Council of Canada.

## REFERENCES

[1] A. CAYLEY, *On the analytical forms called trees*, Philos. Mag., 28 (1858), pp. 374–378. (Collected Mathematical Papers, Cambridge, 4 (1891), pp. 112–115.)
[2] R. BELLMAN, *A Brief Introduction to Theta Functions*, Holt, New York, 1961.
[3] C. BERGE, *The Theory of Graphs and its Applications*, Methuen, London, 1962.
[4] T. J. BROMWICH, *An Introduction to the Theory of Infinite Series*, Macmillan, London, 1931.
[5] F. HARARY AND E. PALMER, *Graphical Enumeration*, Academic Press, New York, 1973.
[6] R. E. HORTON, *Erosional development of streams and their drainage basins: Hydrophysical approach to quantitative morphology*, Geol. Soc. Amer. Bull., 56 (1945), pp. 275–370.
[7] A. E. SCHEIDEGGER, *Theoretical Geomorphology*, Springer-Verlag, Berlin, 1961.
[8] R. L. SHREVE, *Statistical law of stream numbers*, J. Geology, 74 (1966), pp. 17–37.
[9] J. S. SMART, *Channel networks*, Advances in Hydrosciences, vol. 8, V. T. Chow, ed., Academic Press, New York, 1972, pp. 305–346.

[10] J. S. SMART AND C. WERNER, *Applications of the random model of drainage basin composition*, Earth Surface Processes, 1 (1976), pp. 219–233.

[11] G. SZEGÖ, *Orthogonal Polynomials*, Colloquium Publications, vol. 23, American Mathematical Society, Providence, RI, 1975.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0006 $01.00/0

# A COMBINATORIAL PROBLEM ARISING IN THE STUDY OF REACTION-DIFFUSION EQUATIONS*

JAMES GREENBERG,† CURTIS GREENE,‡ AND STUART HASTINGS¶

**Abstract.** We study a discrete model based on the observed behavior of excitable media. This model has the basic properties of an excitable medium, that is, a threshold phenomenon, at refractory period, and a globally stable rest point. We are mainly interested in two dimensional periodic patterns. We characterize the initial conditions which lead to such patterns, by introducing a basic invariant, the "winding number of a continuous cycle."

The problem we shall consider bears a superficial resemblance to the well-known "game" of Life, as devised by J. H. Conway [1], in which a set of simple rules determines the step-by-step evolution of certain patterns in an infinite planar grid. Our problem is also set on an infinite grid of square "cells" and proceeds in discrete time steps. However it differs from its predecessor in having a natural physical interpretation, in terms of reaction-diffusion processes. These are of current interest because of their importance in a variety of biological phenomena, including nerve conduction and morphogenesis. A related paper [2] continues the study of discrete models of such processes which was begun in [3]. However, in our opinion, the combinatorial aspects of the problem have sufficient interest to warrant separate treatment.

To describe our process we label cells $c = c_{i,j}$ with integer coordinates $(i, j)$, $-\infty < i, j < \infty$, and consider an infinite sequence $t = 0, 1, 2, 3, \cdots$ of discrete time steps. To each triple $(i, j, t)$ associate an integer $u^t_{i,j}$, called the "state" of cell $c_{i,j}$ at time $t$. These integers will come from a fixed finite set $S = \{0, 1, 2, \cdots, N\}$, where $N \geqq 2$. The initial states $u^0_{i,j}$ are chosen arbitrarily from $S$. Subsequent states $u^t_{i,j}$, $t > 0$, are then determined inductively, according to rules $A$ and $B$ below.

The inductive procedure to be described requires choosing, initially, a fixed integer $K$ with

$$1 \leqq K \leqq N/2. \tag{1}$$

The states $u = 1, 2, \cdots, K$ are called "excited," while states $K+1, \cdots, N$ are called "refractory." Also, $u = 0$ is sometimes referred to as the "rest" state. The rules for our game are

(A) If $1 \leqq u^t_{i,j} \leqq N-1$, then $u^{t+1}_{i,j} = u^t_{i,j} + 1$, while if $u^t_{i,j} = N$ then $u^{t+1}_{i,j} = 0$.

(B) Suppose that $u^t_{i,j} = 0$. To determine $u^{t+1}_{i,j}$, examine the four "adjacent" cells $c_{i',j'}$, where $|i-i'| + |j-j'| = 1$. If one or more of these cells is excited at time $t$, then $u^{t+1}_{i,j} = 1$. Otherwise, $u^{t+1}_{i,j} = 0$.

The motivation for these rules is, roughly, that excitation "diffuses" from an excited region into an adjacent resting region, but not into a refractory region. Also, once a cell is excited, its state evolves according to fixed dynamics with no diffusion effects from neighboring cells, until it returns to rest. We remark that a number of previous authors, starting with Wiener and Rosenblueth [4], have studied similar processes, usually on a computer and again in a biological context. In [2] it is shown how

* Received by the editors March 1, 1978. This work was supported by the National Science Foundation.

† Department of Mathematics, State University of New York at Buffalo, Buffalo, New York 14214.

‡ Department of Mathematics, State University of New York at Buffalo, Buffalo, New York. Currently at the Department of Mathematics, Haverford College, Haverford, PA 19041.

¶ Department of Mathematics, State University of New York at Buffalo, Buffalo, New York 14214. The work of this author was supported by the United States Army under Contract DAAG 29-75-C-0024.

these rules are related to a certain singular limit of some widely studied continuous models of reaction-diffusion processes.

The problem, broadly, is to describe how a given initial pattern $P_0 = \{(i, j, u^0_{i,j}), -\infty < i, j < \infty\}$ evolves as $t$ increases. In particular, what sorts of patterns can develop, and will the process continue indefinitely without all cells returning eventually and permanently to rest. In [3] it is observed that for $N = 2$ and $K = 1$ there is a complete solution, provided only that the number of non-zero cells at $t = 0$ is finite. There are two possibilities.

I. The pattern dies out. By this we mean that for any $(i, j)$ there is a $T_{i,j}$ such that $u^t_{i,j} = 0$ if $t \geqq T_{i,j}$. Equivalently,

$$\liminf_{t\to\infty} \{|i| + |j| \mid u^t_{i,j} \neq 0\} = \infty.$$

In other words, the pattern becomes identically zero in any finite region in finite time.

II. The pattern persists. Thus there is at least one cell $c_{i,j}$ such that

$$\{t \mid u^t_{i,j} \neq 0\}$$

is unbounded. Equivalently this is the case for each $c_{i,j}$. (This is not hard to show.) Furthermore, the pattern is eventually periodic in any finite region, and can be described as a set of rotating spirals and concentric waves radiating periodically from fixed centers. (See, for example, Fig. 4 of [3].)

In addition, one can easily determine which of I or II will occur, and locate the centers of all rotating spirals and concentric rings, by examining the initial configuration. Since this paper is devoted to $N > 2$, and no particular insight is gained by studying $N = 2$, we refer the reader to [3] for a more thorough description of the three state model.

In considering the many state version we concentrate on determining whether a pattern will persist or die out. We shall only consider patterns with a finite number of nonzero states at $t = 0$. Ideally one would like to find a necessary and sufficient condition for persistence which can be checked at $t = 0$. We have not found such a condition, though we do have nontrivial necessary conditions and sufficient conditions of this type. In addition, we give a necessary and sufficient condition for persistence which can be checked after a certain number $T = T(P_0)$ of iterations have been carried out, where $T$ depends, roughly, on the size of the initial nonzero set. (See Theorem 5.)

In order to state our results we need a measure of the distance between states in $S$. We use the metric $d(\cdot, \cdot)$ defined by

$$d(m, n) = \min \{|m - n|, N + 1 - |m - n|\} \tag{2}$$

for any $m, n$ in $S$. Equivalently, identify each $k$ in $S$ with the point

$$\hat{k} = e^{(k/(N+1))\cdot 2\pi i}$$

on the unit circle $C$ in the complex plane. Then

$$d(m, n) = \frac{N+1}{2\pi} \{\text{shorter distance from } \hat{m} \text{ to } \hat{n} \text{ on } C\}.$$

Observe that for any cell $c_{i,j}$ and any $t \geqq 0$,

$$d(u^t_{i,j}, u^{t+1}_{i,j}) \leqq 1. \tag{3}$$

Assume that $N \geqq 3$ and let

$$L = \min\left\{K+1, \frac{N+1}{4}\right\}. \tag{4}$$

THEOREM 1. *If there is a $t_0 \geqq 0$ such that $d(u_{i,j}^{t_0}, u_{i',j'}^{t_0}) < L$ whenever $c_{ij}$ and $c_{i',j'}$ are adjacent, then the pattern dies out. (Recall that $c_{i,j}$ and $c_{i',j'}$ are adjacent if $|i-i'|+|j-j'|=1$.)*

In particular, if the process is persistent, then there must be adjacent cells $c_{i,j}$ and $c_{i',j'}$ such that $d(u_{i,j}^0, u_{i',j'}^0) \geqq L$. In fact, this can be strengthened a bit.

THEOREM 2. *If the process persists, then there is a fixed pair of adjacent cells $c_{i,j}$ and $c_{i',j'}$ such that*

$$d(u_{i,j}^t, u_{i',j'}^t) \geqq L$$

*for all $t \geqq 0$.*

If $L < (N+1)/4$, then eventually an even stronger discontinuity must develop.

THEOREM 3. *If the process persists, then for any sufficiently large $t$ there is a pair of adjacent cells $c_{i,j}$ and $c_{i',j'}$ such that $d(u_{i,j}^t, u_{i',j'}^t) \geqq (N+1)/4$.*

An estimate will be given for when (at what $t$) this inequality must hold.

In order to give sufficient conditions for persistence we introduce the concept of a cycle.

DEFINITION. A *cycle* is an ordered $(M+1)$-tuple $\mathscr{C} = (c^1, c^2, c^3, \cdots, c^M, c^{M+1})$ of cells such that $c^1, \cdots, c^M$ are distinct, $c^{M+1} = c^1$, and $c^i$ is adjacent to $c^{i+1}$ for $1 \leqq i \leqq M$.

DEFINITION. A cycle $\mathscr{C}$ is said to be continuous at time $t$ if

$$d(u_i^t, u_{i+1}^t) \leqq K \quad \text{for } 1 \leqq i \leqq M,$$

where $u_i^t$ is the state of cell $c^i$ at time $t$.

For such a cycle we then define a "winding number" at time $t$. For this purpose recall the previous identification of the states $k = 0, 1, 2, \cdots, N$ with the points $\hat{k} = e^{(k/(N+1))\cdot 2\pi i}$ on the unit circle. If $m, n \in S$, let $\overline{mn}$ denote the shorter directed arc from $\hat{m}$ to $\hat{n}$. If both arcs from $\hat{m}$ to $\hat{n}$ are of the same length, let $\overline{mn}$ be the arc connecting $\hat{m}$ to $\hat{n}$ in the counter-clockwise direction. Then, for an ordered pair $(m, n)$ of integers in $S$, let

$$\sigma(m, n) = \begin{cases} d(m, n) & \text{if } \overline{mn} \text{ connects } \hat{m} \text{ to } \hat{n} \text{ in the counterclockwise direction,} \\ -d(m, n) & \text{otherwise.} \end{cases}$$

The winding number of a continuous cycle $\mathscr{C}$ at time $t$ is then defined by

$$W_t(\mathscr{C}) = \frac{1}{N+1} \sum_{i=1}^{M} \sigma(u_i^t, u_{i+1}^t).$$

It is not hard to show that $W_t(\mathscr{C})$ is an integer, and represents the net number of times the unit circle is traversed in the counterclockwise direction by the points $\widehat{u_i^t}$ as $i$ runs from 1 to $M+1$. We now give a necessary and sufficient condition for persistence.

THEOREM 4. *The pattern is persistent if and only if there is a $T \geqq 0$ and a continuous cycle $\mathscr{C}$ at time $T$ such that $W_T(\mathscr{C}) \neq 0$.*

Obviously this includes a sufficient condition for persistence which can be checked at $t = 0$.

It is desirable to find an upper bound for the smallest $T$ satisfying the conditions in Theorem 4. It can be shown by example that $T$ may be arbitrarily large if the size of the

initial nonzero set is not restricted. Our result in this direction is probably not the best possible.

THEOREM 5. *Let $R_1$ be a "diamond" shaped set of the form*

$$R_1 = \{c_{i,j} | \quad |i| + |j| \leqq m + 1\}$$

*for some m, and suppose that*

$$u^0_{i,j} = 0 \quad \text{if } |i| + |j| > m.$$

*Then the pattern persists if and only if there is a continuous cycle with nonzero winding number at time $T = n(R_1) \cdot K$, where $n(R_1)$ is the number of cells in $R_1$.*

Our final result is proved in almost the same way as Theorem 5.

THEOREM 6. *There exists an integer $p \geqq 1$ such that the process is eventually periodic with period p in any bounded region.*

In other words, there is a $p$ such that for each $m$ there is a $T_m$ with

$$u^{t+p}_{i,j} = u^t_{i,j}$$

whenever $t \geqq T_m$ and $|i| + |j| \leqq m$..

One can extend these results in several directions. For instance, arrays of cells in more than two dimensions, or with nonrectangular geometry could be considered. Other definitions of "adjacent," or "neighboring" cells might be used. For example in the plane with square cells we could say that $c_{i,j}$ and $c_{i',j'}$ are neighbors if $1 \leqq |i - i'| + |j - j'| \leqq 2$. In any of these cases Theorem 4 goes over without change. However Theorem 1 may need modification. The alternative definition of adjacent cells given above requires the number $(N+1)/3$ instead of $(N+1)/4$ in Theorem 1. On the other hand, with rectangular geometry and the definition of two cells as adjacent if they have common faces, Theorems 1–6 are essentially unchanged in higher dimensions.

*Proofs.* A basic observation is that discontinuities in a cycle do not appear spontaneously. This is implied by the following result.

LEMMA 1. *Suppose that c and d are adjacent cells with states $u^t$ and $v^t$ at time t. If*

$$d(u^{t_0}, v^{t_0}) \leqq K \tag{5}$$

*for some $t_0 \geqq 0$, then*

$$d(u^t, v^t) \leqq \max\{d(u^{t_0}, v^{t_0}), 1\}$$

*for all $t \geqq t_0$.*

COROLLARY. *If a cycle is continuous at $t_0$, then it is continuous for all $t \geqq t_0$.*

*Proof of Lemma* 1. If $u^{t_0} \neq 0$ and $v^{t_0} \neq 0$, then rule $B$, together with (2), implies that

$$d(u^{t_0+1}, v^{t_0+1}) = d(u^{t_0}, v^{t_0}). \tag{6}$$

We can therefore assume that at least one cell, say $u$, is in the resting state at $t = t_0$, i.e. that $u^{t_0} = 0$. If, in addition, $v^{t_0} = 0$ then rule $B$ implies that $d(u^{t_0+1}, v^{t_0+1}) \leqq 1$.

Next suppose that $u^{t_0} = 0$, $1 \leqq v^{t_0} \leqq K$. Again use rule $B$, to conclude that $u^{t_0+1} = 1$ and again (6) holds. Finally, if $K + 1 \leqq v^{t_0} \leqq N$ then (1), (2), and (5) imply that $v^{t_0} \geqq N - K + 1 > N/2$. But from this it follows that

$$d(v^{t_0+1}, u^{t_0}) = d(v^{t_0}, u^{t_0}) - 1.$$

From (3) and the triangle inequality we get $d(u^{t_0+1}, v^{t_0+1}) \leqq d(u^{t_0}, v^{t_0})$, completing the proof of Lemma 1.

It turns out that Theorem 4 is the key result so we prove it first.

*Proof of Theorem* 4. We begin by showing that if $\mathscr{C}$ is a continuous cycle at $t_0$, and hence for $t \geqq t_0$, then $W_t(\mathscr{C}) = W_{t_0}(\mathscr{C})$ for $t \geqq t_0$. It suffices to show that $W_{t_0+1}(\mathscr{C}) = W_{t_0}(\mathscr{C})$.

For each pair of integers $j$ and $k$ with $1 \leqq j < k \leqq M$, let

$$Q_t(j, k) = \sum_{i=j}^{k-1} \sigma(u_i^t, u_{i+1}^t).$$

In particular, $W_t(\mathscr{C}) = (1/(N+1))Q_t(1, M+1)$.

LEMMA 2. *Suppose that $j$ and $k$ are integers with $1 \leqq j < k \leqq M+1$ and that*

$$u_j^t \neq 0, \qquad u_k^t \neq 0, \quad \text{and} \quad u_i^t = 0 \quad \text{if } j < i < k. \tag{7}$$

*Then $Q_t(j, k) = Q_{t+1}(j, k)$.*

*Proof.* If $k = j+1$, then

$$Q_t(j, k) = \sigma(u_j^t, u_k^t) = \sigma(u_j^{t+1}, u_k^{t+1}) = Q_{t+1}(j, k)$$

because both $\widehat{u_j^s}$ and $\widehat{u_k^s}$ move one step counterclockwise around the unit circle as $s$ goes from $t$ to $t+1$.

Now suppose that $k \geqq j+2$. From (7) we see that

$$u_i^{t+1} = 0 \text{ or } 1 \quad \text{if } j < i < k.$$

Hence $\sigma(u_i^{t+1}, u_{i+1}^{t+1}) = u_{i+1}^{t+1} - u_i^{t+1}$ if $j < i < k-1$ and so

$$Q_{t+1}(j, k) = \sigma(u_j^{t+1}, u_{j+1}^{t+1}) + \sum_{i=j+1}^{k-2} (u_{i+1}^{t+1} - u_i^{t+1}) + \sigma(u_{k-1}^{t+1}, u_k^{t+1}),$$

where the summation term on the right is not present if $k = j+2$. If $k > j+2$ then the summation term collapses. We conclude that for any $k \geqq j+2$,

$$Q_{t+1}(j, k) = \sigma(u_j^{t+1}, u_{j+1}^{t+1}) + u_{k-1}^{t+1} - u_{j+1}^{t+1} + \sigma(u_{k-1}^{t+1}, u_k^{t+1}). \tag{8}$$

From (7) it follows that $u_j^{t+1}$ and $u_k^{t+1}$ lie in the set

$$[N-K+2, N] \cup [0, K+1].$$

Since $u_{k-1}^{t+1}$ and $u_{j+1}^{t+1}$ are 0 or 1, it is easily seen that

$$\sigma(u_j^{t+1}, u_{j+1}^{t+1}) - u_{j+1}^{t+1} = \sigma(u_j^{t+1}, 0),$$
$$\sigma(u_{k-1}^{t+1}, u_k^{t+1}) + u_{k-1}^{t+1} = \sigma(0, u_k^{t+1})$$

and from (8),

$$\begin{aligned} Q_{t+1}(j, k) &= \sigma(u_j^{t+1}, 0) + \sigma(0, u_k^{t+1}) \\ &= \sigma(u_j^t, 0) + \sigma(0, u_k^t) \\ &= Q_t(j, k), \end{aligned}$$

where we again use (7). This proves Lemma 2.

The first part of Theorem 4 then follows quickly by letting $i_1 < i_2 < \cdots < i_p$ be those $i$ such that $u_i^{t_0} \neq 0$, and observing that

$$W_t(\mathscr{C}) = \sum_{l=1}^{p} Q_t(i_l, i_{l+1}),$$

where we set $i_{p+1} = i_1$. The desired result follows by applying Lemma 2 to $j = i_m$, $k = i_{m+1}$ for $1 \leqq m \leqq p$.

*Remark.* An alternative proof that the winding number of a continuous cycle is constant proceeds as in the following sketch: The states of the cells of $\mathscr{C}$ at time $t$ can be used to define a continuous map $F_t$: $C \to C$ of the unit circle into itself. To do this, identify the cells $c^1, \cdots, c^M$ of $\mathscr{C}$ with the points $\widehat{c^j} = e^{(j/(M+1))\cdot 2\pi i}$ on $C$. Map $\widehat{c^j}$ into $\widehat{u_j^t}$. Extend this map to a continuous one from $C$ to $C$ by interpolation, using shortest arcs along $C$ and taking the counterclockwise arc in case of ties. This defines $F_t$, and similarly one can define $F_{t+1}$, using the states at time $t+1$. These two maps are easily seen to be homotopic, and hence have the same winding number.

We next consider the second part of Theorem 4, namely, that if a pattern persists, then eventually there must be a continuous cycle with nonzero winding number. To prove this some additional concepts are helpful.

DEFINITION. A *path* is an $M$-tuple $P = c^1, \cdots, c^M$ of cells such that $c^i$ is adjacent to $c^{i+1}$ for $1 \leqq i \leqq M-1$.

DEFINITION. A path $P$ is said to be *continuous at time* $t$ if $d(u_i^t, u_{i+1}^t) \leqq K$ for $1 \leqq i \leqq M-1$, where $u_i^t$ is the state of $c^i$ at time $t$.

By Lemma 1, if $P$ is continuous at $t_0$, then it is continuous for all $t \geqq t_0$.

We shall say that a cell $c$ is external to the pattern at time $t$ if there is a rectangle $R$ in the plane such that $c$ is outside $R$ but all cells which have nonzero states at time $t$ lie inside $R$.

Under the assumption that there are no continuous cycles with nonzero winding number for any $t \geqq 0$, we can define the potential of a cell $c_{i,j}$ at time $t$ for any cell which is connected to an external cell by some continuous path at time $t$. Let the path be $c^1, \cdots, c^M$, where $c^M = c_{i,j}$ and $c^1$ is external. Then we set

$$h_t(i,j) = \sum_{k=1}^{M-1} \sigma(u_i^t, u_{i+1}^t).$$

This will be the same for any continuous path connecting $c_{ij}$ to any external cell, for otherwise we could find a continuous cycle with nonzero winding number. All cells external to the initial pattern have a potential for all $t$. Also, if $h_{t_0}(i,j)$ is defined, then so is $h_t(i,j)$ for $t \geqq t_0$, and $h_t(i,j) \geqq h_{t_0}(i,j)$, $t \geqq t_0$. Among those cells which are not external to the initial pattern (a finite number), we can allow the possibility that some may never have a potential defined. In any case there must be a $t_0 \geqq 0$ such that no cells have a potential defined for the first time at some $t \geqq t_0$.

Let $c_{i,j}$ be any cell with a potential defined for $t \geqq t_0$. Suppose that for some $t \geqq t_0$, $u_{i,j}^t = 0$ and $1 \leqq u_{i',j'}^t \leqq K$ for an adjacent cell $c_{i',j'}$. Then $c_{i',j'}$ has a defined potential at time $t$ which must be higher than that of $c_{i,j}$, since we can connect $c_{i',j'}$ to the outside by a path going through $c_{i,j}$.

Let $\Delta'$ be the highest potential of any cell at time $t_0$. Let $\Delta$ be the next higher multiple of $N+1$. We claim that no cell can achieve a potential greater than $\Delta$. If not, let $c_{k,l}$ be the first cell to achieve a potential of $\Delta+1$, and suppose this occurs at $t = t_1$. Then $u_{k,l}^{t_1-1} = 0$, and $c_{k,l}$ has a neighbor $c_{k',l'}$ which is excited at time $t_1 - 1$, so that $1 \leqq u_{k',l'}^{t_1-1} \leqq K$. But then $c_{k',l'}$ must have a potential greater than $\Delta$ at $t = t_1 - 1$, which is a contradiction.

But this proves that among all those cells with potential $\Delta'$ at time $t_0$, none can ever become excited again, contradicting the persistence of the pattern.

*Proofs of Theorems* 1, 2, *and* 3. Theorems 1, 2, and 3 all follow from Theorem 4. Notice that as a consequence of Lemma 1, Theorems 1 and 2 are corollaries of Theorem 3. To prove Theorem 3, choose $t_0$ large enough to insure that at $t_0$ there is a continuous cycle $\mathscr{C} = (c^1, \cdots, c^M)$ with $W_{t_0}(\mathscr{C}) \neq 0$. This is possible by Theorem 4. We

shall show that

$$d(u_{i,j}^{t_0}, u_{i',j'}^{t_0}) \geqq \frac{N+1}{4}$$

for some pair of adjacent cells $c_{i,j}$ and $c_{i',j'}$.

For this purpose we use a different kind of continuity for a cycle. We say that a cycle $E = (e^1, \cdots, e^Q)$ is mildly continuous at $t_0$ if

$$d(u_i, u_{i+1}) \leqq \frac{N+1}{4}$$

for $1 \leqq i \leqq Q$, where $u_i$ is the state of $e^i$ at $t_0$. (We shall only be concerned with states at $t = t_0$, and so we suppress the time in this notation.)

For any mildly continuous cycle $E$, the winding number $W_{t_0}(E)$ is defined just as before. We are assuming that $W_{t_0}(E) \neq 0$.

Let $r = \sup \{d(u_{i,j}^{t_0}, u_{i',j'}^{t_0}) | c_{i,j} \text{ is adjacent to } c_{i',j'}\}$ and suppose that $r < (N+1)/4$.

LEMMA 4. *There is a mildly continuous cycle $E$ at $t_0$ which consists of exactly four cells and has nonzero winding number.*

*Proof.* Consider $\mathscr{C}$ as an ordered $M$-tuple of closed squares in the plane. Let $\gamma$ be the Jordan curve obtained by joining the centers of consecutive squares of $\mathscr{C}$. Suppose that the inside of $\gamma$ contains squares which are not in $\mathscr{C}$. Such cells we call interior to $\mathscr{C}$. Then there must be three consecutive cells $c^i, c^{i+1}, c^{i+2}$ in $\mathscr{C}$ which form three quarters of a square of four cells such that the fourth cell $c^*$ in this square lies inside $\gamma$. Since $r < (N+1)/4$, the cycle obtained by replacing $c^{i+1}$ in $\mathscr{C}$ by $c^*$ is still mildly continuous. Furthermore,

$$\begin{aligned}\sigma(u_i, u_{i+2}) &= \sigma(u_i, u_{i+1}) + \sigma(u_{i+1}, u_{i+2}) \\ &= \sigma(u_i, u^*) + \sigma(u^*, u_{i+2})\end{aligned}$$

where $u^*$ is the state of $c^*$ at $t_0$. Therefore the new cycle $\mathscr{C}^*$ has the same (nonzero) winding number as $\mathscr{C}$. However, $c^*$ has one less interior cell than $C$.

Continuing this shrinking process, we see that there must be a mildly continuous cycle $\mathscr{C}^1$ at $t_0$ with nonzero winding number and no interior cells. Again there must be consecutive cells $c^i, c^{i+1}, c^{i+2}$ of $\mathscr{C}^1$ which comprise three quarters of a square of cells, and now these can be chosen so that the fourth cell $\hat{c}$ in this square is also a cell of $\mathscr{C}^1$. We can renumber the cycle so that $\hat{c} = c^j$ for some $j > i+2$.

If $j = i+3$, then there are two cases. The cycle $\hat{\mathscr{C}} = c^i, c^{i+1}, c^{i+2}, c^{i+3}$ may have nonzero winding number, in which case we are done. If, on the other hand, $\hat{\mathscr{C}}$ has winding number 0, then omit $c^{i+1}$ and $c^{i+2}$ in $\mathscr{C}^1$ and there results a cycle $\mathscr{C}^2$ with $W_{t_0}(\mathscr{C}^2) = W_{t_0}(\mathscr{C}^1)$. Thus $\mathscr{C}^1$ has been reduced to a still smaller cycle.

Next suppose that $j > i+3$. Proceed according to whether the cycle $\mathscr{D} = (c^{i+2}, c^{i+3}, \cdots, c^j)$ has nonzero winding number or not. If $W_{t_0}(\mathscr{D}) \neq 0$, replace $\mathscr{C}^1$ with $\mathscr{C}$. If $W_{t_0}(\mathscr{D}) = 0$, eliminate $c^{i+2}, \cdots, c^j$ from $\mathscr{C}^1$.

It is thus clear that we arrive eventually at a cycle with the desired properties at $t_0$, proving Lemma 4.

But from this result we obtain a contradiction. Since

$$d(u_i, u_{i+1}) \leqq r < \frac{N+1}{4},$$

and assuming as we may that $u_1 = 0$, we have the inequalities

$$1 \leqq u_2 \leqq r, \qquad u_3 \leqq 2r, \qquad u_4 \leqq 3r.$$

On the other hand, $u_4 \geqq N - r + 1$. This gives

$$3r \geqq N - r + 1, \quad \text{or} \quad r \geqq \frac{N+1}{4},$$

which contradiction proves Theorems 1, 2, and 3.

*Proof of Theorems* 5 *and* 6. Let $R_0$ be the diamond shaped region

$$R_0 = \{c_{i,j} \,||i| + |j| \leqq m\},$$

where $m$ is chosen so large that $R_0$ contains the entire nonzero set at $t = 0$. Also, let

$$R_n = \{c_{i,j} \,||i| + |j| \leqq m + n\}$$

for $n = 1, 2, 3, \cdots$.

LEMMA 5. *Within any given* $R_n$, $n \geqq 1$, *the process proceeds independently of cells outside* $R_n$. *More precisely, if* $c_{i,j} \in R_n$ *and* $u_{i,j}^{t_0} = 1$, *then* $2 \leqq u_{i',j'}^{t_0} \leqq K + 1$ *for some cell* $c_{i',j'}$ *adjacent to* $c_{i,j}$ *and also contained in* $R_n$.

This result implies that if $\hat{u}_{i,j}^t$ is a second process, obeying the rules $A$ and $B$ if $c_{i,j} \in R_n$ but following the rule $u_{i,j}^t \equiv 0$ if $c_{i,j} \notin R_n$, then $\hat{u}_{i,j}^t = u_{i,j}^t$ in $R_n$.

*Proof of Lemma* 5. Suppose, then, that the Lemma is false, and let $t_0$ be the first time where a cell in some $R_n$, $n \geqq 1$, is excited by a cell outside $R_n$, and not simultaneously excited by a cell in $R_n$. Thus there is a cell $c_{i,j} = c^0$ in $R_n$ with $u_{i,j}^{t_0} = u_0^{t_0} = 1$, while $2 \leqq u_1^{t_0} \leqq K + 1$ for some adjacent cell $c'$ in $R_{n+1}$. Furthermore,

$$u_2^{t_0} \notin [2, K+1] \tag{8}$$

for any cell $c^2$ in $R_n$ which is adjacent to $c^0$.

However $u_0^0 - u_1^0 = 0$, so Lemma 1 implies that $u_1^{t_0} = 2$. Also, $c^1$ must in turn have been excited by a cell $c^3 \in R_n$. This follows from our hypotheses on $t_0$, since cells of $R_{n+1}$ which are not in $R_n$ cannot be adjacent to each other. Then $u_3^{t_0} = 3$. Also, $c^0$, $c^1$ and $c^3$ form three cells out of a square of four cells. Let $c^2$ be the fourth cell in this square. Then $c^2 \in R_n$ and $c^2$ is adjacent to $c^0$ and $c^3$. Since $u_3^{t_0} = 3$, $u_0^{t_0} = 1$, and $d(u_0^{t_0}, u_2^{t_0}) \leqq 1$, $d(u_3^{t_0}, u_2^{t_0}) \leqq 1$, it is seen that $u_2^{t_0} = 2$, a contradiction of (8). This proves the Lemma.

Theorem 5 follows from Lemma 5 by the following argument. Suppose the process is persistent, and let $T$ be defined as in the statement of Theorem 5. By Lemma 5, there must exist a cell $c^1$ in $R_1$ such that $1 \leqq u_1^T \leqq K$. (Otherwise the process dies out.) Cell $c^1$ must have been excited by an adjacent cell $c^2$, no more than $K$ time units before; $c^2$ in turn must have been excited by an adjacent cell $c^3$, and so on. Continuing back to time $t = 0$, we obtain a path $c^1, c^2, \cdots, c^J$ of adjacent cells in $R_1$, which is continuous at time $T$. The cells $c^i$ cannot be distinct, since each cell represents at most $K$ time units, and this would imply $T < JK \leqq n(R_1)K$. Thus a cycle must exist, and it is easy to see that this cycle has nonzero winding number.

To prove Theorem 6, observe that the process $\hat{u}_{i,j}^t$ (defined above) must be eventually periodic, since it is on a finite grid. Next observe that for sufficiently large $t$, the states of the border cells in $R_{n+1} - R_n$ are uniquely determined by the states of their neighbors in $R_n$ at times $t$ and $t-1$, by an easy argument using Lemmas 1 and 5. It follows that the process has the same period in $R_{n+1}$ and $R_n$, and hence for all $n \geqq 1$. This completes the proof.

## REFERENCES

[1] M. GARDNER, *Mathematical games*, Scientific American, Oct. 1970, p. 120; Feb. 1971, p. 112.

[2] J. M. GREENBERG, B. D. HASSARD AND S. P. HASTINGS, *Pattern formation and periodic structures in systems modeled by reaction-diffusion equations*, Bull. Amer. Math. Soc., 84 (1978), pp. 1296–1327.

[3] J. M. GREENBERG AND S. P. HASTINGS, *Spatial patterns for discrete models of diffusion in excitable media*, SIAM J. Appl. Math., 34 (1978), pp. 515–523.

[4] N. WIENER AND A. ROSENBLUETH, *The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle*, Arch. Inst. Cardiol. Mexico, 16, pp. 205–265.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0007 $01.00/0

# ON THE AVERAGE SHAPE OF BINARY TREES*

FRANK RUSKEY†

**Abstract.** The average level numbers of the leaves of a binary tree are studied, where each binary tree is regarded as being equally likely. A formula is derived for the number of binary trees with $j$th leaf at a prescribed level. The asymptotic behavior of the average level number of the $j$th leaf is determined. The average level numbers are shown to first increase and then decrease.

**1. Introduction.** The analysis of many algorithms is directly related to the level numbers of the nodes in a binary tree. In this paper we study the average level numbers of the nodes of a binary tree, where each tree is regarded as being equally likely. The average height of ordered trees was studied by de Bruin, Knuth, and Rice [1]. Although there is a direct correspondence between ordered trees and binary trees (i.e. Knuth [4]), the heights of the two types of trees do not appear to be directly related. In [1] complex variable theory was used extensively; here the methods used are entirely elementary.

By *binary tree* we mean what Knuth [4] calls an extended binary tree; that is, each node either has two sons or has no sons. The nodes with two sons are called *internal nodes* and the nodes with no sons are called *leaves*. The *level number* of a node is the length of the path from the root to that node. Let $B_n$ be the set of binary trees with $n$ internal nodes (and thus $n+1$ leaves) and let $b_n$ be the number of trees in $B_n$. It is well-known that

$$b_n = \frac{1}{n+1}\binom{2n}{n}.$$

Let $T(n, k, j)$ denote the set of trees in $B_n$ with $(j+1)$st leaf (counting from left to right) at level $k$. If $j=0$ then we shorten this to $T(n, k)$. Figure 1 shows the 4 trees in $T(4, 2, 1)$

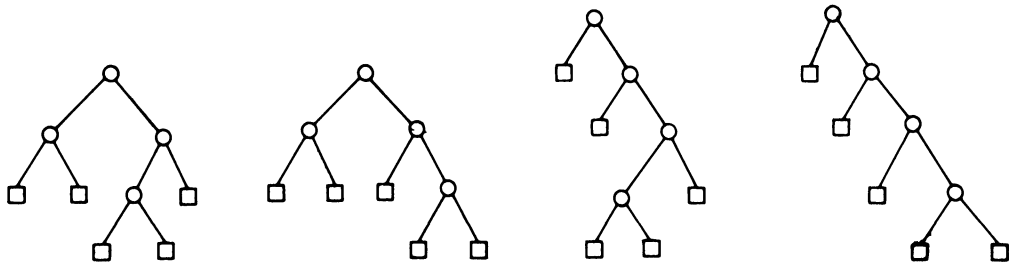


FIG. 1

Let $t(n, k, j)$ and $t(n, k)$ denote the number of trees in $T(n, k, j)$ and $T(n, k)$, respectively. We wish to become better acquainted with the numbers $t(n, k, j)$. Table 1 shows $t(6, k, j)$ for $1 \leqq k \leqq 6$ and $0 \leqq j \leqq 6$.

* Received by the editors June 5, 1979.

† Department of Mathematics, University of Victoria, Victoria, British Columbia V8W 2Y2. This work was supported in part by the National Research Council of Canada under Grant A-3379.

TABLE 1

| k \ j | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 1 | 42 | 0 | 0 | 0 | 0 | 0 | 42 |
| 2 | 42 | 28 | 10 | 8 | 10 | 28 | 42 |
| 3 | 28 | 42 | 30 | 24 | 30 | 42 | 28 |
| 4 | 14 | 36 | 42 | 40 | 42 | 36 | 14 |
| 5 | 5 | 20 | 35 | 40 | 35 | 20 | 5 |
| 6 | 1 | 6 | 15 | 20 | 15 | 6 | 1 |

The numbers $t(n, k)$ have appeared in scattered places throughout the literature. In particular, in Ruskey and Hu [6] and in Ruskey [5] it was shown that

$$t(n,k) = \sum_{\nu_1+\cdots+\nu_k=n-k} b_{\nu_1}\cdots b_{\nu_k} = \frac{k}{2n-k}\binom{2n-k}{n-k}. \tag{1}$$

We will let $\alpha(n, j)$ be the average level number of the $(j+1)$st leaf, taken over all trees in $B_n$, i.e.,

$$\alpha(n,j) = \frac{1}{b_n}\sum_k k \cdot t(n,k,j).$$

Also, $\alpha_j$ will denote the limit as $n \to \infty$ of $\alpha(n, j)$. In [6] it was shown that $\alpha_0 = 3$. Figure 2 shows $\alpha(n, j)$ for $n = 0, 1, \cdots, 6$. Observe the shape of the curves.

In the next section we will derive an expression for $t(n, k, j)$ in terms of $t(n, k)$, derive an expression for $\alpha(n, j)$, determine $\alpha_j$ exactly, and show that for fixed $n$ and

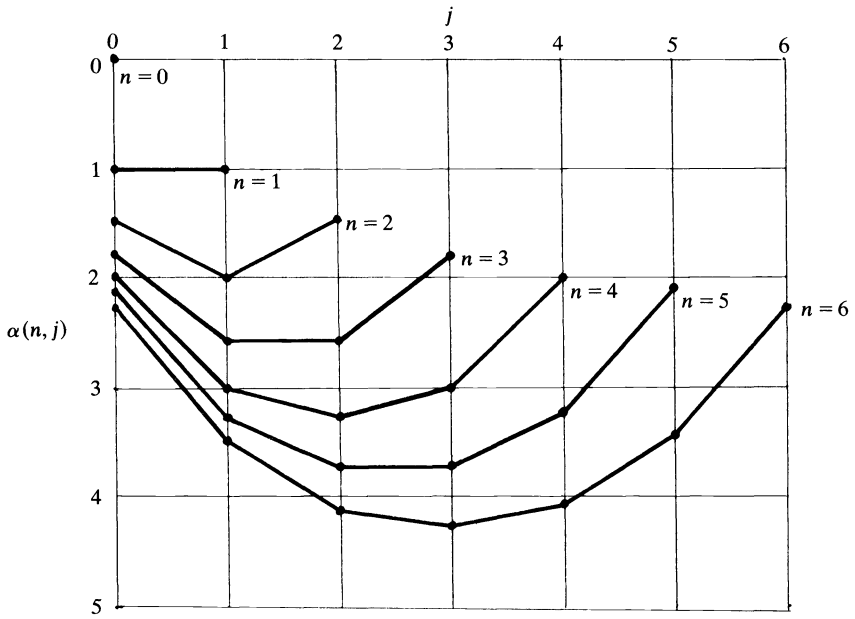



FIG. 2

$0 \leqq j \leqq n$, $\alpha(n, j)$ first increases and then decreases. The proofs of these results rely on a number of identities. These identities will be proven in the final section.

**2. Results.** We begin by trying to express $t(n, k, j)$ in terms of $t(n, k)$. Consider a tree in $T(n, k, j)$. There are $k$ internal nodes along the path from the root to the $(j+1)$st leaf. The sons of each of these internal nodes are either a path node or the root of some subtree. The number of leaves in the subtrees whose nodes are to the left of the $(j+1)$st leaf is $j$, and the number to the right is $n-j$. Thus, to get all of $T(n, k, j)$, we have to decide how many subtrees are on the left, which internal nodes are the roots of those subtrees and how many leaves are in each subtree. Let $l$ denote the number of subtrees on the left of the $(j+1)$st leaf and let $\nu_i$ represent the number of internal nodes in the $i$th subtree. Figure 3 illustrates the situation.

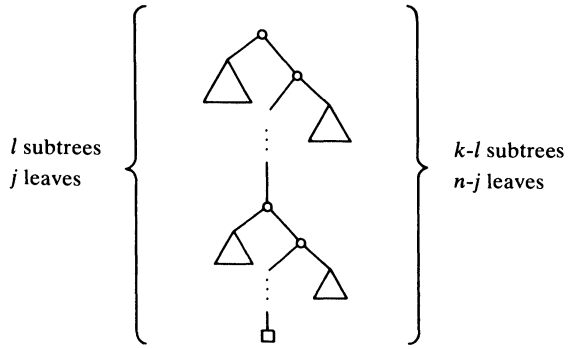



FIG. 3

Thus $t(n, k, j)$ is equal to

$$\sum_{l=0}^{k} \binom{k}{l} \Bigg( \sum_{\substack{\nu_1+\cdots+\nu_l=j-l \\ \nu_i \geqq 0}} b_{\nu_1} \cdots b_{\nu_l} \Bigg) \Bigg( \sum_{\substack{\nu_{l+1}+\cdots+\nu_k=n-j-k+l \\ \nu_i \geqq 0}} b_{\nu_{l+1}} \cdots b_{\nu_k} \Bigg).$$

By (1) we have proven

LEMMA 1.

$$t(n, k, j) = \sum_{l=0}^{k} \binom{k}{l} t(j, l) t(n-j, k-l).$$

The next two theorems rely on a number of (ultimately) binomial coefficient identities which we list here for reference and prove in the next section.

Identities:

$$(k+l)\binom{k+l}{l} = \binom{k+l}{l+1} + l \sum_{p=0}^{l+1} \binom{k+p-1}{p}, \tag{2}$$

$$\sum_{k \geqq 1} \binom{k+p-1}{p} t(n, k) = t(n+p+1, p+2), \tag{3}$$

$$\sum_{p=1}^{l} t(n+p, p+1) = t(n+l, l), \tag{4}$$

$$\lim_{n \to \infty} \frac{t(n+l, k)}{b_n} = \frac{k}{2^{k-2l+1}}, \tag{5}$$

$$\sum_{k\geq 1} t(n,k)2^k = \binom{2n}{n}, \tag{6}$$

$$\sum_{k\geq 1} k\cdot t(n,k)2^k = 4^n - \binom{2n}{n}, \tag{7}$$

$$\sum_{k\geq 1} k^2 t(n,k)2^k = (n+1)\binom{2n+2}{n+1} - 3\cdot 4^n + \binom{2n}{n}. \tag{8}$$

We now try to simplify $\sum k\cdot t(n,k,j)$ as much as possible.

THEOREM 1.

$$\sum_k k\cdot t(n,k,j) = \sum_{l=0}^{j} t(j,l)[t(n-j+l+2,l+3)+l\cdot t(n-j+l+2,l+2)].$$

*Proof.* By Lemma 1

$$\sum_{k=1}^{n} k\cdot t(n,k,j) = \sum_{k=1}^{n} k \sum_{l=0}^{j} \binom{k}{l} t(j,l)t(n-j,k-l).$$

Interchanging the sums and then shifting the inner sum by $l$ yields

$$\sum_{l=0}^{j} t(j,l) \sum_{k=1}^{n-j} (k+l)\binom{k+l}{l} t(n-j,k),$$

which by (2) is equal to

$$\sum_{l=0}^{j} t(j,l) \sum_{k=1}^{n-j} \binom{k+l}{l+1} t(n-j,k) + \sum_{l=0}^{j} l\cdot t(j,l) \sum_{k=1}^{n-j} \sum_{p=0}^{l+1} \binom{k+p-1}{p} t(n-j,k).$$

We now interchange the innermost pair of sums in the second term and apply (3) to both terms to get

$$\sum_{l=0}^{j} t(j,l)\left[ t(n-j+l+2,l+3) + l \sum_{p=0}^{l+1} t(n-j+p+1,p+2)\right].$$

A final application of (4) gives us the theorem. □

Note that $t(0,0)=1$. Using Theorem 1, particular values of $\alpha(n,j)$ can be obtained. For example, $\alpha(n,0)=3n/(n+2)$, $\alpha(n,1)=(5n-2)/(n+2)$, $\alpha(n,2)=(13n^2-18n+2)/[(n+2)(2n-1)]$.

We now determine the exact value of $\alpha_j$.

THEOREM 2.

$$\alpha_j = \frac{2}{4^j}(j+1)\binom{2j+2}{j+1} - 1.$$

*Proof.* By Theorem 1 and the definition of $\alpha_j$ we have

$$\alpha_j = \lim_{n\to\infty} \frac{1}{b_n} \sum_{l\geq 1} t(j,l)[t(n-j+l+2,l+3)+l\cdot t(n-j+l+2,l+2)],$$

which by (5) is equal to

$$\sum_{l\geq 1} t(j,l)\left[\frac{l+3}{2^{2j-l}} + \frac{l(l+2)}{2^{2j-l-1}}\right] = \frac{1}{4^j} \sum_{l\geq 1} t(j,l)2^l(2l^2+5l+3).$$

Substituting (6), (7), (8) and simplifying yields the desired result. □

COROLLARY.

$$4\sqrt{j+1}-1<\alpha_j<4\sqrt{2}\sqrt{j+1}-1.$$

*Proof.* This follows at once from the estimate

$$\frac{2^{2n}}{2\sqrt{n}}<\binom{2n}{n}<\frac{2^{2n}}{\sqrt{2n}}$$

which is given in [7, Chap. VII, § 3] (also see [8] for an even better estimate). □

We now prove that the average level numbers first increase and then decrease. Clearly the average level numbers are symmetric, i.e., $\alpha(n, j)=\alpha(n, n-j)$. Also, the average height of a tree with $n$ internal nodes is at least $\alpha(n, n/2)$. A recurrence for the $t(n, k, j)$ is given below.

LEMMA 2. *If* $k \geqq 2$ *then*

$$t(n, k, j)=\sum_{l=0}^{j-1} b_l t(n-l-1, k-1, j-l-1)+\sum_{l=j}^{n-1} b_{n-l-1} t(l, k-1, j).$$

*Proof.* Consider a tree in $T(n, k, j)$ with left subtree $T_L$ and right subtree $T_R$. Suppose that there are $l$ internal nodes in the left subtree. If $j \leqq l$ then the $(j+1)$st leaf is in $T_L$; thus $T_L \in T(l, k-1, j)$ and $T_R \in B_{n-l-1}$. This accounts for the second term. If $j>l$ then the $(j+1)$st leaf is in $T_R$; thus $T_L \in B_l$ and $T_R \in T(n-l-1, k-1, j-l-1)$. This accounts for the first term. □

Putting the result of Lemma 2 in a slightly different form:

$$t(n, k, j)=\sum_{l=0}^{n-1} b_{n-l-1}[t(l, k-1, n-j)+t(l, k-1, j)].$$

Letting $\beta(n, j)=b_n \alpha(n, j)$ we have the following corollary to Lemma 2.

COROLLARY. *For all* $0 \leqq j \leqq n$, $\beta(0, j)=0$ *and if* $n \geqq 1$ *then*

$$\beta(n, j)=b_n+\sum_{l=0}^{n-1} b_{n-l-1}[\beta(l, j)+\beta(l, n-j)].$$

*Proof.* The proof follows easily from Lemma 2 and the classic recurrence for binary trees, viz., $b_n=\sum b_l b_{n-l-1}$. □

We now prove that $\alpha(n, j)$ (equivalently, $\beta(n, j)$) is a concave function of $j$. Together with the symmetry, this shows that the level numbers first increase and then decrease.

LEMMA 3. *For all* $n \geqq 2$ *and* $0<j<n$,

$$\beta(n, j)>\tfrac{1}{2}[\beta(n, j-1)+\beta(n, j+1)].$$

*Proof.* The proof is by induction on $n$. Figure 2 shows it to be true for small values of $n$. We can use the corollary to Lemma 2 to get

$$\begin{aligned}\beta(n, j) &= b_n+\sum_{l=0}^{n-1} b_{n-l-1}[\beta(l, j)+\beta(l, n-j)] \\ &> b_n+\frac{1}{2}\sum_{l=0}^{n-1} b_{n-l-1}[\beta(l, j-1)+\beta(l, j+1)] \\ &\quad +\frac{1}{2}\sum_{l=0}^{n-1} b_{n-l-1}[\beta(l, n-j-1)+\beta(l, n-j+1)] \\ &= \frac{1}{2}[\beta(n, j-1)+\beta(n, j+1)]. \quad \square\end{aligned}$$

**3. Proof of identities.** In this section we prove the formulas (2) through (8). First we recall a few basic facts about the numbers $t(n, k)$. The proof of these may be found in [5] or [6].

$$t(n, n) = 1 \quad \text{and} \quad t(n, 0) = 0 \qquad (n \geqq 1), \tag{9}$$

$$t(n, k) = t(n-1, k-1) + t(n, k+1) \qquad (n > k \geqq 0), \tag{10}$$

$$\sum_{k \geqq l} t(n, k) = t(n+1, l+1) \qquad (l \geqq 1). \tag{11}$$

Proof of (2). This follows at once from equations 7 and 10 given on pages 53 and 54 of [4].

*Proof of* (3). This is a special case of the following equation (valid for $l \geqq 1$, $p \geqq 0$) which will be proven by induction on $p$.

$$\sum_{k \geqq l} \binom{k+p-l}{p} t(n, k) = t(n+p+1, p+l+1).$$

The base step $p = 0$ is (11), which is obtained by iterating the recurrence (10). Otherwise let $p > 0$ and consider

$$t(n+p+2, p+l+2) = \sum_{i \geqq 1} t(n+p+1, p+l+i).$$

This is true by (11). Proceeding inductively, the right-hand side of the above equation is equal to

$$\sum_{i \geqq 0} \sum_{k \geqq l+i} \binom{k+p-l-i}{p} t(n, k) = \sum_{k \geqq l} t(n, k) \sum_{i \geqq 0} \binom{k+p-l-i}{p}$$

$$= \sum_{k \geqq l} \binom{k+p+1-l}{p+1} t(n, k).$$

The last equality follows from equation 11, page 54 of [4].

*Proof of* (4). This can be proven by using induction on $l$ and the recurrence (10).

*Proof of* (5). If $0 \leqq l \leqq k$ then

$$\frac{t(n+l, k)}{b_n} = \frac{k}{2n+2l-k} \binom{2n+2l-k}{n+l-k} \Big/ \frac{1}{n+1} \binom{2n}{n}$$

$$= \frac{k(2n+2l-k-1)!n!(n+1)!}{(n+l-k)!(n-l)!(2n)!}$$

$$= \frac{k(2n+2l-k-1)_{2l-k-1}(n)_{k-l+2}}{(n+l)_{l+1}}$$

as $n \to \infty$ this becomes $k2^{2l-k-1}$

*Proof of* (6). This is the special case $j = a = 0$, $b = 2$ of equation 6.8 of Gould and Kaucky [3], which is given below.

$$\sum_{k=j}^{n} \frac{a+bk-k}{a+bn-k} \binom{a+bn-k}{n-k} b^k = \binom{a+bn-j}{n-j} b^j. \tag{12}$$

*Proof of* (7). Clearly,

$$\sum_{k=1}^{n} k \cdot t(n, k) 2^k = \sum_{k=1}^{n} \sum_{l=k}^{n} t(n, l) 2^l.$$

By (12) this is equal to

$$\sum_{k=1}^{n} \binom{2n-k}{n-k} 2^k.$$

This sum arises in "Banach's matchbox problem" and is known to be (i.e., Eisen [2, p. 150])

$$4^n - \binom{2n}{n}.$$

*Proof of* (8).

$$\sum_{k=1}^{n} k^2 t(n,k) 2^k = \sum_{k=1}^{n} \sum_{l=k}^{n} l \cdot t(n,l) 2^l$$

$$= \sum_{k=1}^{n} \sum_{l=k}^{n} \left[ (k-1) t(n,l) 2^l + \sum_{j=l}^{n} t(n,j) 2^j \right].$$

We now use (12) again to get

$$\sum_{k=1}^{n} \left[ (k-1) \binom{2n-k}{n-k} 2^k + \sum_{l=k}^{n} \binom{2n-l}{n-l} 2^l \right] = \sum_{k=1}^{n} (2k-1) \binom{2n-k}{n-k} 2^k.$$

To finish the proof we need only verify the identity given below

$$\sum_{k=1}^{n} (k+1) \binom{2n-k}{n-k} 2^k = \frac{1}{2}(n+1) \binom{2n+2}{n+1} - \binom{2n}{n}. \tag{13}$$

Using (13) and the proof of (7) finishes the proof of (8).

*Proof of* (13). Setting $n' = n+1$ yields

$$\sum_{k=1}^{n} (k+1) \binom{2n-k}{n-k} 2^k = \sum_{k=2}^{n'} k \binom{2(n'-1)-(k-1)}{n'-1} 2^{k-1}$$

$$= n' \sum_{k=2}^{n'} \frac{k}{2n'-k} \binom{2n'-k}{n'-k} 2^{k-1}$$

$$= \frac{1}{2} n' \left[ \sum_{k=1}^{n'} t(n',k) 2^k - \frac{2}{n'} \binom{2n}{n} \right]$$

$$= \frac{1}{2}(n+1) \binom{2n+2}{n+1} - \binom{2n}{n}.$$

**Acknowledgment.** The author would like to thank George Lueker for pointing out an error in an earlier version of this paper.

## REFERENCES

[1] N. G. DE BRUIN, D. E. KNUTH AND S. O. RICE, *The average height of planted plane trees*, Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1972, pp. 15–22.

[2] M. EISEN, *Elementary Combinatorial Analysis*, Gordon and Breach, 1969.

[3] H. W. GOULD AND J. KAUCKY, *Evaluation of a class of binomial coefficient summations*, J. Combinatorial Theory (1966), pp. 233–247.

[4] D. E. KNUTH, *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, Addison-Wesley, New York, 1968.

[5] F. RUSKEY, *Algorithmic solution of two combinatorial problems*, Ph.D. dissertation, APIS Dept., Univ. of California, San Diego, CA, 1978.
[6] F. RUSKEY AND T. C. HU, *Generating binary trees lexicographically*, SIAM J. Comput. 6 (1977), pp. 745–758.
[7] K. CHANDRASEKHARAN, *Introduction to Analytic Number Theory*, Springer-Verlag, New York, 1968.
[8] R. E. SHAFER, *Solution of AMM problem 6019*, Amer. Math. Monthly 84A (1977), p. 63.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0008 $01.00/0

# WORST-CASE ANALYSIS OF NETWORK DESIGN PROBLEM HEURISTICS*

RICHARD T. WONG†

**Abstract.** The Optimal Network problem (as defined by A. J. Scott, *The optimal network problem: Some computational procedures*, Trans. Res., Vol 3 (1969) pp. 201-210) consists of selecting a subset of arcs that minimizes the sum of the shortest paths between all nodes subject to a budget constraint. This paper considers the worst-case behavior of heuristics for this problem. Let $n$ be the number of nodes in the network and $\varepsilon$ be a constant between 0 and 1. For a general class of Optimal Network Problems, we show that the question of finding a solution which is always less than $n^{1-\varepsilon}$ times the optimal solution is *NP*-complete. This indicates that all polynomial-time heuristics for the problem most probably have poor worst-case performance. An upper bound for worst-case heuristic performance of 2n times the optimal solution is also derived. For a restricted version of the Optimal Network problem we describe a procedure whose maximum percentage of error is bounded by a constant.

**1. Introduction.** This paper discusses the "optimal" network problem which can be described in the following way: select a subset of arcs in a network so that the total weighted sum of the shortest paths in the network is minimized subject to the constraint that the total cost of the arcs selected does not exceed a given budget. More formally, the optimal network problem can be formulated as the following mixed integer programming problem:

$$\text{minimize} \quad \sum_{(i,j)\in A} \sum_{(k,l)\in(D\times D)} c_{ij} x_{ij}^{kl}$$

$$\text{subject to} \quad \sum_j x_{ij}^{kl} - \sum_q x_{qi}^{kl} = \begin{cases} r_{kl} & \text{if } i = k, \\ -r_{kl} & \text{if } i = l, \\ 0 & \text{otherwise,} \end{cases}$$

$$x_{ij}^{kl} \leqq r_{kl} y_{ij}, \qquad \sum_{(i,j)\in A} d_{ij} y_{ij} \leqq B,$$

$$x_{ij}^{kl} \geqq 0 \qquad (i,j)\in A \text{ and } (k,l)\in(D\times D), \qquad y_{ij} = 0 \text{ or } 1 \qquad (i,j)\in A,$$

where the decision variables are $x_{ij}^{kl}$, the amount of commodity $(k, l)$ routed on arc $(i, j)$, and $y_{ij}$, a binary variable indicating whether or not arc $(i, j)$ is to be constructed. Let D be the set of nodes and A be the set of possible arcs (*undirected*). Define $r_{kl}$ to be the amount of commodity $(k, l)$ that must be routed, $d_{ij}$ to be the construction cost of arc $(i, j)$and $c_{ij}$ to be the per unit routing cost of arc $(i, j)$. Let $B$ be the construction budget. All data $d_{ij}$ and $c_{ij}$ are assumed to be nonnegative. For technical purposes (and without any real loss of generality), we assume that all $c_{ij}$ and $d_{ij}$ are integer valued and that all problems under consideration have an optimal solution greater than zero.

This type of network design problem has potential uses in designing air, rail or highway transportation networks. Although such systems are usually much more complex than the above problem, this model could be useful in screening network configurations for more detailed study [4].

Previous work done on the optimal network problem has indicated that it is a very difficult optimization problem. Johnson, Lenstra, and Rinooy Kan have shown that the

* Received by the editors November 13, 1978. This paper constitutes part of the author's doctoral research performed at the Massachusetts Institute of Technology. This work was supported in part by the Department of Transportation Advanced Research Program (TARP) Contract No. DOT-TSC-1058.

† Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12181.

optimal network problem is *NP*-complete [8], which means that there is very probably no efficient method for solving problems of this type. Computational studies by several authors [1], [3], [4], [7] using branch and bound techniques have shown that for optimal network problems with more than 50 or 75 arcs, solution times are prohibitive. So suboptimal heuristic methods appear to be the only methods available for generating solutions to large scale network design models. Scott [16] and Dionne and Florian [4] have proposed heuristics for the optimal network problem. In the next section we review some of these procedures (for a more complete survey of the optimal network problem and related design models see Wong [17]).

An important question that arises in using heuristic techniques is the accuracy of the answers generated. One technique for evaluating heuristics is to analyze their worst-case performance. That is, we compute the maximum possible percentage deviation from the optimal solution when using the heuristic. This type of analysis is conservative in that only the worst possible error is computed, but can be useful in terms of evaluating performance guarantees for heuristics. Many researchers have analyzed heuristics for various combinatorial problems in terms of their worst-case error performance. See Garey and Johnson [6] for a survey of these results.

In this paper we analyze the worst-case behavior of a wide class of optimal network problem heuristics. The next section reviews some past work in designing such heuristics. Also some examples are given which demonstrate worst-case behavior for some of these procedures. The third section contains our main results which show that even finding an approximate optimal network solution is *NP*-complete. These results indicate that all polynomial-time heuristics for the optimal network problem probably have poor worst-case error bounds. The fourth section describes a particular heuristic algorithm whose worst-case error ratio for a restricted version of the optimal network problem is bounded by a constant that does not depend on the size of the input problem. The last section provides a summary and overview of the paper's results.

We should note that most of the previous work in this area (see [1], [4], [7], [16]) dealt with a restricted version of the optimal network problem where all required flows $r_{kl}$ were one and every arc routing cost $c_{ij}$ was equal to its construction cost $d_{ij}$. In this paper, unless otherwise noted, we assume that all required flows $r_{kl}$ are one *but* that an arc routing cost may be different from its construction cost.

**2. Previous work in optimal network problem heuristics.** Scott [16] and Dionne and Florian [4] have presented some optimal network problem heuristics which we consider here.

The first heuristic that we review is due to Dionne and Florian and was stated as follows:

(H1) (1) Construct the minimal cost spanning tree (using the construction costs $d_{ij}$ as the arc costs) as the initial network configuration.

(2) As long as the budget constraint is not violated, add to the network configuration the arc whose construction cost is the least of all arcs not yet included in the network design.

Note that if the minimal cost spanning tree is infeasible because of the construction budget constraint then the problem is infeasible.

Dionne and Florian also presented another heuristic that is a modified version of one described by Scott. It has the following description:

(H2) (0) Let $M$ be the set of arcs in the current network design. For $k \in M$, define $Q_k(M)$ as the increase in the total routing cost if arc $k$ is deleted from $M$.

(1) Initialize $M$ so it contains all arcs in the network.

(2) Find $k^*$ such that

$$L_{k^*}(M) = \frac{Q_{k^*}(M)}{d_{k^*}} = \min_{k \in M} \frac{Q_k(M)}{d_k},$$

where $d_k$ is the construction cost of arc $k$. If $L_{k^*}(M) = \infty$, then the removal of any link will disconnect the network and computation should be restarted using heuristic H1. Otherwise, delete arc $(k^*)$ from $M$ and continue with step 3.

(3) If $\sum_{k \in M} d_k > B$, i.e., the current network exceeds the construction budget, go to step 2; otherwise continue with step 4.

(4) If $B - \sum_{k \in M} d_k \geqq 0$, then introduce as many arcs as possible so that the routing cost decrease is maximized and the budget constraint is satisfied.

The quantity $L_k(M)$ can be viewed as the normalized "loss" due to deleting arc $k$. At each iteration we delete the arc whose loss is the minimum of all arcs; the process continues until a feasible solution is reached. This procedure is related to the "greedy" heuristic that has been studied previously [2].

Dionne and Florian performed computational tests to compare both heuristics. H2 performed noticeably better than H1. In fact, for many test problems H2 was able to find the optimal solution.

Now we consider the worst-case performance for these heuristics. Let us define the following terms:

$V_h(\cdot)$ = the value of the solution computed by heuristic $h$ for problem $(\cdot)$.

$V(\cdot)$ = the optimal solution value for problem $(\cdot)$.

$S(n)$ = the set of optimal network problems containing $n$ nodes.

$R_h(n) = \max_{s \in S(n)} V_h(s)/V(s)$.

$R_h(n)$ is the worst possible error ratio when heuristic $h$ is applied to optimal network problems consisting of $n$ nodes. The goal of our worst-case performance anaylsis is to compute $R_h(n)$.

We show that for both of the above heuristics, the worst-case error ratio essentially behaves as a linear function of $n$, the number of nodes in the network. Therefore the error ratio is unbounded as the size of the network increases.

Consider the following canonical example depicted in Fig. 1. Let $t_1$ and $t_2$ represent a subnetwork consisting of $Z$ nodes. Figure 2 contains a diagram of this subnetwork. Any arc connected to $t_1$ or $t_2$ is considered to be connected to the center node in the corresponding subnetwork.

The label associated with each arc in Figure 1 denotes the arc's routing cost and the construction cost respectively. The construction budget B is $13Z^5 + 8$.

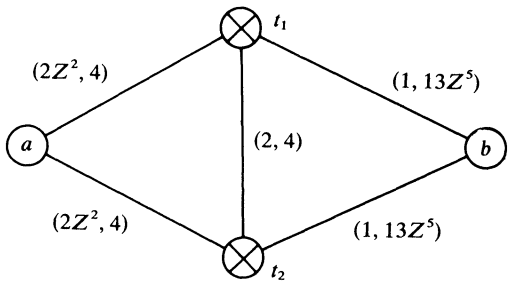



FIG. 1. *Optimal network problem example for heuristic* H2.

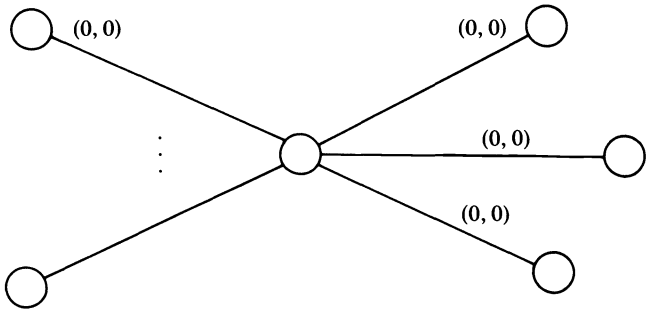



FIG. 2. *Star network representing a node.*

Using heuristic H2, we start with all arcs in the network. Then we drop arc $(t_1, t_2)$. Next, we drop arc $(t_1, b)$ or $(t_2, b)$ (the analysis is the same regardless of which arc is deleted). This leaves us with the following network depicted in Fig. 3. Recalling that all

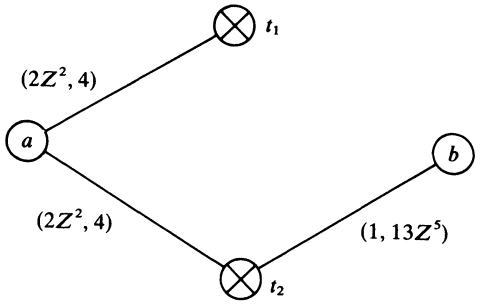



FIG. 3. *Solution computed by heuristic* H2 *for the example.*

required flows $r_{ij}$ are equal to one, we compute the cost of the above solution as

$$V_{H2} = 8Z^4 + 16Z^3 + 4Z^2 + 4Z + 2.$$

Figure 4 depicts the optimal solution to the above problem. The optimal solution has

$$V = 8Z^3 + 8Z^2 + 12Z + 6.$$

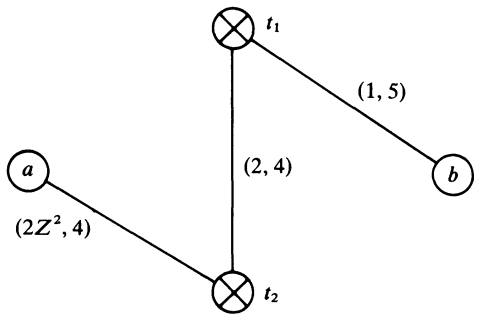



FIG. 4. *Optimal solution for optimal network example.*

The total number of nodes in the network is $2Z+2$.

$$R_{H2}(2Z+2) \geqq \frac{8Z^4+16Z^3+4Z^2+4Z+2}{8Z^3+8Z^2+12Z+6}$$

$$R_{H2}(2Z+2) \geqq Z \quad \text{for} \quad Z \geqq 1.$$

This implies

$$R_{H2}(n) \geqq \tfrac{1}{2}n - 1 \quad \text{for} \quad n = 6, 8, 10, \ldots.$$

So our example shows that the worst-case error ratio for H2 must be at least linear since our canonical example exhibits such behavior for an infinite number of network sizes.

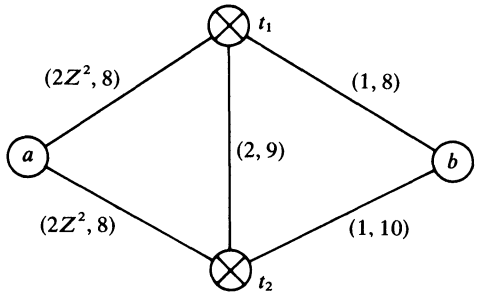



FIG. 5. *Optimal network problem example for heuristic* H1.

Heuristic H1 behaves similarly. Consider the canonical example represented by Fig. 5. Let the budget $B$ be 25. An analysis that closely follows the one given above tells us that

$$R_{H1}(n) \geqq \tfrac{1}{2}n - 1 \quad \text{for} \quad n = 6, 8, 10, \ldots.$$

So the worst-case error ratio for H1 must also be at least linear.

The above results lead us to question if there are optimal network heuristics whose worst-case behavior is better than the ones given above. The next section gives a result which indicates that all "reasonable" heuristics must probably perform nearly as badly in terms of worst-case error margins. Also we show that the worst-error ratio for the above heuristics is no worse than a linear function of network size. So the examples given above show essentially the worst possible behaviour of heuristics H1 and H2.

**3. Two theorems on the accuracy of optimal network problem heuristics.** The first result that we consider concerns the class of polynomial-time heuristics for optimal network problems, that is, the set of all optimal network design heuristics whose worst-case computation time is a polynomial function of the problem input size. As we stated previously, Johnson, Lenstra and Rinooy Kan [8] showed that the optimal network problem is *NP*-complete. Next we show that the problem of finding an optimal network design heuristic whose worst-case error ratio is less than $n^{1-\varepsilon}$, where $n$ is the number of nodes in the network and $\varepsilon$ is between 0 and 1, is also *NP*-complete. So finding a polynomial-time optimal network design heuristic that is always "close" to the optimal solution is as hard as finding a polynomial-time procedure that is always optimal. Sahni and Gonzales [15] demonstrated similar results for the traveling salesman problem (without the triangle inequality restriction), the multi-commodity

network flow problem and other combinatorial problems. Garey and Johnson [5] derived a related result for the graph-coloring problem.

Our first result can be stated in the following terms:

DEFINITION. The *approximate optimal network problem* is the following: let $\varepsilon$ be any fixed positive constant between 0 and 1, for any optimal network problem $s$ find a solution whose value is less than or equal to $n^{1-\varepsilon}V(s)$, where $n$ is the number of nodes in the problem $s$.

THEOREM 1. *The approximate optimal network problem is NP-complete.*

*Proof.* Since the optimal network problem belongs to *NP* (see [8], [9], [10]), the approximate optimal network problem must also belong to *NP*. Now we show that if the approximate problem could be solved in polynomial-time, that is, if there existed a polynomial-time heuristic $h^*$ and a constant $\varepsilon$, $0<\varepsilon<1$, such that $R_{h^*}(n)<n^{1-\varepsilon}$ for all $n$, then all of the *NP*-complete problems could be solved in polynomial-time.

Let us define a useful auxillary problem. The Steiner tree problem [9] has the following description: given a network (D, A) with node set $D$ and arc set $A$ and the data (i) $\{d_{ij}\}_{(i,j)\in A}$, the set of arc construction costs, (ii) $B$, the construction budget, and (iii) $S$, a set of nodes which is a subset of $D$, determine if there is a subtree of the network whose construction cost is less than the given budget $B$ with the property that all nodes in $S$ are connected by the subtree. Karp [9] has shown that the Steiner tree problem is *NP*-complete.

We next demonstrate that if the heuristic $h^*$ defined above exists, then the Steiner tree problem could be solved in polynomial-time. It would then follow [9], [10] that every *NP*-complete problem could be solved in polynomial-time.

Given any Steiner tree problem, transform it into an approximate optimal network problem in the following way: replace each node in the set $S$ by a subnetwork of the type pictured in Fig. 2. Each of these subnetworks should have $M^k$ nodes, where $M$ is the number of nodes in the original Steiner tree problem and $k$ is an integer constant that will be specified later. All routing and construction costs for arcs in the subnetwork should be zero.

Attach a special node $T$ to the Steiner problem network. Every "special" arc between $T$ and the set of nodes $D$ has a construction cost of zero and routing cost of one. Every arc between $T$ and a node in $S$, which is represented by a star network corresponding to Fig. 2, is connected to the center of the star network. All arcs originally in the Steiner problem network have zero routing cost and *retain their original* construction costs.

Figures 6 and 7 illustrate such a transformation. $S'$ is the set $(D\text{-}S)$. The arc labels in the original Steiner tree problem network are the arc construction costs. The arc labels

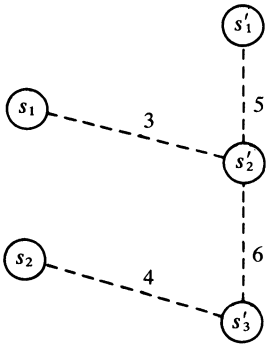



FIG. 6. *Example of a Steiner network problem (before the transformation).*

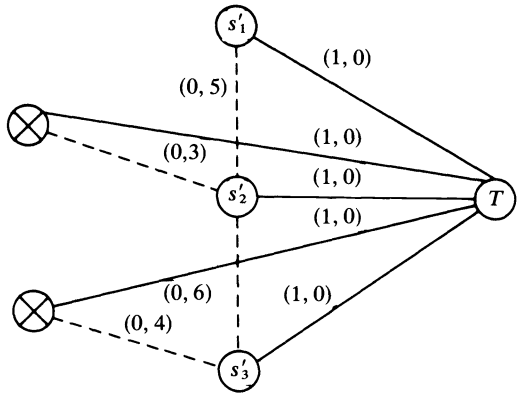



FIG. 7. *Example of a Steiner network problem (after the Transformation).*

in the modified optimal network problem indicate the arc routing and construction costs.

The construction budget for the optimal network problem is the same as the Steiner problem budget. As we have assumed throughout this paper, all required flows in the optimal network problem are equal to one.

It is important to note that this transformation to create an optimal network problem from a Steiner tree problem is a polynomially-time bounded procedure for any finite value of the parameter $k$. Also note that the size $n$ of the optimal network problem created by our transformation is at most $(M^{k+1}+1)$ nodes.

Now if one of the special arcs is utilized in the optimal network design to connect two nodes that are in $S$,

$$\text{routing cost} \geqq 4M^{2k}.$$

If all nodes in $S$ are connected with arcs from the original Steiner tree problem,

$$\text{routing cost}^1 \leqq 2M^{k+2}, \qquad k \geqq 3 \text{ and } M \geqq 4.$$

Now suppose there is a polynomial-time heuristic $h^*$ for the optimal network problem such that for some $0<\varepsilon<1$

$$R_{h^*}(n) < n^{1-\varepsilon} \quad \text{for all } n \geqq 1.$$

Since there exists a $k \geqq 3$ such that $(k-2)/(k+2) \geqq 1-\varepsilon$ we have

$$R_{h^*}(n) < n^{1-\varepsilon} < n^{(k-2)/(k+2)} \quad \text{for some } k \geqq 3.$$

Next we examine the implications of the above statement on the class of optimal network problems consisting of our transformed Steiner problems. Note that $n \leqq M^{k+1}+1$, where $M$ is the number of nodes in the original Steiner problem. Therefore, *for this class of optimal network problems*

$$R_{h^*}(n) < n^{(k-2)/(k+2)} \leqq (M^{k+1}+1)^{(k-2)/(k+2)};$$

[1] Let $RC(N_1, N_2)$ represent the cost of routing between every pair of nodes in the set $(N_1 \times N_2)$. Then we can say total routing cost $= RC(S,S)+2RC(S,S')+RC(S',S')+2RC(S,\{T\})+2RC(S',\{T\})$, where the factors of 2 are a result of the symmetry of the required flows in the network. Since all arcs from the original Steiner tree problem have routing cost zero, $RC(S,S)=0$. We can always utilize the special arcs connecting $T$ to the rest of the network so we have $RC(S,S') \leqq M^{k+2}/2$, $RC(S',S') \leqq 4M^2$, $RC(S,\{T\}) \leqq M^{k+1}$ and $RC(S',\{T\}) \leqq M$. Therefore, total routing cost $\leqq M^{k+2}+4M^2+2M^{k+1}+2M \leqq 2M^{k+2}$, $k \geqq 3$ and $M \geqq 4$.

for $M \geqq 4$,

$$(M^{k+1}+1)^{(k-2)/(k+2)} < (M^{k+2})^{(k-2)/(k+2)} = M^{k-2};$$

and,

$$R_{h^*}(n) < M^{k-2}, \qquad M \geqq 4, \quad k \geqq 3.$$

The above inequality implies that for $M \geqq 4$ the Steiner tree problem could be solved in polynomial-time by first using our polynomial-time transformation to create an optimal network problem and then applying the heuristic $h^*$ to it. The existence of a subtree satisfying the conditions of the Steiner problem could be verified by examining whether the heuristic gave a routing cost solution that was less than $4M^{2k}$.

Since the finite number of cases where $M < 4$ will not effect the polynomial-time bound of this procedure, the above inequality implies that the Steiner tree problem could be solved in polynomial-time.

Finding a heuristic $h^*$ as defined above is equivalent to solving an *NP*-complete problem, so we can say that the approximate optimal network problem is also *NP*-complete. □

We have seen that all polynomial-time bounded heuristics most probably have a worst-case error ratio that grows almost linearly with the size of the network, or at a faster rate. Next we see that for reasonable heuristics the error ratio grows no faster than linearly with the size of the network.

Before presenting this result we introduce some additional notation. Let $T$ be any spanning tree of a network and arbitrarily choose a node $R$ with degree one from $T$ and designate it as the root node. A node $f$ is the father of node $N$ if $f$ lies on the (unique) path in $T$ between $N$ and $R$ and if there is an arc in $T$ that connects $f$ and $N$. Node $s$ is the son of node $f$ if $f$ is its father. Let $w_i$ be the number of nodes which are descendants of node $i$ (i.e., nodes other than $i$ whose path to $R$ in $T$ must pass through $i$). Des($N$) is the set of nodes which are descendants of $N$.

Figure 8 contains an example illustrating these definitions. Node 1 is the root node. In this example node 2 is the father of node 5. Also $w_2 = 5$ and $w_6 = 0$.

THEOREM 2. *For optimal network problems whose routing costs satisfy the triangle inequality, any heuristic h which always produces a feasible solution will have a worst-case error ratio*

$$R_h(n) \leqq 2n \quad \textit{for all } n,$$

*where n is the number of nodes in the input network.*

*Proof.* We will show that

$$\frac{\text{routing cost of any spanning tree network}}{\text{routing cost of the complete network}} \leqq 2n.$$

(The complete network contains every arc in A, the set of all possible arcs. Note that $A$ may *not* have an arc for every pair of nodes in the network.) The theorem immediately follows from this fact since the above ratio is greater than or equal to $R_h(n)$ for any heuristic $h$ which always produces a feasible solution.

Let $T$ be any spanning tree for an optimal network problem and $C$ denote the complete network. Let $RC(T)$ represent the routing cost of network design $T$ and $n$ be the total number of nodes in $s$. Consider an arc $(i, j)$ belonging to $T$ (since we are dealing with undirected arcs, assume that for any arc $(i, j)$ in $T$, $i$ is the father of $j$). Its contribution to the total routing cost is $S(i, j) = 2(w_j + 1)(n - (w_j + 1))c_{ij}$ (that is, the

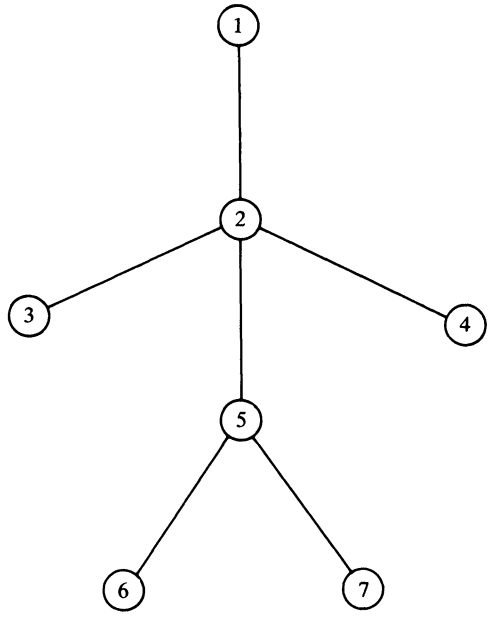



FIG. 8. *Example of a tree with root node* 1.

number of origin-destination pairs whose travel path passes through arc $(i, j)$ multiplied by the routing cost of arc $(i, j)$).

Therefore,

$$RC(T) = \sum_{(i,j)\in T} S(i, j).$$

For the routing cost of the complete network, let $a_{ij}$ be the minimum routing cost between nodes $i$ and $j$ on the complete network. Since all required flows are one we have

$$RC(C) = \sum_{(i,j)\in(D\times D)} a_{ij}.$$

Let us define the following quantity

$$C(i, j) = \sum_{k\in \mathrm{Des}(j)\cup\{j\}} 2(a_{jk} + a_{ki}) \geqq 2(w_j + 1)c_{ij}, \qquad (i, j) \in T$$

where the inequality follows from the triangle inequality for the routing costs and symmetry of the routing costs (since the arcs are undirected).

Therefore,

$$\frac{S(i, j)}{C(i, j)} \leqq \frac{2(w_j + 1)(n - (w_j + 1))c_{ij}}{2(w_j + 1)c_{ij}} \leqq n, \qquad (i, j) \in T.$$

Combining these inequalities for all $(i, j) \in T$ we have

$$\frac{\sum_{(i,j)\in T} S(i, j)}{\sum_{(i,j)\in T} C(i, j)} \leqq n$$

and since $\sum_{(i,j)\in T} S(i,j) = RC(T)$

$$\frac{RC(T)}{\sum_{(i,j)\in T} C(i,j)} \leqq n.$$

Next we show that $\sum_{(i,j)\in T} C(i,j) \leqq 2RC(C)$ and thus complete the proof.

We argue that each arc cost term $a_{st}$ appears in at most two expressions of the form $C(i,j)$ (without loss of generality assume that node $t$ is a descendent of node $s$). $a_{st}$ appears in the expression $C(i,j)$ only if

(1) $j$ equals $s$. Recall that since $i$ must be the father of $j$, $i$ must be the father of $s$.

(2) $i$ equals $s$ and $t$ belongs to $\text{Des}(j)\cup\{j\}$.

The first situation can only happen once since node $s$ must have a unique father. The second situation can only occur once since if it happened twice, for example, with $C(i,j_1)$ and $C(i,j_2)$, $j_1 \neq j_2$, then between $s$ and $t$ there would be two distinct paths in the tree T.

Since $\text{RC(C)} = \sum_{(s,t)\in D\times D} a_{st}$ and the term $a_{st}$ occurs in at most two terms of the form $C(i,j)$ we have

$$\sum_{(i,j)\in T} C(i,j) \leqq 2RC(C).$$

Therefore,

$$\frac{RC(T)}{2RC(C)} \leqq n$$

or

$$\frac{RC(T)}{RC(C)} \leqq 2n. \qquad \square$$

Notice that the optimal network problem used in the proof of Theorem 1 had routing costs which satisfy the triangle inequality. Therefore Theorem 1 also holds if we impose the triangle inequality for the routing costs of the optimal network problem.

With these two theorems we have demonstrated probable lower and upper bounds on the worst-case error ratio for all reasonable polynomial-time heuristics for the optimal network problem with the triangle inequality for all routing costs.

The above results can also be extended to situations in which the required flows $r_{kl}$ are not necessarily equal to one. Suppose that all the $r_{kl}$ are positive integers such that

$$\max_{i,j,k,l} \frac{r_{kl}}{r_{ij}} \leqq n^P \quad \text{for some } P \geqq 3.$$

Then Theorem 1 is modified by changing the worst-case error ratio from $n^{1-\varepsilon}$ to $n^{P-2}$. Theorem 2 is modified by changing the upper bound of $2n$ to $2n^{P+1}$. The proofs of such generalizations are straightforward modifications of the ones given above and will not be given here.

**4. A heuristic for a special case of the optimal network problem.** In this section we consider a special case of the optimal network problem where all construction costs $d_{ij}$ are one. The budget constraint for this type of problem essentially limits the number of arcs allowed in the optimal network design. We will not have to assume that the triangle inequality holds for the routing costs. Johnson, Lenstra, and Rinooy Kan [8] have also shown that this restricted problem is *NP*-complete.

With these new restrictions on the problem, the result of Theorem 1 is no longer valid. We will describe a polynomial-time heuristic $h$ whose worst-case error ratio

$$R_h(n) \leqq 2 \quad \text{for all } n.$$

Let TREE($i$) be the tree network of minimum routing cost paths between node $i$ and every other node in the network. COST($i$) is the sum of the minimum routing costs from node $i$ to every other node in the network.

Our third heuristic can be defined as:

(H3) (1) Find $i$ such that

$$\text{COST}(i) = \min_{j \in D} \text{COST}(j).$$

(2) TREE($i$) is the proposed network configuration.

THEOREM 3. *For optimal network problems having all construction costs equal to one*

$$R_{H3}(n) \leqq 2 \quad \textit{for all } n.$$

*Proof.* We demonstrate this result by proving the stronger fact that

$$\frac{V_{\text{H3}}(s)}{RC(C)} \leqq 2 \quad \text{for all } s,$$

where $V_{\text{H3}}(s)$ is the value of the solution computed by heuristic H3 for optimal network problem $s$ and $RC(C)$ is the routing cost (and solution cost) of the complete network (i.e., the network with all arcs in $A$).

As before define $a_{ij}$ to be the minimum routing cost between nodes $i$ and $j$ on the complete network. Therefore,

$$\text{COST}(i) = \sum_{j \in D} a_{ij}.$$

The routing cost for connecting node $j \neq i$ to all other nodes in the problem using the network TREE($i$) is at most $(n-2)a_{ij} + \text{COST}(i)$.

So

$$V_{H3}(s) \leqq n \cdot \text{COST}(i) + (n-2) \sum_{j \neq i} a_{ij},$$

and

$$V_{\text{H3}}(s) \leqq (2n-2)\,\text{COST}(i).$$

For the complete network we have

$$RC(C) = \sum_{j \in D} \text{COST}(j).$$

Since $\text{COST}(i) \leqq \text{COST}(j)$ for all $j$,

$$RC(C) \geqq n \cdot \text{COST}(i).$$

This implies

$$\frac{V_{\text{H3}}(s)}{RC(C)} \leqq \frac{(2n-2)\,\text{COST}(i)}{n \cdot \text{COST}(i)} \leqq 2. \qquad \square$$

Note that heuristic H3 has a polynomially bounded computation time so that it is possible to have a polynomial-time approximation procedure for a restricted class of

optimal network problems whose worst-case error ratio is bounded by a constant. Theorem 1 shows that is unlikely that such a heuristic exists for a broader class of network design problems.

We believe that combining some local improvement heuristic (perhaps one which added arcs in a "greedy" manner) with H3 could lead to a useful optimal network problem heuristic. It would be necessary to perform additional worst-case analyses or some computational tests in order to verify this conjecture.

**5. Conclusions.** The results of this paper indicate some unusual aspects concerning the complexity of the optimal network problem. Theorem 1 shows that even getting "close" to the optimal solution is an *NP*-complete problem. So, in a sense, this network design problem is more difficult than many other *NP*-complete problems. Similar results of this nature have been developed by Sahni and Gonzalez [15] and Garey and Johnson [6].

Theorem 1 also applies to other discrete network design problems such as the one treated by Leblanc [13] and Morlok and Leblanc [14]. This problem is similar to the optimal network problem except that more complex routing costs and strategies are allowed. So a variety of network design problems appear to be inherently very difficult.

For optimal network problems where the routing costs ($c_{ij}$'s) satisfy the triangle inequality, we have an even stronger result. A strengthened version of Theorem 1 along with Theorem 2, implies that the upper and lower bounds on the worst-case behavior of all reasonable optimal network heuristics (i.e., polynomial-time heuristics that always produce a feasible solution) must be very close unless $P = NP$.

In addition, we explored the relation between various problem parameters and heuristic accuracy. By allowing the required flows ($r_{ij}$'s) to assume different values we were again able to obtain probable (unless $P = NP$) upper and lower bounds on the worst-case behavior of reasonable heuristics. We also saw that by restricting all the construction costs ($d_{ij}$'s) to be equal, it is then possible to find heuristics whose worst-case error is bounded by a constant independent of problem size.

Although most optimal network heuristics probably have a bad worst-case error, there may be some heuristics whose "average" case behavior is quite good. In § 2 we saw that heuristics used by Dionne and Florian [3] can be very inaccurate in terms of worst-case error even though computational tests have indicated that their relative margins of error are usually quite small. Many heuristics, especially ones for complicated real world problems (such as telephone network optimization), also appear to behave in a similar way. An interesting area of future work would be to explore probabilistic analyses of optimal network heuristics. See Karp [11], [12] for some examples of probabilistic analyses for various combinatorial problems.

**Acknowledgments.** I am indebted to Professor Thomas L. Magnanti of MIT for his encouragement and suggestions concerning this paper. Paulo Villela of MIT also provided useful comments. This paper was originally presented at the ORSA/TIMS Conference, New York, May, 1978.

REFERENCES

[1] D. E. BOYCE, A. FARHI, AND R. WEISCHEDEL, *Optimal network problem: A branch-and-bound algorithm*, Environment and Planning, 5 (1973), pp. 519–533.

[2] G. CORNUEJOLS, M. L. FISHER, AND G. L. NEMHAUSER, *An analysis of heuristics and relaxations for the uncapacitated location problem*, Management Sci., 23 (1977), pp. 789–810.

[3] R. DIONNE, *Une analyse théorique et numérique du problème du choix optimal d'un réseau de transport sans congestion*, Publication No. 198, Départment d'informatique, Université de Montreal, October 1974.

[4] R. DIONNE AND M. FLORIAN. *Exact and Approximate Algorithms for Optimal Network Design*, Publication No. 41, Centre de Recherche Sur les Transports, Université de Montreal, 1977.

[5] M. R. GAREY AND D. S. JOHNSON, *The Complexity of Near-Optimal Graph Coloring*, J. Assoc. Comput. Mach., 23 (1976), pp. 43–49.

[6] ———, *Approximation algorithms for combinatorial problems: An annotated bibliography*, Algorithms and Complexity, J. F. Traub, ed., Academic Press, New York, 1976.

[7] H. H. HOANG, *A Computational Approach to the Selection of an Optimal Network*, Management Sci., 19 (1973) pp. 488–498.

[8] D. S. JOHNSON, J. K. LENSTRA, AND A. H. G. RINNOOY KAN, *The Complexity of the Network Design Problem*, Networks, to appear.

[9] R. M. KARP, *Reducibility among combinatorial problems*, Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, pp. 85–104.

[10] ———, *On the computational complexity of combinatorial problems*, Networks, 5 (1975), pp. 45–68.

[11] ———, *The Probabilistic Analysis of Some Combinatorial Search Algorithms*, Algorithms and Complexity, J. F. Traub, ed., Academic Press, New York, 1976.

[12] ———, *Probabilistic analyses of partitioning algorithms for the traveling salesman problem in the plane*, Math. of Operations Res., 2 (1977), pp. 209–224.

[13] L. J. LEBLANC, *An Algorithm for the Discrete Network Design Problem*, Trans. Sci., 9, (1975), No. 3, pp. 283–287.

[14] E. K. MORLOK AND L. J. LEBLANC, *A marginal analysis technique for determining improvements to an urban road network*, Paper presented at ORSA/TIMS National Meeting, Las Vegas, NV, Fall 1975.

[15] S. SAHNI AND T. GONZALEZ, *P-Complete approximation problems*, J. Assoc. Comput. Mach., 23 (1976), pp. 555–565.

[16] A. J. SCOTT, *The optimal network problem: Some computational procedures*, Trans. Res., 3 (1969), pp. 201–210.

[17] R. WONG, *A survey of network design problems*, Working Paper OR 080–78, M.I.T. Operations Research Center, Cambridge, MA, 1978.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0009 $01.00/0

# A PROOF OF TUTTE'S TRINITY THEOREM AND A NEW DETERMINANT FORMULA*

KENNETH A. BERMAN†

**Abstract.** A new proof of Tutte's trinity theorem (Proc. Cambridge Phil. Soc., 1948), (North-Holland, 1973) is presented. The proof is based on a new determinant formula for the number of spanning arborescences of a digraph. This formula generalizes the determinant formula given by Maurer (SIAM J. Appl. Math., 1976) for the number of spanning trees of an undirected graph.

**1. Introduction.** In his paper "The dissection of equilateral triangles into equilateral triangles", Tutte generalizes the concept of dual planar maps to a trinity of alternating planar dimaps [7]. This concept is further developed by Tutte in [8]. An *alternating dimap* $A$ is a planar Eulerian map which is oriented so that the edges around each vertex are directed alternately into and out of that vertex. Such a map has the property that the number of spanning in arborescences rooted at any vertex equals the number of spanning out arborescences rooted at that vertex. Further, the number of spanning out arborescences rooted at every vertex is the same [7]. The *arborescence number* of $A$, denoted by $a(A)$, is defined to be this common number. Tutte proves in his papers [7], [8] that each dimap of a trinity of alternating dimaps has the same arborescence number. These proofs are based on graph theoretical considerations. In this paper an alternate approach employing determinants is given.

The paper is in two parts. The first part, which includes §§ 2 and 3, deals with trinity and gives a proof of Tutte's trinity theorem employing determinants. In the second part, §§ 4 and 5, a new determinant formula for enumerating spanning arborescences is developed.

In § 2, trinity is defined and it is shown how trinity generalizes duality. Two approaches are discussed. In § 3, Tutte's trinity theorem is deduced from a determinant formula for the arborescence number of an alternating dimap. Following a discussion of some results from the literature in § 4 the enumeration of spanning arborescences of a digraph is considered in § 5 and a new determinant formula is obtained. A special case of this formula was used in § 3 to derive Tutte's trinity theorem and another special case is Maurer's determinant formula [5] for the number of spanning trees of an undirected graph.

**2. Trinity.** Consider an alternating dimap $A$ with edge set $E$, vertex set $V$ and face set $F$. Let $P$ denote the set of faces of $A$ directed counterclockwise and let $N$ denote the set of faces directed clockwise. We associate with dimap $A$ two alternating dimaps $A_P$ and $A_N$ as follows. In the interior of each face of $P$ place a vertex. These vertices which we denote by $V_P$ will be the vertex set of $A_P$. Consider any vertex $v \in V$. Let $f_1^v, f_2^v, \cdots, f_k^v$ denote the faces from $P$ which are incident with $v$ taken clockwise around $v$. Let $u_1, u_2, \cdots, u_k$ denote the vertices from $V_P$ contained in the interior of faces $f_1^v, f_2^v, \cdots, f_k^v$ respectively. Join vertices $u_i$ and $u_{i+1}$ with an edge directed toward $u_{i+1}$, $i = 1, 2, \cdots, k$ where $u_{k+1} = u_1$. Repeat this for every vertex $v \in V$. The resultant dimap $A_P$ is an alternating dimap. By an analogous construction using the set $N$ in place of $P$ we obtain the alternating dimap $A_N$. Dimaps $A_P$ and $A_N$ are the *derived alternating dimaps* of $A$.

* Received by the editors December 5, 1978.

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

It can be verified that the derived alternating dimaps of $A_P$ are $A_N$ and $A$ and the derived alternating dimaps of $A_N$ are $A$ and $A_P$. We call $A$, $A_P$, $A_N$ a *trinity of alternating dimaps* and speak of the property of *trinity* for alternating dimaps in analogy with the property of "duality" for undirected maps [7], [8].

Trinity is a true generalization of duality. To see this let $M$ be a planar map with dual map $M'$. Associated with the map $M$ is the alternating dimap $\hat{M}$ defined as follows. Double each edge of $M$ such that the bivalent face created contains no edges in its interior and orient all the edges such that each of these bivalent faces is directed clockwise. The resultant map is $\hat{M}$. Apply the same construction on the dual map $M'$ to obtain dimap $\hat{M}'$ except direct the new bivalent faces counterclockwise. The derived alternating dimap of $\hat{M}$ associated with the faces directed counterclockwise is the dimap $\hat{M}'$ and the derived alternating dimap of $\hat{M}$ associated with the faces directed clockwise is a 4-valent map $\bar{M}$ known as the medial map of $M$ and $M'$. Thus $\hat{M}$, $\hat{M}'$, $\bar{M}$ form a trinity of alternating dimaps. This shows that duality is a special case of trinity.

Tutte in [8] gives the following approach to trinity. Let $T$ be a plane Eulerian triangulation whose vertices have been colored with colors 1, 2, 3. Let $V_1$, $V_2$, $V_3$ be the sets of vertices colored 1, 2, 3, respectively. We now construct three alternating dimaps $A_1$, $A_2$, $A_3$ having vertex sets $V_1$, $V_2$, $V_3$.

We construct $A_1$ as follows. (See Fig. 1 taken from [8]). Bi-color the triangle faces of $T$ black and white such that the outside face is colored white. This can be done since $T$ is Eulerian. For $x \in V_1$ and $f_b$ a black triangle incident with $x$ draw a directed edge of $A_1$ along a median of $f_b$ to the midpoint of the opposite side. Continue the edge along a median of the adjacent white face $f_w$ to the opposite vertex $y$. It is clear that $y \in V_1$. (In the case $x$ and $y$ are coincident the edge drawn is a loop taken with the specified sense of description.) Thus we have the dimap $A_1$ and by a similar construction using vertex sets $V_2$ and $V_3$ in place of $V_1$ we obtain dimaps $A_2$ and $A_3$.

We now define mappings $\sigma_{12}$, $\sigma_{23}$, $\sigma_{31}$. Let $E_1$, $E_2$, $E_3$ be the edge sets of $A_1$, $A_2$, $A_3$. Consider the mapping $\sigma_{12}$: $E_1 \to E_2$ defined as follows. An edge $e \in E_1$ intersects two edges from $E_2$, one at the center of a black triangle and one at the center of a

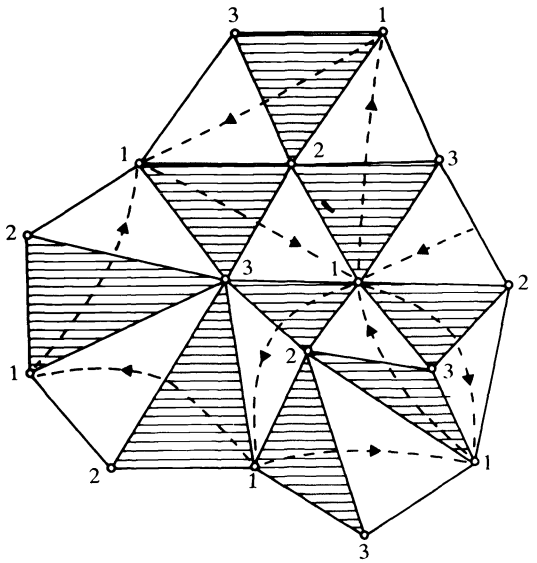



FIG. 1. (*Tutte* [8].)

white triangle. Let $\sigma_{12}(e)$ be the edge from $E_2$ which intersects $e$ at the center of a white triangle. Similarly define functions $\sigma_{23}$, $\sigma_{31}$.

*The three dimaps $A_1$, $A_2$, $A_3$ form a trinity of alternating dimaps [8]. Conversely, given any trinity of alternating dimaps there exists a 3-colored plane Eulerian triangulation which yields them by the above construction.*

**3. Tutte's trinity theorem.** In this section, a new proof of Tutte's trinity theorem [7], [8] is given. The proof employs a new determinant formula for the number of spanning arborescences.

THEOREM 3.1 (Tutte). *Each dimap of a trinity $A_1, A_2, A_3$ of alternating dimaps has the same arborescence number, that is*

$$a(A_1) = a(A_2) = a(A_3).$$

We deduce this theorem from the following determinant formula for the arborescence number. This formula is a special case of a more general formula derived in § 5, Theorem 5.1.

THEOREM 3.2. *Let $A$ be an alternating dimap with edge set $E$, vertex set $V$ and face set $F$ of cardinalities, $m$, $n$, $l$ respectively. Let $\mathbb{V}$ be the $n \times m$ matrix such that*

$$\mathbb{V}(v, e) = \begin{cases} 1, & e \text{ directed into } v, \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

*$v \in V$, $e \in E$ and let $\mathbb{F}$ be the $l \times m$ matrix such that*

$$\mathbb{F}(f, e) = \begin{cases} 1, & e \text{ belongs to } f, \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

*$f \in F$, $e \in E$. For $v \in V$ and $f \in F$ let $\mathbb{V}_v$ be the matrix obtained from $\mathbb{V}$ by deleting the row corresponding to vertex $v$ and let $\mathbb{F}_f$ be the matrix obtained from $\mathbb{F}$ by deleting the row corresponding to face $f$. Then the arborescence number $a(A)$ of $A$ is given by*

$$a(A) = \pm \det \begin{pmatrix} \mathbb{V}_v \\ \cdots\cdots \\ \mathbb{F}_f \end{pmatrix}_{m \times m}. \tag{3.3}$$

We now prove Theorem 3.1. Let $A_1, A_2, A_3$ be a trinity of alternating dimaps with vertex sets $V_1$, $V_2$, $V_3$, and face sets $F_1$, $F_2$, $F_3$. Let $P_1$, $P_2$, $P_3$ denote the sets of faces directed counterclockwise and let $N_1$, $N_2$, $N_3$ denote the sets of faces directed clockwise.

Consider the function $\sigma_{12}$ defined at the end of § 2. Function $\sigma_{12}$ maps, (i) the set of edges around a face from $P_1$ onto the set of edges directed into a vertex of $V_2$, (ii) the set of edges around a face from $N_1$ onto the set of edges around a face from $P_2$, (iii) the set of edges directed into a vertex of $V_1$ onto the set of edges around a face from $N_2$. Let $\mathbb{V}_1$, $\mathbb{V}_2$ be the matrices corresponding to dimaps $A_1$ and $A_2$ defined by (3.1) and let $\mathbb{F}_1$, $\mathbb{F}_2$ be the matrices corresponding to dimaps $A_1$ and $A_2$ defined by (3.2). It follows that the rows of the matrix

$$\begin{pmatrix} \mathbb{V}_1 \\ \cdots\cdots \\ \mathbb{F}_1 \end{pmatrix}$$

are a permutation of the rows of the matrix

$$\begin{pmatrix} \mathbb{V}_2 \\ \cdots\cdots \\ \mathbb{F}_2 \end{pmatrix}.$$

By Theorem 3.2 this implies that dimaps $A_1$ and $A_2$ have the same arborescence number. Similarly, dimaps $A_2$ and $A_3$ have the same arborescence number. This proves Theorem 3.1.

**4. Results from the literature.** Let $G$ be an undirected graph with edge set $E$ and vertex set $V$. Let $D$ be a digraph obtained by directing the edges of $G$. A 1-*chain* over a ring $R$ is a mapping from $E$ into the elements of $R$. In this paper we take $R$ to be the field of real numbers. Let $C$ denote the set of 1-chains. For $c_1, c_2 \in C$ and $\lambda \in R$ we define the operation $+$ and scalar multiplication $\cdot$ as follows:

$$(c_1 + c_2)(e) = c_1(e) + c_2(e) \tag{4.1}$$

$$(\lambda \cdot c_1)(e) = \lambda(c_1(e)) \tag{4.2}$$

for all $e \in E$. The system $(C, +, \cdot)$ is a vector space over $R$.

For $v \in V$ let $\delta^+(v)$ and $\delta^-(v)$ denote the set of edges directed into $v$ and out of $v$ respectively. A *cycle* over $R$ is a 1-chain $z$ such that

$$\sum_{e^+ \in \delta^+(v)} z(e^+) = \sum_{e^- \in \delta^-(v)} z(e^-) \tag{4.3}$$

for each vertex $v \in V$. A *coboundary* over $R$ is a 1-chain $b$ such that

$$b(e) = \phi(h(e)) - \phi(t(e)) \tag{4.4}$$

for all $e \in E$ where $\phi$ is a mapping from $V$ into $R$ and $h(e)$, $t(e)$ denote the head and tail of $e$ respectively. Let $Z$ denote the set of cycles and let $B$ denote the set of coboundaries. It is immediate that $Z$ and $B$ are subspaces of $C$ and it is easily verified that they are orthogonal.

For $v \in V$ define 1-chains $\delta_v^+$, $\delta_v^-$, $\delta_v$ as follows.

$$\delta_v^+(e) = \begin{cases} 1, & v = h(e), \\ 0, & \text{otherwise}, \end{cases} \tag{4.5}$$

$$\delta_v^-(e) = \begin{cases} 1, & v = t(e), \\ 0, & \text{otherwise}, \end{cases} \tag{4.6}$$

$$\delta_v(e) = \delta_v^+(e) - \delta_v^-(e) \tag{4.7}$$

for $e \in E$. Set

$$\Delta_v^+ = \{\delta_u^+ \mid u \in V - v\}, \tag{4.8}$$

$$\Delta_v = \{\delta_u \mid u \in V - v\}. \tag{4.9}$$

It is immediate that $\delta_v$ is a coboundary and that $\Delta_v$ is a basis for the coboundary space $B$. Let $\mathbb{B}_v$ be the matrix whose rows correspond to the vectors in $\Delta_v$.

A matrix is *unimodular* if all its full square submatrices have determinants 0, +1, −1. A *unimodular cycle matrix* $\mathbb{Z}$ is a unimodular matrix whose rows correspond to a basis of the cycle space $Z$. There is a unimodular cycle matrix associated with every spanning tree of $G$ [1]. Note that the matrix $\mathbb{F}_f$ defined by (3.2) is a unimodular cycle matrix.

The following two determinant formulas for the number $t(G)$ of spanning trees of $G$ are well-known [1], [5], [6]. (The transpose of a matrix $M$ will be denoted by $M'$.)

THEOREM 4.1 (Kirchhoff–Trent). $t(G) = \det(\mathbb{B}_v \mathbb{B}_v')$.

THEOREM 4.2. $t(G) = \det(\mathbb{Z}\ \mathbb{Z}')$.

A third formula is given in Maurer [5].

THEOREM 4.3 (Maurer).

$$t(G) = \pm\det \begin{pmatrix} \mathbb{B}_v \\ \cdots \\ \mathbb{Z} \end{pmatrix}.$$

Theorem 4.1 has been extended to spanning arborescences of a digraph $D$ by Tutte [7]. This can be formulated as follows. Let $\mathbb{B}_v^+$ be the matrix whose rows correspond to the vectors in $\Delta_v^+$. (Note that $\mathbb{V}_v = \mathbb{B}_v^+$ where $\mathbb{V}_v$ is defined in Theorem 3.2). Let $a_v(D)$ denote the number of spanning out arborescences rooted at vertex $v$ of $D$.

THEOREM 4.4 (Tutte). $a_v(D) = \det(\mathbb{B}_v^+ \ \mathbb{B}_v')$.

Theorem 4.1 is the matrix-tree theorem and Theorem 4.4 is the matrix-arborescence theorem.

**5. Enumeration of spanning arborescences.** In this section the following new determinant formula for enumerating spanning arborescences is derived.

THEOREM 5.1. *The number $a_v(D)$ of spanning out arborescences rooted at a vertex $v$ of a digraph $D$ is given by*

$$a_v(D) = \pm\det \begin{pmatrix} \mathbb{B}_v^+ \\ \cdots \\ \mathbb{Z} \end{pmatrix}, \tag{5.1}$$

*where $\mathbb{B}_v^+$ is the matrix whose rows correspond to the vectors in $\Delta_v^+$ and $\mathbb{Z}$ is a unimodular cycle matrix.*

Theorem 5.1 is an immediate corollary of the following stronger result. Associate the indeterminate weights $x_1, x_2, \cdots, x_m$ with the edges $e_1, e_2, \cdots, e_m$ of $D$. The *weight* of a spanning arborescence of $D$ is defined to be the product of the weights on the edges of the arborescence.

THEOREM 5.2. *Let $D$ be a digraph with the indeterminate weights $x_1, x_2, \cdots, x_m$ associated with the edges $e_1, e_2, \cdots, e_m$ respectively. The polynomial $A_v(x_1, x_2, \cdots, x_m)$ which is the sum of the weights over all the spanning out arborescences rooted at vertex $v$ of $D$ is given by*

$$A_v(x_1, x_2, \cdots, x_m) = \pm\det \begin{pmatrix} \mathbb{B}_v^+\mathbb{X} \\ \cdots \\ \mathbb{Z} \end{pmatrix}, \tag{5.2}$$

*where $\mathbb{B}_v^+, \mathbb{Z}$ are defined as in Theorem 5.1 and $\mathbb{X}$ is the diagonal matrix $(x_1, x_2, \cdots, x_m)$.*

To prove Theorem 5.2 we employ an extension of Theorem 4.4 to weighted digraphs. This result which is sometimes called the Bott–Mayberry theorem [2] may be formulated as follows.

$$A_v(x_1, x_2, \cdots, x_m) = \det(\mathbb{B}_v^+ \ \mathbb{X} \ \mathbb{B}_v'). \tag{5.3}$$

Now let $t$ denote the number of spanning trees of $D$ (of the underlying graph $G$). Then by Theorem 4.3

$$t \det \begin{pmatrix} \mathbb{B}_v^+\mathbb{X} \\ \cdots \\ \mathbb{Z} \end{pmatrix} = \pm\det \begin{pmatrix} \mathbb{B}_v^+\mathbb{X} \\ \cdots \\ \mathbb{Z} \end{pmatrix} \det \begin{pmatrix} \mathbb{B}_v' \vdots \ \mathbb{Z}' \end{pmatrix}.$$

Since the cycle space $Z$ and the coboundary space $B$ are orthogonal, $\mathbb{Z}\mathbb{B}_v' = 0$ and we have

$$t \det \begin{pmatrix} \mathbb{B}_v^+\mathbb{X} \\ \cdots \\ \mathbb{Z} \end{pmatrix} = \pm\det \begin{pmatrix} \mathbb{B}_v^+\mathbb{X}\mathbb{B}_v' \vdots \ \mathbb{B}_v^+\mathbb{X}\mathbb{Z}' \\ \cdots\cdots\cdots\cdots \\ 0 \ \vdots \ \mathbb{Z}\mathbb{Z}' \end{pmatrix}$$

$$= \pm\det(\mathbb{B}_v^+\mathbb{X}\mathbb{B}_v') \det(\mathbb{Z}\mathbb{Z}').$$

Employing Theorem 4.2 and (5.3) gives (5.2) of Theorem 5.2.

Theorem 3.2 used in the proof of Tutte's trinity theorem is obtained as a special case of Theorem 5.1 with the observation that $\mathbb{V}_v = \mathbb{B}_v^+$ and $\mathbb{F}_f$ is a unimodular cycle matrix.

Maurer's formula, Theorem 4.3 is also a special case of Theorem 5.1. To see this consider the digraph $\hat{G}$ obtained by replacing each edge of $G$ with two edges directed in opposite directions. It is immediate that

$$t(G) = a_v(\hat{G})$$

for all vertices $v$ of $\hat{G}$. It can be shown by simple manipulations that the determinant for $\hat{G}$ on the right-hand side of (5.1) is equal to the determinant for $G$ on the right-hand side of the equation of Theorem 4.3. Thus Maurer's formula follows from Theorem 5.1.

## REFERENCES

[1] J. A. BONDY AND U. S. R. MURTY, *Graph Theory and Applications*, American Elsevier, New York, 1976, pp. 218–219.

[2] R. BOTT AND J. P. MAYBERRY, *Matrices and trees, Economic Activity Analysis*, Wiley, New York, 1954, pp. 391–400.

[3] R. L. BROOKS, C. A. B. SMITH, A. H. STONE AND W. T. TUTTE, *The dissection of rectangles into squares*, Duke Math. J. (1940), pp. 312–340.

[4] R. L. BROOKS, C. A. B. SMITH, A. H. STONE AND W. T. TUTTE, *Leaky electricity and triangulated triangles*, Philips Res. Rep., 30 (1975), pp. 205–219.

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

[6] H. TRENT, *A note on the enumeration and listing of all possible trees in a connected linear graph*, Proc. Nat. Acad. Sci. U.S.A., 40 (1954), pp. 1004–1007.

[7] W. T. TUTTE, *The dissection of equilateral triangles into equilateral triangles*, Proc. Cambridge Phil. Soc., 44 (1948), pp. 463–482.

[8] ———, *Duality and trinity*, Infinite and Finite Sets (Colloq. M. Soc. J. Bolyai, 10, Keszthely, Hungary, 1973), vol. 3, North-Holland, Amsterdam, 1975, pp. 1459–1472.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0010 $01.00/0

# RECTILINEAR STEINER TREES IN RECTANGLE TREES*

ARTHUR M. FARLEY†, STEPHEN T. HEDETNIEMI† AND SANDRA L. MITCHELL†

**Abstract.** The rectilinear Steiner tree (RST) problem is known to be *NP*-complete for an arbitrary set of points in the plane. In this paper we extend the known cases for which solutions can be determined in linear time to include sets of points which generate types of minimum-distance rectangle trees. Rectangle trees are rectilinear, plane $C_4$-trees. A minimum-distance rectangle tree is a rectangle tree such that the length of a shortest path between any two vertices in the graph is equal to the rectilinear distance (in the plane) between the two vertices. A complete enumeration of minimum-distance rectangle trees is included. The results have application in the design of terminal patterns and wiring layouts for electronic circuitry.

**Introduction.** A *rectilinear Steiner tree* (RST) for a set $A$ of points in the plane is a tree which contains all the points of $A$ and which contains only horizontal and vertical edges. An *optimal RST* for a set of points is an RST in which the edges have minimum total length. Although the problem of determining an optimal RST for an arbitrary set of points in the plane has been shown to be *NP*-complete [2], an efficient (polynomial time) algorithm has been designed [5] which determines good approximate solutions by constructing rectilinear spanning trees. Furthermore, efficient algorithms which provide optimal solutions in several special cases have been determined [1], [3]. In particular, an $O(n)$ algorithm exists which determines an optimal RST for a set of points forming a subset of the grid points of a $2\times n$ grid [1]. In this paper we extend this result to include sets of points which generate certain classes of rectangle trees.

**1. Definitions.** A *rectilinear plane graph* is a plane graph [4] having only horizontal and/or vertical edges. A *rectangle tree* (RT) is a rectilinear plane graph which can be constructed from an initial rectangle by a finite number of applications of the following operation:

Add a new rectangle, identifying one of its edges with an edge $e$ on the exterior face of the existing graph such that neither vertex of edge $e$ had degree 4.

Rectangle trees form a planar subclass of the class of $C_4$-trees, where a $C_n$-tree is either a cycle $C_n$ with $n$ vertices or a graph obtained by identifying an edge of a new $C_n$ with an edge of an existing $C_n$-tree [6]. If each cycle were considered to be a vertex and its neighbors to be those cycles with which it shared an edge, the resultant graph would be a tree. Several examples of RTs are presented in Fig. 1. A *ladder* is a single edge or an RT containing no vertices of degree 4. An $L$ is an RT containing exactly one vertex of degree 4. A *cross* is an RT containing exactly four vertices of degree 4 which form a $C_4$. A $T$ is an RT containing exactly two vertices of degree 4 which are also adjacent.

In the results that follow we will need to refer to the various components of a given rectangle tree. They are defined as follows. An *end-edge* is an edge both vertices of which have degree 2; a rectangle containing an end-edge is an *end-rectangle*. A *maximal induced ladder* of an RT is a ladder subgraph of the RT such that the ladder is not a subgraph of another ladder contained in the RT. An *arm* of an RT is a maximal induced ladder of the RT. Figure 2 illustrates the arms of one RT. A *branch* of an RT is a maximal induced ladder of the RT containing an end-edge $(a', b')$ but no vertices of degree 4. Every branch of an RT such that the RT itself is not a ladder is attached to the rest of the RT at two adjacent vertices $r_a$, $r_b$, at least one of which has degree 4. We refer

* Received by the editors May 17, 1978, and in revised form March 16, 1979.
† Computer Science Department, University of Oregon, Eugene, Oregon 97403.

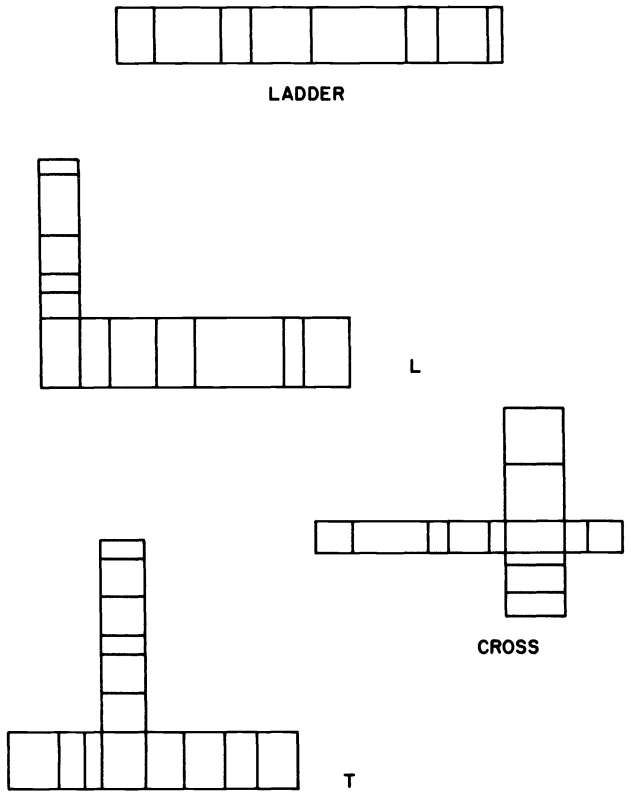



FIG. 1

to the edge $(r_a, r_b)$ as the *remote edge* of the branch; the edge of the ladder $(a, b)$ which forms a rectangle with $(r_a, r_b)$ is called the *initial edge* of the ladder. Figure 3 illustrates the branches of one RT.

Finally, a set $A$ of points in the plane *generates* an RT, denoted RT($A$), if there exists an RT which can be constructed according to the definition above such that (1) at least two nonadjacent vertices of the initial rectangle coincide with points in $A$, and (2) at least one vertex of each new end-edge and at least one vertex of the existing edge

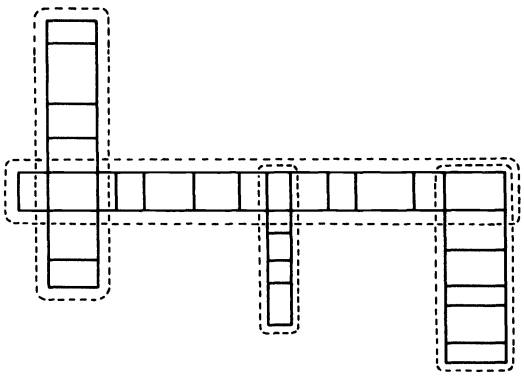


FIG. 2

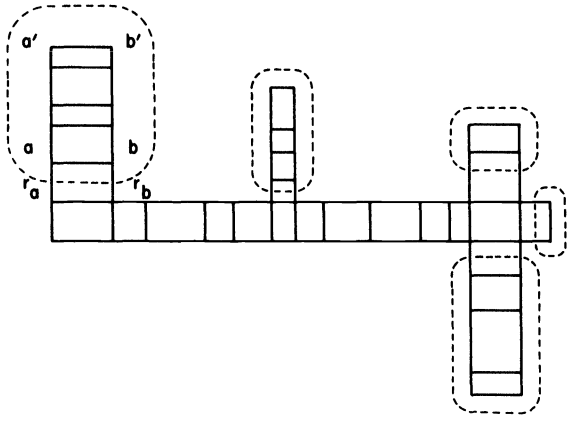



FIG. 3

which is identified with an edge of the new rectangle are in $A$. Figure 4 demonstrates that not every set of points in the plane generates an RT.

The problem of determining whether a set of points generates an RT is likely to be NP-complete in the general case. The problem can be solved in linear time for certain special cases as is shown later. A necessary property of a set of points which generates an RT follows from the definitions above and the following one. A point $p$ of a set of points $A$ is *rectilinearly isolated* if there are no other points of $A$ on the horizontal and vertical lines through $p$. If $A$ contains more than two points and $A$ generates an RT, then at most one point of $A$ is rectilinearly isolated. Unfortunately, this is not a sufficient condition for a set of points to generate an RT.

**2. Separable RTs.** Let $l(e)$ denote the length of edge $e$, and let the width of a branch $B$ with end-edge $(a', b')$ be $w(B) = l(a', b')$. A *separable RT* is an RT such that for every branch $B$, $w(B) \leqq l(a, r_a)$ (i.e. the length of the edge joining the remote vertex $r_a$ and the initial vertex $a$ is not less than the width of the branch), and if one removes branch $B$ from the RT then the remaining RT is separable. A set $A$ of points generates a separable RT if there exists an RT($A$) which can be constructed such that the length of the edge of the new rectangle which is identified with an existing edge is not greater than the length of the edge between this edge and the new end-edge. This condition is sufficient, but not necessary, for the resulting RT to be separable.

An *RT-optimal RST* for a set of points $A$ which generates a rectangle tree RT($A$) is an optimal RST for $A$ which is a subgraph of RT($A$). In other words, an RT-optimal RST is a solution to the rectilinear Steiner tree problem for the subset of vertices $A$ within the graph RT($A$). Our interest in separable RTs stems from the fact that whenever a set $A$ of points generates a separable RT, we can solve the RST problem for RT($A$) in linear time.

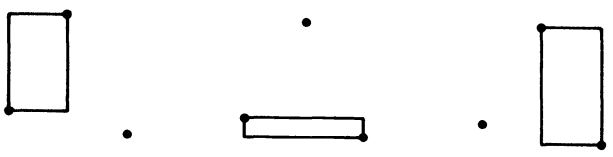


FIG. 4

**3. A linear algorithm for obtaining an RT-optimal RST in a separable RT.** The following Algorithm RT repeatedly determines an RT-optimal RST for each branch of a given separable RT($A$). In the case of a branch which is a single edge, the solution is trivial. For a nontrivial branch, Algorithm RT uses the dynamic programming algorithm of Aho, Garey and Hwang [1] to determine an RT-optimal RST for branch $B$, denoted $S(B)$. After each branch solution is determined, it is joined to a remote vertex and the branch is pruned from the RT. The algorithm halts when a branch has no remote edge (i.e. the RT has been reduced to a ladder). After application of Algorithm RT, the variable ORST has as its value the set of edges constituting an RT-optimal RST for the set $A$ in RT($A$). Algorithm RT is formally defined as follows:

```
ALGORITHM RT
[Initialize]      set G ← RT(A)
                  set ORST ← ∅
[Each branch]     while there exists a branch B in G
                  do [find RST of branch]
                      if B consists of the single edge (a', b')
                          then [both vertices in A?]
                              if a' ∈ A and b' ∈ A
                                  then set ORST ← ORST ∪ {(a', b')}
                              fi
                          else use Aho–Garey–Hwang Algorithm to obtain S(B)
                              set ORST ← ORST ∪ S(B)
                      fi
                      [join to solution for entire tree]
                      if remote edge (r_a, r_b) of B is null
                          then STOP
                          else [join a or b to the remote edge]
                          if only one initial vertex, say a, is an element of A
                              then set ORST ← ORST ∪ {(a, r_a)}
                                  set A ← A ∪ {r_a}
                              else if r_a ∈ A then set ORST ← ORST ∪ {(a, r_a)}
                                          else set ORST ← ORST ∪ {(b, r_b)}
                                  fi
                          fi
                      fi
                      [delete branch from the graph]
                      set G ← G − (B ∪ {(a, r_a), (b, r_b)})
                  od
```

THEOREM 1. *Algorithm RT determines an RT-optimal RST for any set $A$ of points which generates a separable RT, RT($A$). Algorithm RT can be implemented so as to require only $O(n)$ computational steps, where $n$ is the number of points in $A$.*

*Proof.* The correctness of Algorithm RT is readily established. The Aho–Garey–Hwang algorithm has been shown to determine an RT-optimal RST for the set of points in any nontrivial branch. The solution for each branch is joined in a manner which minimizes the length added to the overall solution. The edge which is added is of minimum length and it is added so that no unnecessary edge is added to the rest of the solution. Since RT($A$) is separable, joining the branch solution to the overall solution by a single edge is a globally optimal decision. Any overall solution which includes two edges connecting a branch to the overall solution can be replaced by one using only one such edge and an edge traversing the width of the branch. Since the width of a branch is

less than or equal the length of an edge connecting a branch to a remote vertex, this new solution will have overall length less than or equal to that of the original. The RT $G$ is finally reduced to a ladder with a null remote edge, which leads to the completion of the algorithm's execution.

The Aho–Garey–Hwang algorithm determines a branch solution in time (computational steps) linearly related to the number of points of $A$ in the branch. The joining procedure requires a constant amount of time per branch. It only remains to show that the total time spent finding branches is also $O(n)$. The branches, which are identified by end-edges, can be found by making an initial $O(n)$ pass over the vertices and edges in RT($A$) and forming a list of all adjacent pairs of vertices of degree 2. This list can be updated in constant time as each branch is deleted. Thus, Algorithm RT has $O(n)$ time complexity. □

Algorithm RT produces approximate solutions for nonseparable RTs. In a nonseparable RT, the decision to join a branch solution by a single edge may not be globally optimal. A better solution may use two edges to join the branch solution while eliminating an edge which crosses the branch. The solution produced by Algorithm RT exceeds the optimal solution in length by less than the sum of the widths of the solved branches. The performance of Algorithm RT on nonseparable RT can be improved by the following modifications.

(1) as each branch solution is joined to the remote edge, mark the remote edge "special" if there are elements of $A$ on both sides of the branch whose distance to the remote edge is less than the width of the branch;

(2) after each branch solution is determined, for each edge which is marked "special" and is a member of that solution eliminate the nearest parallel edge in the overall solution which lies in the branch adjacent to the special edge and add edges to reconnect the disconnected side of that branch to its remote vertex.

Figure 5 illustrates this adjustment procedure, where "$X$" marks the special edge. In the case of a nonseparable L, T, or cross, all special edges will be elements of the same rectangle of the RT. As such, Algorithm RT can be further modified to determine the overall RT-optimal RST for such a graph while maintaining the linear time bound.

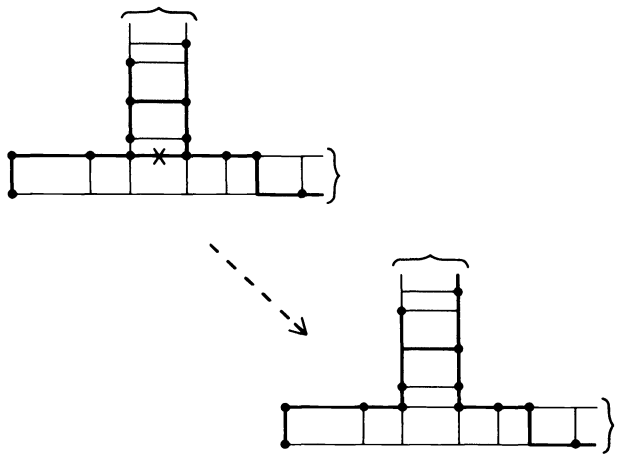


FIG. 5

**4. Minimum-distance RTs.** A *minimum-distance RT* is an RT such that the length of a shortest path in the RT between any two vertices of the RT is equal to the rectilinear distance (in the plane) between the two vertices.

THEOREM 2. *Any ladder, L, T, or cross is a minimum-distance RT.*

*Proof.* Let $G$ be such a graph and $u$ and $v$ be two vertices of $G$. If a path can be found between $u$ and $v$ which contains at most one right-angle turn, then that path has length equal to the rectilinear distance between the two vertices. It is easy to see that any two vertices in a ladder, $L$, $T$, or a cross are either joined by a straight path or a path with exactly one right-angle turn. □

Let $A$ be a set of points in the plane. Consider the rectangular grid formed by drawing a horizontal and a vertical line through each point. Let $G(A)$ be the plane graph denoting that portion of this grid enclosed by the largest rectangle formed by the grid lines. Consider the intersection of a horizontal and vertical grid line to be a vertex and a line segment connecting two such vertices to an edge.

Let $A$ be a set of points generating an RT such that RT($A$) is a minimum-distance RT. We want to show that there exists an optimal RST (in the plane) for $A$ which consists solely of vertices and edges of RT($A$). To accomplish this we present the following results. The first is immediate from the definitions, the second is due to Hanan [3].

LEMMA 3.1. RT($A$) *is a subgraph of* $G(A)$.

LEMMA 3.2. *There is an optimal RST for A which is contained in* $G(A)$.

Let $R$ denote an optimal RST for $A$ contained in $G(A)$. Let the *interior* of RT($A$) be the vertices and edges of RT($A$). Let the *exterior* of RT($A$) be the rest of $G(A)$. Let $R_{ext}$ be that part of $R$ lying in the exterior of RT($A$).

The following two lemmas relate points or paths in the exterior of RT($A$) to the interior of RT($A$).

LEMMA 3.3. *A straight path in* $R_{ext}$ *has at most one end which is a vertex of* RT($A$).

*Proof.* Since RT($A$) is a minimum distance RT, any straight path whose ends are both vertices of RT($A$) must be in RT($A$). □

LEMMA 3.4. *Vertices in the interior of* RT($A$) *lie in only two of the four possible rectilinear directions from any vertex in the exterior of* RT($A$). *Furthermore, one direction is horizontal (left or right) and the other is vertical (up or down).*

*Proof.* Since a vertex $v$ in the exterior is within $G(A)$, points of $A$ lie in at least one vertical and one horizontal direction from $v$. If points on the interior were to lie in opposite directions from $v$, there would exist a path through $v$ which would interconnect two vertices of RT($A$) such that this path would have length less than that of the shortest path connecting them in RT($A$). This would contradict the assumption that RT($A$) is a minimum-distance RT. □

Lemma 3.4 implies that each connected subgraph of the exterior of RT($A$) contains a corner vertex of the rectangle which bounds the exterior face of $G(A)$. Therefore, there can be at most four such graphs in the exterior of RT($A$). Furthermore, a subgraph contains edges on two sides of the rectangle bounding the exterior face of $G(A)$. The specification of an operation which will eliminate all edges and vertices of $R_{ext}$ while maintaining an optimal RST for the set $A$ will complete preparation for our theorem.

A *segment* of $R_{ext}$ is a maximal straight path in $R_{ext}$. The operation we shall employ is the *segment shift.* The segment shift operation is applicable to a segment of $R_{ext}$ nearest the exterior boundary of $G(A)$. The operation moves the segment so that it lies on the adjacent parallel grid line of $G(A)$ which is nearer the interior of RT($A$). The shifted segment can only have other segments of $R_{ext}$ attached to it on the side toward

the interior of RT($A$), since it is nearest the outside boundary of $G(A)$. These attached segments are shortened to remain connected to the shifted segment on the same side. The shifted segment is merged with any segment at its new position which it abuts or overlays. If one end of the shifted segment is a vertex of RT($A$), an edge is added to $R$, interior to RT($A$), connecting the new end vertex to the previous end vertex. If the addition of that edge creates a cycle in $R$, the edge is not added. The segment shift operation maintains $R$ as a tree, since connectivity is maintained and no cycles are introduced. Furthermore, the operation does not increase the overall length of the tree, as only one edge may be added and this is offset by the removal of at least one edge of the same length. Thus, $R$ is maintained as an RST.

LEMMA 3.5. *The shifted segment has one end which is a vertex of* RT($A$) *and has only one segment attached to it in the direction of the interior of* RT($A$). *Furthermore, the shifted segment does not overlay a segment of* $R_{\text{ext}}$ *at its new grid position.*

*Proof.* Any other situation would result in a reduction in the length of $R$. This would contradict the assumption that $R$ is an optimal RST. □

We are now in a position to state and prove our major result concerning the inclusion of an optimal RT within RT($A$).

THEOREM 3. *Given a set $A$ of points generating a rectangle tree* RT($A$) *such that* RT($A$) *is a minimum-distance RT, there exists an optimal RST (in the plane) for $A$ consisting solely of vertices and edges of* RT($A$).

*Proof.* Let $R$ denote an optimal RST for $A$ contained in $G(A)$. Assume that $R$ is not strictly contained in RT($A$). Let $R_{\text{ext}}$ denote that part of $R$ lying in the exterior of RT($A$), as defined above.

Since there is a finite number of grid lines in each subgraph of the exterior of RT($A$) and the segment shift operation reduces the maximum distance of segments of $R_{\text{ext}}$ from the interior of RT($A$) by one grid line, a finite number of applications of this operation eliminates all segments of $R_{\text{ext}}$ in a given subgraph of the exterior of RT($A$). Since there are at most four such subgraphs, $R_{\text{ext}}$ can be eliminated, yielding an optimal RST consisting only of edges and vertices of RT($A$). □

To look at this another way, Lemma 3.5 implies that in each subgraph of $R_{\text{ext}}$ there is at most one simple path connecting two vertices of RT($A$). Since RT($A$) is a minimal distance RT($A$), the path in $R_{\text{ext}}$ can be replaced by one within RT($A$) of equal length.

This now allows us to extend the usefulness of Algorithm RT.

THEOREM 4. *Given a set $A$ of points in the plane generating* RT($A$) *such that* RT($A$) *is a minimum-distance, separable RT, Algorithm RT determines an optimal RST (in the plane) for set $A$.*

*Proof.* Since there exists an optimal RST for $A$ which consists solely of vertices and edges of RT($A$), Algorithm RT not only determines an RT-optimal RST but also determines an optimal RST for set $A$. □

**5. Characterization of minimum-distance RT's.** Since Algorithm RT determines an optimal RST for sets of points generating separable, minimum-distance RT's, a further characterization of minimum-distance RT's would be useful.

Two vertices of an RT *interact* if there exists a straight path between them in the RT.

THEOREM 5. *Let $T$ be an RT. $T$ is a minimum-distance RT if and only if every pair of vertices of degree 4 which interact in $T$ are connected by an interior edge of $T$.*

*Proof.* Let $T$ be a minimum-distance RT. Let $u$ and $v$ be interacting vertices of degree 4. Let $p$ be the straight path between $u$ and $v$ in $T$. There are two edges incident to both $u$ and $v$ which are perpendicular to path $p$. If $T$ is a minimum-distance RT, all

four of these edges must be elements of the same arm of $T$. This can only be the case when path $p$ is an interior edge of $T$.

Let $T$ be an RT such that every pair of vertices of degree 4 which interact are connected by an interior edge of $T$. Let $u$ and $v$ be any two vertices of $T$. If a directed path $p$ exists from $u$ to $v$ which has edges directed in only one or two of the four rectilinear directions then $p$ is a minimum-distance path. Let $A_1, \cdots, A_n$ be the sequence of arms within which the directed path $p$ from $u$ to $v$ must lie, where $u \in A_1$ and $v \in A_n$ ($n \geqq 1$). Certainly, if $n = 1$ or 2, then $p$ has minimum distance. If $n > 2$, then the path is continued by connecting the path to the nearest vertex in $A_3$ and similarly continued for greater $n$. If $p$ must contain a third rectilinear direction, it will be due to opposite directions which must be chosen in $A_i$ and $A_{i+2}$, $1 \leqq i \leqq n-2$. This implies that the path chosen to connect $A_i$ and $A_{i+2}$ must connect two vertices of degree 4. Furthermore, the path through $A_{i+1}$ is not an interior edge. This contradicts the assumption about interacting vertices of degree 4 in $T$. □

We next show that every minimum-distance RT falls into one of 11 classes of graphs. In order to do this, we give the following, equivalent definition of a rectangle tree. A graph $G$ is a *rectangle tree* if and only if it can be constructed from a (non-trivial) ladder by a finite number of applications of the following operation:

Form a new RT by adding a new ladder, joining it to the existing RT by identifying one rectangle of the ladder with one rectangle of the RT, in such a way that the new ladder becomes an arm of the new RT.

This definition is equivalent to our initial characterization. Each rectangle added by that definition either extends an existing arm or begins a new arm which shares a rectangle with an existing arm. Therefore, any given RT can be constructed by repeatedly sharing a rectangle of an existing RT with a rectangle of a new ladder. There are only four types of rectangle-sharing operations. These are as follows:

(1) a CROSS-share: the shared rectangle is not an end-rectangle in either the existing RT or in the new ladder;
(2) an END-$T$ share: the shared rectangle is an end-rectangle in the existing RT but not in the new ladder;
(3) a SIDE-$T$ share: the shared rectangle is an end-rectangle in the new ladder but not in the existing RT;
(4) an L-share: the shared rectangle is an end-rectangle in both the existing RT and the new ladder.

Figure 6 illustrates each of the above operations, the new ladder being shown as dotted in each case.

A *segment* of an RT is a straight path of maximal length $\geqq 2$. An *open segment* is a segment containing no vertices of degree 4. When a new ladder is added to an existing RT by one of the above four sharing operations, either one vertex of degree 4 or two adjacent vertices of degree 4 are created on one or two segments of the existing RT.

THEOREM 6. *Let $T$ be an RT. $T$ is a minimum-distance RT if and only if $T$ can be constructed in such a way that when new ladders are added, vertices of degree 4 are created only on open segments of the existing RT.*

*Proof.* The proof follows immediately from Theorem 5. □

By definition, a rectangle has no segments. An *extended ladder* is a ladder containing more than one rectangle. An extended ladder has two open segments. The $L$-share operation leaves the number of open segments unchanged. The END-$T$ share and SIDE-$T$ share operations reduce the number of open segments by one, while the CROSS-share operation reduces the number of open segments by two. As such, there are never more than two open segments in any RT.

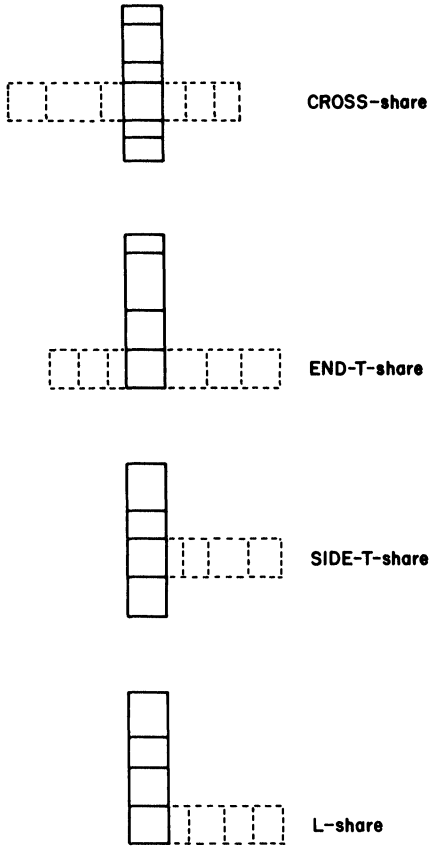



FIG. 6

The above result and Theorem 6 allow us to completely enumerate all classes of minimum-distance RT's. Figure 7 presents a generation graph of the classes of minimum-distance RT's. The extended ladder is the root. Each directed arc of the diagram is labeled with the number of the share operation producing the new class of minimum-

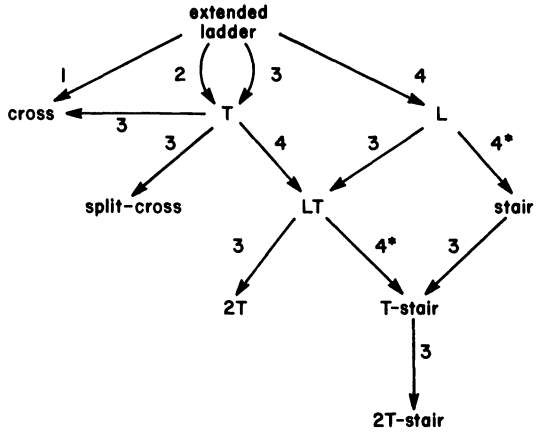



FIG. 7

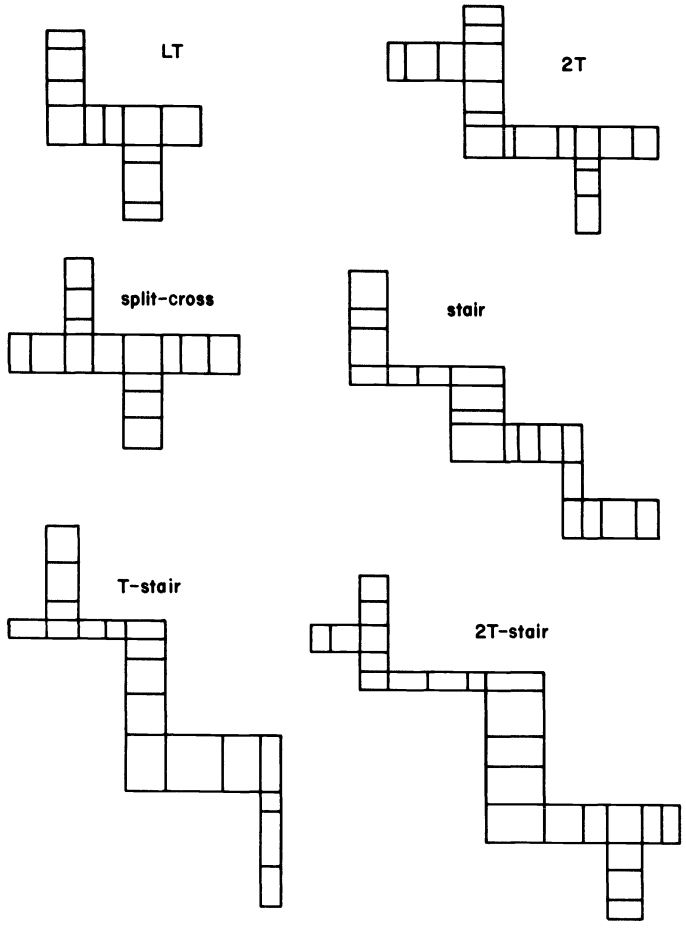



FIG. 8

distance RTs—where operation 1 is the CROSS-share, 2 is the END-$T$-share, 3 is the SIDE-$T$-share, and 4 is the $L$-share. A star indicates one or more applications of the operator. Each class of RTs with no exiting arc contains no open segment upon which to build. Figure 8 presents an illustration of each class of minimum-distance RT introduced in Figure 7.

COROLLARY 6. *A graph $G$ is a minimum-distance RT if and only if $G$ is a member of one of the following classes of graphs*:

1. *rectangle*
2. *extended ladder*
3. $L$
4. $T$
5. $LT$
6. $2T$
7. *cross*
8. *split-cross*
9. *stair*
10. *T-stair*
11. *2T-stair*

**6. The recognition problem for minimum-distance RT's.** The usefulness of Algorithm RT would be greatly enhanced if it could be coupled with a polynomial time algorithm for deciding whether an arbitrary set of points in the plane generates a minimum-distance RT.

Although we have not been able to produce a polynomial algorithm for recognizing an arbitrary minimum-distance RT, it is easy to construct a linear algorithm for recognizing rectangles, extended ladders, L's, T's, LT's, crosses, and split-crosses. These graphs all contain a determinate number of vertices of degree 4.

The recognition algorithm requires (expects) one or two columns of points (a rectangle or extended ladder), followed by one or two rows of several points, none of which lie between the prior two columns (an L or T), followed by one or two columns of points which are aligned with earlier points (a T, LT, cross, or split-cross). The algorithm is applied in each of the four rectilinear directions, accepting the set if succeeding in any direction. The number of steps is linearly related to the number of lines in $G(A)$ which is linearly related to the number of points in $A$. This recognition algorithm is formally described elsewhere and other more general algorithms are under study [7].

Once a minimum-distance RT has been recognized it is a simple matter to determine whether or not it is separable. A traversal of the edges on the exterior face of the graph is sufficient.

**7. Summary.** In this paper we have introduced the class of rectangle trees and two special subclasses: separable and minimum-distance RTs. We have shown

(i) that a linear algorithm exists for finding an RT-optimal RST in a separable RT;

(ii) that a linear algorithm exists which determines optimal RSTs for sets of points in the plane generating minimum-distance, separable RTs;

(iii) that the class of minimum-distance RTs can be characterized completely by eleven subclasses of minimum-distance RTs;

(iv) that a linear algorithm exists for determining whether an arbitrary set of points in the plane generates a ladder, cross, T, L, LT, or split-cross.

A natural application for these results is in the design of wiring or printed circuit layouts interconnecting a subset of terminals embedded in the plane. We extend the set of allowable terminal patterns to include separable, minimum-distance RTs. These

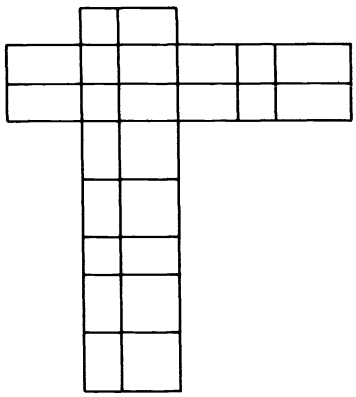


FIG. 9

results appear extendible to RTs in which neighboring, parallel arms are separated by distances sufficient to guarantee that the space between them should be traversed at most once. Interesting questions based upon generalizations of our work suggest themselves. For example, a rather conservative generalization leads to the question: Can the RT-optimal RST problem be solved in polynomial time for sets of points which generate two-wide rectangle trees? The basic building block of two-wide rectangle trees is the ladder containing two rectangles. Figure 9 presents an example of a two-wide rectangle tree.

## REFERENCES

[1] A. AHO, M. R. GAREY AND F. K. HWANG, *Rectilinear Steiner Trees: Efficient special case algorithm*, Networks, 7 (1977), pp. 37–58.

[2] M. R. GAREY AND D. S. JOHNSON, *The Rectilinear Steiner Tree problem is NP-complete*, SIAM J. Appl. Math., 32 (1977), pp. 826–834.

[3] M. HANAN, *On Steiner's problem with rectilinear distance*, SIAM J. Appl. Math., 14 (1966), pp. 255–265.

[4] F. HARARY, *Graph Theory*, Addison-Wesley, MA, 1966.

[5] F. K. HWANG, *On Steiner minimal trees with rectilinear distance*, SIAM J. Appl. Math., 30 (1976), pp. 104–114.

[6] K. L. MCANANEY, *The number and stability of $C_n$-trees*, Proceedings of the Fourth Australian Conference, Lecture notes in Mathematics No. 560, Dold and Eckman, eds., Springer-Verlag, New York, 1975.

[7] S. MITCHELL, S. HEDETNIEMI AND A. FARLEY, *Recognition of minimum-distance rectangle trees*, in preparation.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0011 $01.00/0

# ON THE SEQUENTIAL SEARCH FOR SPATIALLY-DISTRIBUTED EVENTS*

ARNOLD BARNETT† AND JOHN MAZZARINO‡

**Abstract.** Events arise at two points according to independent Poisson processes; their durations are independent and identically distributed random variables. An observer can make visits to the two points, and, on any particular visit to either point, can detect all events then going on there. There is a "dead time" associated with travel from one point to the other. The problem is: What should the observer's visiting strategy be if his goal is to maximize the steady-state fraction of events he observes at least once? We prove a series of theorems about an optimal search strategy that ultimately provide the basis of an algorithm shown to converge to that strategy.

**Introduction.** In a recent paper [1], one of the current authors considered the following problem:

Events arise at a series of discrete points according to independent Poisson processes. The durations of different events are independent and identically distributed random variables, and several events can occur simultaneously at any point. At instants of time exactly one unit apart, an observer can visit any one of the points; when visiting a point he detects all events then going on there. Under what strategy should he make his visits so as to maximize the steady-state fraction of events he observes at least once?

The problem was solved completely when events could arise at only two points and, more generally, it was shown that there is a cyclic optimal search policy. An "exclusion" theorem identified some points the observer need never visit, while a "coexistence" theorem showed that, in some circumstances, a search problem with $N$ points could properly be decoupled into a series of two-point problems.

The paper just described was motivated by a concern for search efforts for events that arise randomly in time and space and, after a short period, disappear. Such search problems differ substantially from those whose targets are stationary. Police patrols intended to intercept crimes in progress, for example, are successful only if they pass the scene of the crime while it is still going on. A radar scanner that focuses on an area after a missile has passed through it will not detect the penetration. On a desolate road, a state patrol vehicle might have to reach an accident within a certain time of its occurrence; otherwise lives will be lost.

The model in [1] captured many salient features of such search problems, but was strikingly unrealistic in one major respect. It was assumed that the observer could move between *any* two points between successive instants of search, and thus that physical distances imposed no constraints on possible search strategies. The idyllic character of that assumption warrants the paper's admission that it "is not itself greatly applicable to practical problems," and leads one to wonder what the optimal search policies would be if the restrictions on the observer's mobility were explicitly considered. A simple investigation into this matter is the subject of this paper.

We concentrate on search models in which events are generated at two different points. Such models form the simplest discrete approximation to a continuous region consistent with the concept of spacially-distributed events. There is transit-related "dead time" (not necessarily the same in both directions) associated with the observer's moving from one point to the other. We prove a series of theorems that, taken together,

* Received by the editors August 24, 1978. This research was supported in part by the Office of Naval Research under contract N 00014-67-A-0204-0063.

† Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.
‡ Peat, Marwick and Mitchell, Washington, D.C.

provide the basis for an algorithm that is shown to converge rapidly to an optimal sequential search policy. The hope is that the models considered, though highly idealized, capture enough of the dynamics and constraints of certain search problems that their solutions provide useful "rules of thumb" about the proper allocation of effort.

We end the paper with some numerical examples, and a brief discussion of some related work by Chelst [2] and of the difficulties that attend attempts to generalize the results to problems in which events arise at three or more points.

**1. A two point search model.** Events arise at two points, 1 and 2, according to independent Poisson processes with parameters $\lambda_1$ and $\lambda_2$ respectively. The durations of various events are independent and identically distributed random variables, with cumulative distribution function $F(x)$. There is no limitation on the number of simultaneous events at either point. An observer can make periodic visits to the two points; on any particular visit to either point, he notices all events then going on there. After a visit to point $i$, the observer has two options for his next visit: (1) return to point $i$ one unit of time later or (2) go to the other point $j$, and arrive there after a trip time $C_{ij}$. ($j = 3 - i$; we assume $C_{ij} > 1$.)

We make the tacit assumption that, on detecting an event, the observer loses no search time dealing with it. This assumption is sometimes realistic; when an accident is observed from a traffic helicopter, for example, it seems more likely that the pilot will radio for emergency help than land himself.

The problem is: What should the observer's visiting strategy be if, over an extremely long period, his objective is to maximize the expected number of new sightings per unit time? (This goal is equivalent to maximizing the steady-state fraction of events observed at least once.) We will proceed below to prove some theorems about characteristics of the optimal search policy; taken together, they lead directly to the precise optimal policy for given $\lambda_1$, $\lambda_2$ and $F(\cdot)$. We begin with some useful general observations.

Let $a_w$ be the expected number of events seen for the first time when the observer returns to point 1 after an absence of exactly $w$ time units. As is shown in [1], $a_w$ is given by:

$$a_w = \int_0^w \lambda_1[1 - F(t)]\, dt.$$

$F(t)$, a probability distribution, is a nondecreasing function of $t$ that grows from 0 to 1; from this fact is follows that (i) $a_w$ is a nondecreasing function of $w$ and (ii) $d_w \equiv a_{w+1} - a_w$ is a nonincreasing function of $w$. The quantity $b_w$, analogous to $a_w$ for point 2, follows $b_w = Ba_w$, where $B = \lambda_2/\lambda_1$. We will assume for convenience that $\lambda_1 > \lambda_2$, although the special case $\lambda_1 = \lambda_2$ presents no added problems. We also assume that $\lim_{w\to\infty} a_w$ is finite; this is equivalent to assuming that the event durations have a finite mean.

What we want to do is specify the pattern of visits to 1 and 2 that yields the highest expected detection rate in terms of the relevant $a_w$'s and $b_w$'s.

**2. On restricting search to the busier point.** Not surprisingly, the search policy under which the observer remains at (slower) point 2 is inherently nonoptimal. Under this policy, the observer visits 2 at intervals of one time unit, and thus witnesses new events at an average rate $b_1$ per unit time. This quantity is strictly lower than the observation rate $a_1$ that can be achieved if the observer simply stays at point 1; hence an optimal search policy must allocate some portion of the visits to the busier point.

Theorem 1 proved below identifies the circumstances in which *all* the search effort should be devoted to 1.

THEOREM 1. *Let* $b_\infty = \lim_{w\to\infty} b_w$, $c = c_{12} + c_{21}$, $S_0 = b_\infty + a_c - ca_1$, $h_k = b_1 + d_{c+k} - a_1$, *and* $A = \min\{k/h_k \leqq 0\}$ *where* $k$ *is a nonnegative integer. The policy of making observations only at point* 1 *is optimal if and only if* $s_0 + H(A) < 0$, *where* $H(A) = \sum_{j=0}^{A-1} h_j$ *for* $A > 0$ *and* $H(A) = 0$ *if* $A = 0$.

If the observer is currently at 1, he should not make any round-trips to 2 unless, over the duration of the excursion, he can make new sightings at a rate[1] higher than $a_1$, the rate he could achieve by "staying put." Let $c = C_{12} + C_{21}$ be the round-trip travel time for a 1-2-1 journey. Consider a trip under which the observer goes from 1 to 2, stays there for $k$ time units ($k$ is assumed a nonnegative integer), and then returns to 1. He witnesses $b_\infty$ or fewer new events on arriving at 2, then $b_1$ more events in each of his further visits there, then $a_{c+k}$ events on his return to 1. Hence his overall sighting rate is bounded from above by $(b_\infty + kb_1 + a_{c+k})/(c+k)$; this falls below $a_1$ unless $S_k$, defined by $S_k = b_\infty + kb_1 + a_{c+k} - (c+k)a_1$, exceeds zero. Thus if $S_k < 0$ for all $k$, the observer should never leave 1.

Note that $S_0$ (which corresponds to a trip with "no layover" at 2 follows $S_0 = b_\infty + a_c - ac_1$ while for $k \geqq 0$, there is the recursive relationship: $S_{k+1} = S_k + h_k$, where $h_k = b_1 + d_{c+k} - a_1$ (again, $d_w = a_{w+1} - a_w$). Let $A = \min(k | h_k \leqq 0)$. Note that because $d_w$ is nonincreasing in $w$, $h_k$ is nonincreasing in $k$.

Since for $k > 0$ $S_k = S_0 + \sum_{j=0}^{k-1} h_j$, it is clear that if $S_0 < 0$ and $\sum_{j=0}^{A-1} h_j < |S_0|$, then $S_k < 0$ for all $k$. This is the most general condition under which all $S_k$'s are negative and, for this reason, searching at 2 is unwarranted. Is it possible that even if some $S_k$'s exceed zero, point 2 should be ignored by the observer? The answer to this question is "no," as we explain below.

Suppose that $S_k > 0$ for a particular $k$, and consider the search policy under which the observer leaves 1 for 2 every $n$ units, stays at 2 for $k$ units and, on returning to 1, remains there $n - c - k$ units before his next departure. Straightforward calculations make clear that this policy yields a detection rate higher than $a_1$ (i.e. is better than just staying at 1) if

$$S_k - (b_\infty - b_{n-c-k}) > 0.$$

Since $\lim_{w\to\infty} b_w = b_\infty$ and $S_k > 0$, there must exist some $n$ for which $S_k > b_\infty - b_{n-c-k}$, which implies the nonoptimality of devoting all effort to point 1. This completes the proof of Theorem 1.

Because $h_k$ is nonincreasing in $k$ the following corollary to Theorem 1 arises:

COROLLARY 1. *If* $S_0 \leqq 0$ *and* $h_0 \leqq 0$, *the optimal search strategy is to make all observations at point* 1, *and this to achieve a sighting rate of* $a_1$ *new events per unit time.*

**3. Some remarks on the optimal strategy.** The following remarks about the optimal search strategy facilitate the task of finding it. Their proofs, which are fairly straightforward, are omitted.

*Remark* 1. If the optimal search strategy involves visits to both points, the observer should only leave one point for the other immediately after an instant of search.

*Remark* 2. One can achieve the maximum possible sighting rate under a cyclic search strategy, in which (i) the time between successive $1 \to 2$ trips is unchanging and (ii) the time the observer spends at 2 is the same on all visits.

[1] For the remainder of this paper, we will use "rate" to mean expected rate and, in all discussion, consider only average detection rates.

Remark 2 means there is an optimal strategy characterized by two parameters, $n^*$, its complete cycle length and $k^*$, the length of each visit to point 2.

*Remark* 3. Let $\bar{k}$ be the set of integer visit lengths at 2 for which $S_k > 0$. $\bar{k}$ is a finite set, and $k^*$, the length-of-stay under the optimal policy, is contained in it.

*Remark* 4. $n^*$, the cycle length of the optimal search policy, is bounded from below by $c+2k^*$.

Remark 4 reflects the nonoptimality of spending more time at 2 than at 1.

*Remark* 5. $n^*$ is finite unless, because of Theorem 1, the observer should never leave 1.

Remark 3 implies, among other things, that $k^*$ is bounded from above by $k_0$, which is defined as the largest $k$ for which $S_k > 0$. Theorem 2 below gives a better upper bound on $k^*$, in the sense that the bound never exceeds $k_0$ and often falls considerably below it. It is the first of three forthcoming theorems that reduce drastically the number of serious candidates for $k^*$ and $n^*$.

**4. Reducing the candidates for optimal policy.**

THEOREM 2. *$k^*$ is bounded from above by the quantity $A$, where $A$ is the smallest nonnegative integer for which* $(b_1 - a_1) + d_{A+c} \leqq 0$.

The "$A$" just mentioned is, of course, the same $A$ as appeared in Theorem 1. To prove the theorem we must first prove two lemmas:

LEMMA 2A. *Let $k(n)$ be the optimal choice of $k$ given that the cycle length is fixed at $n$. Then* $\lim_{n\to\infty} k(n) = A$.

The expected detection rate $E(k, n)$ under the cyclic search policy with parameters $n$ and $k$ follows:

$$E(k, n) = \frac{kb_1 + b_{n-k} + a_{c+k} + (n-c-k)a_1}{n} = a_1 + \frac{S_k + (b_{n-k} - b_\infty)}{n}.$$

Since $E(k, n)$ can exceed $a_1$ only if $S_k > 0$, we know that $k(n)$ cannot be greater than $k_0$ defined above. As noted earlier, successive $S_k$'s are related by the equation: $S_{k+1} = S_k + h_k$. Because A is the smallest integer for which $h_k < 0$, $S_A$ must be the largest of the $S_k$'s.

As $n$ gets very large, it is clear that $b_{n-k} - b_\infty \to 0$ for all $k \leqq k_0$. Thus beyond some threshold $n$, the $E(n, k)$'s for $k \leqq k_0$ are sufficiently close to $a_1 + S_k/n$ that they are maximized when $k = A$; this completes Lemma 2A.

LEMMA 2B. *If $n > n'$, then $k(n) \geqq k(n')$.*

For notational ease, we write $k(n) = k$ and $k(n') = k'$ in this proof. If, for a given $n$, there are two or more $k$-values that tie for optimal, we define $k(n)$ as the smallest.

Suppose that $n > n'$ but $k(n') > k(n)$. From the definitions of $k(n)$ and $k(n')$ we have the two relationships.

(A) $$b_{n'-k'} + k'b_1 + a_{c+k'} + (n'-c-k')a_1 > b_{n'-k} + kb_1 + a_{c+k} + (n'-c-k)a_1$$

and

(B) $$b_{n-k} + kb_1 + a_{c+k} + (n-c-k)a_1 \geqq b_{n-k} + k'b_1 + a_{c+k'} + (n-c-k')a_1.$$

Let $L_A(R_A)$ be the left (right) hand side of (A), and let $L_B$ and $R_B$ be the corresponding quantities of (B). Since $L_A - R_B > L_B - R_A$, it follows that:

(C) $$b_{n-k} - b_{n'-k} > b_{n-k'} - b_{n'-k'}.$$

Given the concavity of the function $b$, inequality (C) is inconsistent with the assumption that $k' > k$. Thus we can conclude that $k' \leqq k$ as claimed.

Taken together, Lemmas 2A and 2B immediately imply that $k(n^*) \leqq A$, which was to be proved.

THEOREM 3. *Let $E(k, n)$ be the detection rate under the cyclic search policy with parameters $k$ and $n$. If $E(k, n) > E(k, n+1)$, then $E(k, n) > E(k, n+q)$ where $q$ is any positive integer.*

As noted earlier, $E(k, x)$ follows:

$$E(k, x) = \frac{kb_1 + a_{c+k} + (x-c-k)a_1 + b_{x-k}}{x}. \tag{D}$$

That $E(k, n)$ exceeds $E(k, n+1)$ readily implies the relationship

$$a_1 + Bd_{n-k} < E(k, n) \tag{E}$$

where $B = \lambda_2/\lambda_1$ as before.

From Equation (D), we see that $E(k, n)$ will exceed $E(k, n+q)$ if and only if:

$$qa_1 + B \sum_{j=0}^{q-1} d_{j+n-k} < qE(k, n). \tag{F}$$

Since $d_{j+n+k} \leqq d_{n+k}$ for $j > 0$, the left-hand side of (F) is bounded from above by $q(a_1 + Bd_{n+k})$. Therefore, multiplying (E) by $q$ implies that condition (F) must be satisfied if $E(k, n) > E(k, n+1)$, which proves Theorem 3.

Theorem 3 is useful in that it suggests an orderly procedure for finding an optimal policy. For a given "serious" value of $k$, one first calculates $E(k, n)$ for $n = c + 2k$. One then proceeds iteratively both to increase $n$ by 1 and calculate $E(k, n)$, *until* one reaches $\tau$, the first cycle length whose detection rate is not the maximum achieved so far. (i.e. $E(k, \tau-1) > E(k, \tau)$). Because of Remark 5, we can be confident that $\tau$ is finite; from Theorem 3, we know that $E(k, \tau-1) > E(k, \tau+q)$ for all positive integers $q$. Thus we can treat the cyclic policy with parameters $(k, \tau-1)$ as the "candidate" for optimal policy tied to the given value of $k$. We could obtain the "candidates" for best policy for each $k$ that satisfies both Remark 3 and Theorem 2; the one with the highest detection rate must be an optimal policy.

The approach just described figures prominently in an algorithm for finding $n^*$ and $k^*$ to be presented in the next section. One last theorem, proved below, allows further simplification of the search for an optimal strategy. We have already shown that in searching for $k^*$, one can exclude all $k$-values outside a certain range. Theorem 4 implies that, even within that range, it is often unnecessary to consider all values of $k$.

THEOREM 4. *Let $v(k)$ be the highest attainable sighting rate under the constraints that* (i) *all visits to* 2 *must be exactly k units long and* (ii) *each cycle length must be at least $c + 2k$. If $v(k) > v(k+1)$ for a particular $k$, then $v(k) > v(k+q)$ for any positive integer $q$.*

To prove Theorem 4 it is sufficient to show, for any $k$ and any $q > 1$, that if $v(k+q) > v(k)$, then $v(k+1) > v(k)$. For notational and conceptual ease we will establish this result for the special case $k = 0$ and $q = 2$; the reasoning for the general case is wholly analogous.

Let $w_0$ be the maximum cycle length for which $E(0, w_0) = v(0)$. (We say "maximum" to cover the case where $v(0)$ is achieved under consecutive cycle lengths.) By definition of $v(0)$, the inequality $E(0, w_0) > E(0, w_0+1)$ prevails; hence Equation (D) implies that[2]

$$v(0) > a_1 + [b(w_0+1) - b(w_0)]$$

[2] In this proof, we write $b_x = b(x)$ and $d_x = d(x)$ for clarity.

or

$$v(0) > a_1 + Bd(w_o). \tag{G}$$

If $w_0 > c$, we have the further relationship $E(0, w_0) \geqq E(0, w_0 - 1)$, which leads to:

$$v(0) \leqq a_1 + d(w_0 - 1) \quad \text{if } w_0 > c. \tag{H}$$

It turns out that in proving Theorem 4, one must deal separately with the cases $w_0 = c$ (i.e. when the observer shuttles between 1 and 2 with no pauses at either point) and $w_0 > c$. We consider below the case $w_0 > c$; the argument when $w_0 = c$ is similar but simpler. The general approach is to show that if $v(2) > v(0)$, one can construct a policy with $k = 1$ whose detection rate exceeds $v(0)$. Important to the construction is the lemma below.

LEMMA 4A. *Let* $w_0 = \max(n | E(0, n) = v(0))$. *If* $v(2) > v(0)$, *then* $E(2, w_0 + 2) > v(0)$.

We prove Lemma 4A for $w_0 > c$; it is also true for $w_0 = c$. The proof uses indirect reasoning.

Simple manipulations with Equation (D) yield the relationship:

$$E(2, x+1) = \frac{x}{x+1} E(2, x) + \frac{1}{x+1}(a_1 + Bd(x-2)). \tag{J}$$

Equation (J) reminds us that $E(2, x+1)$ is actually a weighted average of $E(2, x)$ and $a_1 + Bd(x-2)$, the latter quantity being the incremental detection rate when the cycle length is raised from $x$ to $x+1$. Now suppose $E(2, w_0+2) \leqq v(0)$ although $v(2) > v(0)$. Taken together, (G) and (J) imply that if $E(2, w_0+2) < v(0)$, then $E(2, w_0+3) < v(0)$ and, because of the concavity of $d(\ )$, it follows inductively that $E(2, y) < v(0)$ for all $y \geqq w_0 + 2$. Let $w_2 = \max(n | E(2, n) = v(2))$. Since $v(2) > v(0)$ by hypothesis, we have just shown that if $E(2, w_0+2) < v(0)$, it must be true that $w_2 < w_0 + 2$.

If $w_2 < w_0 + 2$, one can use (D) to obtain:

$$E(2, w_0+2) = \frac{w_2}{w_0+2} v(2) + \frac{1}{w_0+2}\left[\sum_{j=0}^{w_0+1} (a_1 + Bd(j-2))\right]. \tag{K}$$

Since $v(2) > v(0)$ and $a_1 + Bd(x) \geqq v(0)$ for all $x < w_0$ (Equation (H) and concavity), it follows from (K) that $E(2, w_0+2) > v(0)$, which contradicts the assertion that this inequality does not hold. The lemma is proved.

With Lemma 4A, the proof can be brought to a rapid conclusion. From (D) and the definition of $w_0$, one can easily show that:

$$E(2, w_0+2) = \frac{w_0}{w_0+2} v(0) + \frac{1}{w_0+2}[2b_1 + d(c) + d(c+1)]. \tag{L}$$

If $v(2) > v(0)$, then $E(2, w_0+2) > v(0)$ from Lemma 4A; thus (L) implies that $[2b_1 + d(c) + d(c+1)]/2 > v(0)$ or, by concavity, that

$$b_1 + d(c) > v(0). \tag{M}$$

Equation (D) also allows one to write

$$\text{(N)} \qquad E(1, w_0+1) = \frac{w_0}{w_0+1} v(0) + \frac{1}{w_0+1}(b_1 + d(c)).$$

Combining (M) and (N) shows that $E(1, w_0+1) > v(0)$; since $w_0 > c$, that is enough to show that if $v(2) > v(0)$, then $v(1) \geqq E(1, w_0+1) > v(0)$, which completes the proof.[3]

Theorem 4 means that once $v(k)$ starts to decline as $k$ is increased it can never return to its former level; thus $k^*$ has already been passed. With this and earlier results, we are ready to discuss a recursive procedure for obtaining $k^*$ and $n^*$.

**5. An algorithm for the optimal strategy.** Directly or indirectly, the algorithm below uses all four theorems and all five remarks already discussed. We assume that, at the outset, the user of the algorithm has calculated the needed $a_w$'s from $a_w = \lambda_1 \int_0^w [1-F(t)]\,dt$, the $b_w$'s from $b_w = Ba_w$, the quantity $A$ from $A = \min(k | b_1 - a_1 + d_{k+c} < 0)$, and $S_k = b_\infty + kb_1 + a_{c+k} - (c+k)a_1$ for all $k$ from 0 to $A$. The goal is to obtain the values of $k^*$ and $n^*$, as well as $R^* = E(k^*, n^*)$, the optimal detection rate.

THE ALGORITHM.

(1) If $S_A < 0$, then set $k^* = 0$, $n^* = \infty$, and $R^* = a_1$. (i.e. the observer should never visit 2.)
(2) If $S_A > 0$, then find $a$, the smallest nonnegative integer for which $S_k > 0$.
(3) Go to (4), setting "$a$" as the current value of $k$.
(4) Calculate $E(k, n)$ from (D) for $n = c+2k$. Do the same for $n = c+2k+1$. Continue to do so as $n$ is increased in steps of 1 until reaching $\tau$ the smallest $n \geqq c+2k+1$ for which $E(k, \tau) < E(k, \tau-1)$. Set $w_k = \tau - 1$ and $v(k) = E(k, w_k)$. Record $w_k$ and $v(k)$.
(5) If $k = a$ in (4), set $k = a+1$ and return to (4).
(6) If $k > a$ in (4), then compare $v(k)$ and $v(k-1)$.
(7) If $v(k) > v(k-1)$ then:
  (i) if $k = A$, set $k^* = A$, $n^* = w_A$, and $R^* = v(A)$.
  (ii) if $k < A$, increase the $k$-value by 1 and return to (4).
(8) If $v(k) \leqq v(k-1)$ then set $k^* = k-1$, $n^* = w_{k-1}$, and $R^* = v(k-1)$.

Reasoning in the proofs of Theorem 1 and Remark 5 makes clear that $w_k$ is finite for all $k$ that satisfy $a \leqq k \leqq A$. Thus there is no danger that the algorithm will fail to converge to $k^*$ and $n^*$ in a finite number of calculations.

Now we can turn to some numerical examples. But before doing so, we should make explicit two general properties of the optimal search policy.

(i) *The policy uses $C_{12}$ and $C_{21}$ only through their sum $c = C_{12} + C_{21}$.*

Thus the round-trip time from 1 to 1 via 2 is the only needed travel-time parameter.

(ii) *The policy depends on $\lambda_1$ and $\lambda_2$ only through their ratio.*

To see this, note that if both $\lambda_1$ and $\lambda_2$ were changed by the factor $Q$, all $a_w$'s and $b_w$'s would change by the same factor. Consequently, the $E(k, n)$'s would all change by factor $Q$ so $k^*$ and $n^*$ would remain the same.

**6. Numerical examples.** Since the optimal search strategy depends only on the $\lambda_2/\lambda_1$ ratio we can, without loss of generality, set $\lambda_1 = 1$. We therefore need only the values of $B$ and $c$ and the distribution of $x$, the duration of any given event.

[3] If $w_0 = c$, then $(1, w_0+1)$ is not among the candidates for best policy since it is biased towards point 2. As noted before, a slightly different proof applies for $w_0 = c$.

*Example* 1. $B = 28, c = 4$

$$x = \begin{cases} 1 & \text{w.p. } .1 \\ 2 & \text{w.p. } .1 \\ 4 & \text{w.p. } .6 \\ 6 & \text{w.p. } .1 \\ 7 & \text{w.p. } .05 \\ 8 & \text{w.p. } .05 \end{cases} \quad \text{(w.p. = with probability)}$$

Direct calculations with the $x$-distribution reveal that:

$$\begin{aligned} a_1 &= 1 & a_5 &= 3.7 \\ a_2 &= 1.9 & a_6 &= 3.9 \\ a_3 &= 2.7 & a_7 &= 4.0 \\ a_4 &= 3.5 & a_8 &= 4.05 \end{aligned} \qquad a_w = 4.05 \text{ for all } w > 8.$$

The $b_w$'s follow $b_w = .28a_w$; $A = 0$, and $S_0 = b_\infty + a_4 - 4a_1 = .63$. Use of the algorithm leads rapidly to the conclusion that $k^* = 0$, $n^* = 4$, and $R^* = 1.12$. Under the best search strategy, $\frac{7}{8}$ of the events at both points are sighted at least once.

Since $c = 4$, the optimal strategy has the observer shuttling continuously between the two points. One might have suspected that, since 78% of the events arise at 1, the strategy would have concentrated search effort there. This does not happen because 80% of the events last at least 4 units; hence the losses at 1 associated with quick visits to 2 are minimal.

*Example* 2. $B = .985$, $c = 3$,

$$x = \begin{cases} 1 & \text{w.p. } .5 \\ 11 & \text{w.p. } .5. \end{cases}$$

In this problem, $a_w = .5(1 + w)$ for $w \leqq 11$, and $a_w = 6$ for all $w > 11$. $b_w = .985a_w$; $A = 8$. The algorithm implies that $k^* = 8$ and $n^* = 19$. (i.e. the observer should spend 8 units at 1, go to 2, spend 8 units there, return to 1, then repeat the cycle.) $R^* = 1.46$, so about 73% of the events at each point are sighted.

With events arising at almost the same rate at 1 and 2, it is unsurprising that the optimal strategy is symmetric. Note that, under the strategy, the observer returns to each point exactly 11 units since his last departure; he therefore sees on his return all events with durations greater than 1 that occurred during his absence. The long pauses (8 units) at each point arise because, with half the events only one unit long, the losses while the observer is "in transit" work against frequent excursions.

*Example* 3. $B = .6$, $c = 4$, $x = 1.5$ w.p. 1.

In this case

$$a_w = \begin{cases} w & \text{for } w < 1.5, \\ 1.5 & \text{for } w \geqq 1.5. \end{cases}$$

$A = 0$ and $S_0 = -1.6 < 0$. Hence all search effort should be devoted to point 1; under this policy, all events at 1 but none at 2 will be observed, which means that $\frac{5}{8}$ of all events will be sighted.

Example 2 reminds us that sometimes $k^* > 0$, which means that the observer's initial observations at 2 should be followed by further ones at spacings of one time unit. Since the sighting rate on the additional observations is only $b_1$, their presence in the

optimal strategy might seem perplexing; this is particularly so since, were the observer to return from 2 to 1 immediately, he could be detecting new events at the higher rate $a_1$.

This apparent paradox is resolved if we realize that, even if the observer spends $k^*$ units at 2, he will not miss all the events that arise at 1 during that period; some of them will continue until his return to 1 and thus eventually be noticed. By contrast, if visits to 2 are infrequent, it is possible that the bulk of events at 2 over the $k$-unit period will escape detection unless the observer stays there. Thus remaining at 2 is sometimes justified because foregoing some of the events at 1 over a period is preferable to losing a larger fraction of the concurrent events at 2.

**7. The Chelst paper.** The problem discussed here bears certain similarities to one recently considered by Chelst [2]. He assumed that "targets" arrive in Poisson fashion at two different points[4] at equal rates, and that they depart after an exponentially distributed "visiting time". An observer who can search continuously wishes to observe these targets at the highest possible rate. The observer can not usually find targets immediately; even if he restricts his effort to one point, the time between a target's arrival there and its detection—assuming that it does not leave before being sighted—follows an exponential distribution. There is a time $T$ required for travel from either point to the other.

In this symmetric problem (i.e. $\lambda_1 = \lambda_2$), the critical parameter is $x$, the time between the observer's arrival at either point and his next departure for the other. Chelst obtains a transcendental equation from which $x$ can be obtained, through successive approximation methods.

Because of differences in underlying models, it would be inappropriate directly to compare the current results with those of Chelst. But we might observe that, unlike Chelst, we imposed no restrictions at all on the $\lambda_2/\lambda_1$ ratio or on the form of the event-duration distribution. Perhaps that is because our exploitation of the "diminishing returns" concavity property of the objective function made unnecessary any assumptions of symmetry or memorylessness.

**8. Concluding remarks.** As noted earlier, the two-point problem discussed here could serve as a first approximation to an actual search problem in a continuous region. The region could be broken up into two distinct parts, the $C_{ij}$'s perhaps tied to the times of travel between their centers of gravity, etc. Use of the algorithm might well give some insight into the fraction of time the observer should spend in each region, as well as the frequency with which he should switch regions.

Extending the analysis to the case where events arise at $N$ discrete points, $N > 2$, would allow a more detailed approximation of continuous search problems. Unfortunately, such an extension is anything but straightforward. The results obtained here cannot be generalized directly, and *all* the results in [1] about $N$-point optimal search policies fail to hold up when $C_{ij}$'s are unequal to 1. In an $N$-point problem, for example, it is conceivable that the observer should never visit the busiest point, because it is so far from the other points that travel times to and from it are prohibitively large.

Under the circumstances, those considering the $N$-point problem face a formidable task. While one can imagine progress with heuristic approaches, the outlook with analytical methods does not seem particularly bright. Future work, of course, could make such pessimism seem very foolish.

---

[4] While he uses the word "region" rather than "point", he effectively approximates each region by a point. The time to travel between regions, for example, is assigned a constant value.

**Acknowledgment.** This paper benefited from the careful reading and perceptive comments of Professor Stephen Graves.

## REFERENCES

[1] A. BARNETT, *On searching for events of limited duration*, Operations Res., 24 (1976), pp. 438–451.

[2] K. CHELST, *A differential equation model of search for randomly arriving and departing targets*, Wayne State University College of Engineering Technical Report, TR-78-3, 1978.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0012 $01.00/0

# CANDIDATE KEYS AND ANTICHAINS*

J. DEMETROVICS†

**Abstract.** It is shown that a matrix can be constructed in which a row is determined when specified in any collection of column sets with the property that $B \subset C$, $A \subset B$ implies $A \subset C$. The problem of determining the minimum number of rows needed for a collection $C$ on $n$ rows in worst case is raised. Some bounds are given.

In the relational data model proposed by Codd [1]–[3] data are represented by matrices with rows corresponding to records and columns corresponding to attributes. One can identify a row by examining the values of its elements in certain sets of columns, namely those for which no two rows are identical. We call such sets of columns *keys*, and minimal keys are called *candidate keys*.

The candidate keys, being minimal column sets that are keys, form an antichain. If there are $n$ columns there can be at most $\binom{n}{n/2}$ candidate keys by a well-known theorem of Sperner. In this paper we address the question: given an antichain (collection of subsets of $n$ elements such that no member contains another) is there always a matrix having it as its set of candiate keys? Some related questions are discussed.

THEOREM. *If $A$ is an antichain on $n$ elements which are column indices, there exists a matrix for which the candidate keys consist of the sets of columns that are members of $A$.*

*Proof.* Let the antichain $B$ consist of the maximal sets that do not contain members of $A$. Let the members of $B$ be $B_1, \cdots, B_a$. We define $2a$ rows as follows: For $1 \leqq i \leqq a$ let rows $2i-1$ and $2i$ have zero entries in the columns of $B_i$ and entries $2i-1$ and $2i$ respectively in all other columns. Obviously any set of columns not containing a member of $A$ is not a key, as it leaves an ambiguity between the two rows corresponding to each $B_i$ containing it. On the other hand, the number of any row will be displayed for at least one element of each $A_i$.

Two further questions are suggested here:

1. What is the largest number $r(n)$ of rows needed for some $A$ having $n$ columns?
2. What is the smallest alphabet $a(n)$ that the matrix elements may be restricted to for a matrix on $n$ columns to work here?

These two problems are in general open. However, we make the following remarks.

If one member of the antichain is a one element set, all the rows must differ in that column, so that $a(n) \geqq r(n-1)$. The construction above gives $r(n) \leqq 2\binom{n}{n/2}$. One can easily obtain $(2\binom{n}{n/2})^{1/2}$ as a lower bound for $r(n)$. The upper bound here can be improved somewhat but the gap between these is still wide.

**Acknowledgment.** The author thanks the referee for editorial assistance.

## REFERENCES

[1] E. F. CODD, *A relational model of data for large shared data banks*, Comm. ACM, 13 (1970), pp. 377–387.

[2] ———, *Normalized data base structure: a brief tutorial*, Proc. 1971 ACM-SIGFIDET Workshop on Data Description, Access and Control.

[3] ———, *Further normalization of the data base relational model*, Courant Computer Science Symposis 6 "Data Base System," Prentice-Hall, Englewood Cliffs, NJ, 1971, pp. 33–64.

* Received by the editors August 31, 1978, and in revised form July 10, 1979.

† Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Hungary.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0013 $01.00/0

# EFFICIENT ALGORITHMIC SOLUTIONS TO EXPONENTIAL TANDEM QUEUES WITH BLOCKING*

GUY LATOUCHE† AND MARCEL F. NEUTS‡

**Abstract.** Stable queuing systems consisting of two groups of servers, having exponential service times, placed in tandem and separated by a finite buffer, are shown to have a steady-state probability vector of matrix-geometric form. The queue is stable as long as the Poisson arrival rate does not exceed a critical value, which depends in a complicated manner on the service rates, the numbers of servers in each group, the size of the intermediate buffer and the unblocking rule followed when system becomes blocked. The critical input rate is determined in a unified manner.

For stable queues, it is shown how the stationary probability vector and other important features of the queue may be computed. The essential step in the algorithm is the evaluation of the unique positive solution of a quadratic matrix equation.

**1. Introduction.** The queuing model consisting of two units in series with a finite intermediate waiting room has an extensive literature, dating back to 1956 with the work of G. C. Hunt [9]. The study of blocking in two or more units in series without intermediate waiting spaces was initiated by B. Avi-Itzhak and M. Yadin [2]. Further contributions to this model are due to N. U. Prabhu [18] and A. B. Clarke [3].

Models in which there is a finite waiting room between the two units and the service times in the first unit have a general distribution were discussed by T. Suzuki [21], M. F. Neuts [11], [12] and K. Hildebrand [7], basically using transform methods which are not readily computationally implemented. The thesis by I. Hannibalsson [5] utilizes a buffer model to represent a queue with delayed feedback. The second unit then represents a holding stage for those customers who will rejoin the queue in front of unit I. In this paper and also in that by B. Wong, W. Giffin and R. L. Disney [23], the analysis of finite capacity buffer models is carried out by fairly involved spectral decompositions of the transition probability matrices. Related models, with finite total numbers of customers were treated in the papers by K. L. Arya [1] and O. P. Sharma [20]. These papers also do not have an algorithmic orientation.

In recent years, there has been a growing interest in the development of computational methods to evaluate the stationary probability vector and related quantities for tandem queues with blocking. This interest came primarily from the recognition that these models are useful in the study of the behavior of subsystems of computers. In addition to detailed descriptions of several computer-related applications, A. G. Konheim and M. Reiser [10] propose an algorithm for the solution of a system consisting of two single-server units with exponential service time distributions. They also allow feedback of some departures from the second server to the queue in front of the first unit. In [19], these same authors further considered more elaborate forms of feedback and discussed additional applications in computer modeling. Iterative numerical procedures of the Gauss–Seidel type, such as proposed by F. S. Hillier and R. W. Boling [8], may also be implemented for these models.

* Received by the editors September 20, 1978, and in revised form May 15, 1979. This research was supported in part by the Air Force Office of Scientific Research Air Force Systems Command USAF, under Grant AFOSR-77-3236.

† Laboratoire d'Informatique Théorique, Université Libre de Bruxelles, 1050, Brussels, Belgium and University of Delaware, Newark, Delaware 19711.

‡ Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19711.

In addition, bounds on the blocking probability were investigated more recently by F. G. Foster and H. G. Perros [4]. A particularly detailed study of diffusion approximations in tandem queues is due to G. Newell [16], [17]. The recent paper by J. M. Harrison [6] is also relevant in this context.

It is the purpose of this paper to show that a large number of buffer models with exponential servers may be numerically solved in a unified way. The key result identifies their stationary probability vector in a (modified) matrix-geometric form. Appropriately partitioning that vector $\mathbf{x}$ as $(\mathbf{x}_0, \mathbf{x}_1, \cdots)$, we show that

$$\mathbf{x}_i = \mathbf{x}_{r-1} R^{i-r+1} \quad \text{for } i \geqq r-1,$$

where $r$ is the number of servers in the first unit. The matrix $R$ is the unique positive solution to a matrix quadratic equation. The spectral radius of $R$ is less than one. The $r$ vectors $x_0, \cdots, x_{r-1}$, are also uniquely determined.

The approach, which is used here, is already implicit in the thesis of V. Wallace [22], but the proofs are based on further refinements and generalizations given in Neuts [13], [14], [15].

**Description of the model.** The system consists of units I and II and a finite intermediate buffer. Unit I consists of $r$ parallel exponential servers, processing customers at the same rate $\alpha$. In unit II, $c$ parallel exponential servers process customers at the common rate $\beta$. Arrivals to unit I occur according to a homogeneous Poisson process of rate $\lambda$. (See Fig. 1.)

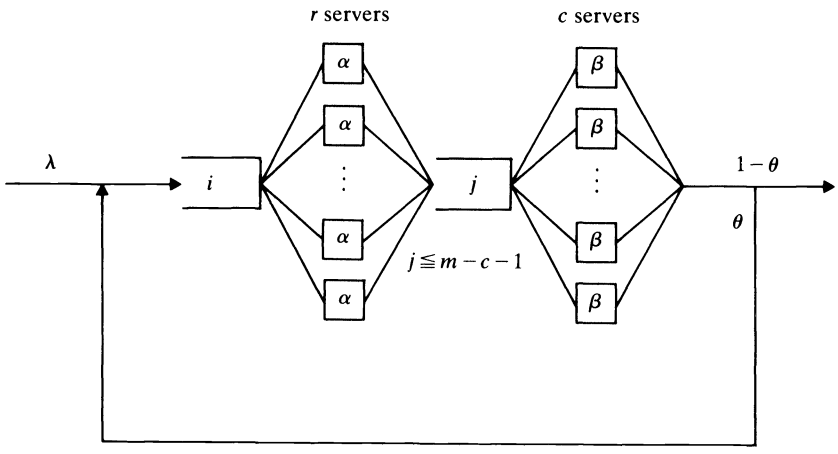



FIG. 1

The servers in unit II can be active as long as there are customers, who have completed a pass through unit I and are requesting their service. There are $M - c - 1 \geqq 0$ places in the buffer, so that at most $M - 1$ customers can be either waiting in the buffer or being processed by one of the servers in unit II. If the number of customers who have completed a pass through unit I but have not cleared unit II reaches $M$, one of the servers in unit I becomes *blocked.*

Depending on the application, the blocking of one or more servers in unit I may affect the ability of the unblocked servers either to accept a customer for service or to complete a service in course. We shall assume that when the number of blocked servers in unit I reaches $r^*$, $1 \leqq r^* \leqq r$, all unblocked servers in unit I also cease service. This situation will be referred to as *full blocking.*

In a partially blocked system, when a service completion in unit II occurs, one of the blocked servers of unit I releases his customer into the buffer. This server may now again initiate a service.

Next we specify *the unblocking rule.* In a fully blocked system there are $M + r^* - 1$ customers who have completed a pass through unit I and are requesting service in unit II. We define an integer $k^*$, $0 \leqq k^* \leqq M + r^* - 2$. When the number of customers, who have not been cleared by unit II, drops to $k^*$, all interrupted services in the servers in unit I resume and any free servers can again initiate services.

We shall allow a feedback loop of departures from unit II back to the queue in front of unit I. With probability $\theta' = 1 - \theta$, $0 \leqq \theta < 1$, a customer who completes a service in unit II leaves the system. Feedback occurs with probability $\theta$.

In order to concentrate only on parameters which have substantial significance, we shall not discuss further extensions in which customers may leave the system from unit I or may enter feedback loops from the buffer to unit I or from unit II to the buffer. The relevant matrices which govern such cases can be constructed easily; the theorems and algorithms discussed below carry over routinely.

We also make the standard independence assumptions. All service times and interarrival times are mutually independent random variables. From a numerical viewpoint, it is routine to consider extensions such as the case where the rate of the Poisson arrival process depends on the number of blocked servers in unit I, but in order not to add to the number of parameters of the model we shall not pursue this topic further.

**Notational convention.** The material in this paper involves a large number of Jacobi matrices, whose detailed definitions require display. A matrix such as

$$\begin{array}{c|cccccccc}
0 & b_0 & c_0 & 0 & 0 & & & & \\
1 & a_1 & b_1 & c_1 & 0 & & & & \\
2 & 0 & a_2 & b_2 & c_2 & & & & \\
\vdots & & & & \ddots & \ddots & & & \\
m-2 & & & & & a_{m-2} & b_{m-2} & c_{m-2} & 0 \\
m-1 & & & & & 0 & a_{m-1} & b_{m-1} & c_{m-1} \\
m & & & & & 0 & 0 & a_m & b_m
\end{array}$$

will be displayed as

$$\left\| \begin{array}{ccccccc}
 & c_0 & c_1 & \cdots & c_{m-3} & c_{m-2} & c_{m-1} \\
b_0 & b_1 & b_2 & \cdots & b_{m-2} & b_{m-1} & b_m \\
a_1 & a_2 & a_3 & \cdots & a_{m-1} & a_m &
\end{array} \right\|.$$

**2. The structure of the Markov process.** Under the assumption of exponential service times for the servers in units I and II, the queuing model may be described as a continuous-parameter Markov chain on the state space $\{(i, j), i \geqq 0, 0 \geqq j \leqq N\}$, where $N$ is a finite nonnegative integer. The index $i$ will denote the number of customers queued up or in service in unit I. Such customers will be called I-*customers.* Upon completion of a pass through unit I, a customer becomes a II-*customer.* We note that because of the possibility of feedback, a customer may be termed a I- or a II-customer several times in succession before leaving the system. The role played by the index $j$ is more complicated to describe and will be spelled out for the specific cases discussed below.

In all cases, however, the infinitesimal generator $P$ of the Markov chain will have the structure of a block-tridiagonal matrix of the form

$$P = \begin{Vmatrix} & A_{02} & A_{12} & \cdots & A_{r-3,2} & A_{r-2,2} & A_2 & A_2 & \cdots \\ A_{01} & A_{11} & A_{21} & \cdots & A_{r-2,1} & A_{r-1,1} & A_1 & A_1 & \cdots \\ A_{10} & A_{20} & A_{30} & \cdots & A_{r-1,0} & A_0 & A_0 & A_0 & \cdots \end{Vmatrix}, \tag{1}$$

where all entries are square matrices of order $N+1$. The rows in the block-partitioned matrix correspond to the sets of states $\{(i, 0), (i, 1), \cdots, (i, N)\}$ for $i \geqq 0$.

We shall now give the detailed definitions of these blocks for various models of increasing complexity.

**Model A.** Unit I blocks as soon as there are $M$ II-customers in the system. All $c$ servers in Unit II are busy; there are $M-c-1$ customers in the waiting room and one server in unit I has completed service of a customer who cannot enter the waiting room. Unblocking occurs as soon as a departure from unit II occurs. In terms of our general description, Model A corresponds to $r^* = 1$, $k^* = M-1$.

In this case $N = M$. The matrices $A_0$, $A_1$ and $A_2$ are given by

$$A_0 = \begin{Vmatrix} & r\alpha & r\alpha & \cdots & r\alpha & r\alpha \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \end{Vmatrix},$$

$$A_1 = \begin{Vmatrix} & 0 & \cdots & 0 & 0 & \cdots \\ -\lambda - r\alpha & -\lambda - r\alpha - \beta & \cdots & -\lambda - r\alpha - (c-1)\beta & -\lambda - r\alpha - c\beta & \cdots \\ \beta\theta' & 2\beta\theta' & \cdots & c\beta\theta' & c\beta\theta' & \cdots \\ & & & & 0 & 0 \\ & & & & -\lambda - r\alpha - c\beta & -\lambda - c\beta \\ & & & & c\beta\theta' & \end{Vmatrix},$$

$$A_2 = \begin{Vmatrix} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \lambda & \lambda & \cdots & \lambda & \lambda & \cdots & \lambda & \lambda \\ \beta\theta & 2\beta\theta & \cdots & c\beta\theta & c\beta\theta & \cdots & c\beta\theta & \end{Vmatrix},$$

and for $1 \leqq i \leqq r-1$,

$$A_{i0} = \begin{Vmatrix} & i\alpha & i\alpha & \cdots & i\alpha & i\alpha \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \end{Vmatrix}.$$

For $0 \leqq i \leqq r-1$, the matrices $A_{i2} = A_2$, and the matrices $A_{i1}$ are given by

$$A_{i1} = \begin{Vmatrix} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ * & * & \cdots & * & * & \cdots & * & * \\ \beta\theta' & 2\beta\theta' & \cdots & c\beta\theta' & c\beta\theta' & \cdots & c\beta\theta' & \end{Vmatrix}.$$

The asterisks correspond to the negative diagonal entries, which are such that the row sums of the matrix $P$ are zero.

**Model B.** This model is as the preceding one, except that full blocking occurs only when $r^*$, $1 \leqq r^* \leqq r$, servers in unit I are blocked. The index $j$ now ranges from 0 to $M + r^* - 1$ and denotes the number of II-customers in the system. Unblocking occurs again upon a subsequent departure from unit II, which corresponds to the case $k^* = M + r^* - 2$.

The blocks in the partitioned matrix $P$ are now of order $M+r^*$. They are obtained by *augmenting* the blocks in Model A in a systematic manner. Specifically

$$A_0 = \left\| \begin{array}{ccccccc} \cdots & r\alpha & (r-1)\alpha & \cdots & (r-r^*+2)\alpha & (r-r^*+1)\alpha \\ \cdots & 0 & 0 & \cdots & 0 & 0 \\ \cdots & 0 & 0 & \cdots & 0 & \\ & \uparrow & & & & \uparrow \\ & M & & & & M+r^*-1 \end{array} \right\|$$

$$A_1 = \left\| \begin{array}{ccccccc} \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & * & * & * & \cdots & * & * \\ \cdots & c\beta\theta' & c\beta\theta' & c\beta\theta' & \cdots & c\beta\theta' & \\ & \uparrow & & & & \uparrow \\ & M & & & & M+r^*-1 \end{array} \right\|,$$

$$A_2 = \left\| \begin{array}{ccccccc} \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \lambda & \lambda & \lambda & \cdots & \lambda & \lambda \\ \cdots & c\beta\theta & c\beta\theta & c\beta\theta & \cdots & c\beta\theta & \end{array} \right\|,$$

where the asterisks stand respectively for the entries $-\lambda - c\beta - (r-1)\alpha, \cdots, -\lambda - c\beta - (r-r^*+1)\alpha, -\lambda - c\beta$, chosen so that the row sums of $P$ are zero.

The matrices $A_{i0}$ are given by

$$A_{i0} = \left\| \begin{array}{ccccccc} \cdots & i\alpha & \min(r-1, i)\alpha & \min(r-2, i)\alpha & \cdots & \min(r-r^*+1, i)\alpha \\ \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & 0 & 0 & 0 & \cdots & \\ & \uparrow & & & & \uparrow \\ & M & & & & M+r^*-1 \end{array} \right\|$$

for $1 \leqq i \leqq r-1$.

For $0 \leqq i \leqq r-1$, the matrices $A_{i2} = A_2$ and the matrices $A_{i1}$ are given by

$$A_{i1} = \left\| \begin{array}{cccccc} \cdots & 0 & 0 & \cdots & 0 & 0 \\ \cdots & * & * & \cdots & * & * \\ \cdots & c\beta\theta' & c\beta\theta' & & c\beta\theta' & \end{array} \right\|.$$

The asterisks correspond to the negative diagonal entries, which are such that the row sums in the matrix $P$ are zero.

**Model C.** In this model, we add further complexity to Model B by assuming that when full blocking occurs, unit I does not become unblocked until the number of II-customers drops to $k^*$. In most cases of interest, we will have $c \leqq k^* \leqq M+r^*-2$, and in order to limit the number of variants, we shall assume that this is the case. The case $k^* = M+r^*-2$ corresponds to Model B, so we only need to discuss the cases where $c \leqq k^* \leqq M+r^*-3$.

We now consider the indices $j$:

$$0, 1, \cdots, M-1, M, \cdots, M+r^*-1, \overline{M+r^*-2}, \overline{M+r^*-3}, \cdots, \overline{k^*+1}.$$

The index values with a bar correspond to the situations where the unit I is blocked, although fewer than $M+r^*-1$ II-customers are in the system. The blocks in the partitioned matrix $P$ are now matrices of order $2M+2r^*-k^*-2$.

The matrices $A_0$ and $A_{i0}$, $1 \leqq i \leqq r-1$, for this model are obtained by adding $M+r^*-k^*-2$ rows and columns to the corresponding matrices for Model B. These rows and columns are identically zero.

The matrices $A_1$ and $A_{i1}$, $0 \leqq i \leqq r-1$, are obtained by adding $M+r^*-k^*-2$ rows and columns to the corresponding matrices for Model B, and changing the row with index $M+r^*-1$. The diagonal elements to be added are all equal to $-\lambda-c\beta$. To the *right* of these diagonal elements we add an entry $c\beta\theta'$, except for the last row, where the entry $c\beta\theta'$ is placed in the column labeled $k^*$. In the row with index $M+r^*-1$ the entry $c\beta\theta'$ should appear immediately to the right, rather than to the left of the diagonal entry. All other added elements are zero.

The matrices $A_2$ and $A_{i2}$, $0 \leqq i \leqq r-1$, are similarly modified, with entries $\lambda$ and $c\beta\theta$ now playing the role of the quantities $-\lambda-c\beta$ and $c\beta\theta'$.

We see that in the present model, the matrices are no longer Jacobi matrices, but remain highly structured sparse matrices. The theoretical results in this paper do not depend on the detailed structure of the blocks in the partitioned matrix $P$, but particularly when the order of these blocks becomes large, their sparsity may be exploited to economize on the storage and processing time requirements of the algorithm.

**3. Quasi-birth-and-death process.** Consider an irreducible continuous parameter Markov chain with state space $\{(i, j);\ i \geqq 0,\ 0 \geqq j \leqq N\}$ and infinitesimal generator $P$ of the form (1).

Let us denote by $\mathbf{x}$ the vector of steady-state probabilities, associated to $P$, $\mathbf{x}P = \mathbf{0}$, $\mathbf{x}\mathbf{e} = 1$, and define the conservative stable matrix $A$ by $A = A_0 + A_1 + A_2$. We assume that $A$ is irreducible and denote by $\boldsymbol{\pi}$ its vector of steady-state probabilities, i.e., $\boldsymbol{\pi}A = \mathbf{0}$, $\boldsymbol{\pi}\mathbf{e} = 1$. Each component of $\boldsymbol{\pi}$ is strictly positive. In the tandem queue models considered here, $A$ will obviously be irreducible.

We partition $\mathbf{x}$ as $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \cdots)$, where each vector $\mathbf{x}_i$ has $N+1$ components. We shall examine below the existence of a solution of the form $\mathbf{x}_i = \mathbf{x}_{r-1}R^{i-r+1}$ for $i \geqq r-1$, where $R$ has a spectral radius strictly less than one ($\mathrm{sp}(R) < 1$). For such a solution to exist, we must have that

$$
\begin{aligned}
&\mathbf{x}_0 A_{0,1} + \mathbf{x}_1 A_{1,0} = \mathbf{0}, \\
&\mathbf{x}_i A_{i,2} + \mathbf{x}_{i+1} A_{i+1,1} + \mathbf{x}_{i+2} A_{i+2,0} = \mathbf{0} \quad \text{for } 0 \leqq i \leqq r-3, \\
&\mathbf{x}_{r-2} A_{r-2,2} + \mathbf{x}_{r-1} A_{r-1,1} + \mathbf{x}_r A_0 = \mathbf{0}, \\
&\mathbf{x}_{r-1} R^{i-r+1}(A_2 + RA_1 + R^2 A_0) = \mathbf{0} \quad \text{for } i \geqq r-1.
\end{aligned} \tag{2}
$$

We shall show that in the positive recurrent case, a strictly positive probability vector $\mathbf{x}$ of the stated form exists, for which the matrix $R$ is a nonnegative irreducible matrix of spectral sp $(R)$ less than one and such that $A_2 + RA_1 + R^2 A_0 = 0$.

We now have to make several technical assumptions that are satisfied for the models we consider:

(a) $A_1$ is nonsingular. By Wallace [22, Thm. 3.1], a sufficient condition is that $A_1\mathbf{e} < \mathbf{0}$ which means that from any state $(i, j)$, $i \geqq r$, it is possible to move in one step to a state $(i+1, j')$ or $(i-1, j')$. By Wallace [22, Lemma 3.4], $A_1^{-1}$ is a nonpositive matrix with strictly negative diagonal elements.

(b) The matrix $C_2 = -A_2 A_1^{-1}$, has at least one nonzero element in each row. A sufficient condition is that all diagonal entries of $A_2$ are strictly positive, which means that arrivals can occur when the system is in any state $(i, j)$, $i \geqq r$.

(c) If we define $C_0 = -A_0 A_1^{-1}$ and $C = C_0 + C_2$, we assume that $C$ is irreducible.

The equation $A_2 + RA_1 + R^2 A_0 = 0$, may now be rewritten in the form

$$
R = C_2 + R^2 C_0. \tag{3}
$$

LEMMA 1. *The matrix $C$ has a maximal eigenvalue equal to one with corresponding left and right eigenvector, respectively proportional to $\boldsymbol{\pi}$ and $A_1\mathbf{e}$.*

*Proof.* (a) $\boldsymbol{\pi}C = -\boldsymbol{\pi}(A_0+A_2)A_1^{-1} = \boldsymbol{\pi}(A_1-A)A_1^{-1} = \boldsymbol{\pi}$, since $\boldsymbol{\pi}A = \mathbf{0}$. (b) $CA_1\mathbf{e} = -(A_0+A_2)\mathbf{e} = (A_1-A)\mathbf{e} = A_1\mathbf{e}$, since $A\mathbf{e} = 0$.

Define the sequence $\{R(n), n \geqq 0\}$ of matrices as follows:

$$R(0) = 0, \qquad R(n+1) = C_2 + R(n)^2 C_0 \quad \text{for } n \geqq 0.$$

THEOREM 1. *If $\boldsymbol{\pi}A_0\mathbf{e} \leqq \boldsymbol{\pi}A_2\mathbf{e}$, the equation $R = C_2 + R^2C_0$ has a unique solution $R$ for which $R \geqq 0$, sp $(R) \leqq 1$. This solution is $\lim_{n\to\infty} R(n)$ and sp $(R) = 1$.*

*If $\boldsymbol{\pi}A_0\mathbf{e} > \boldsymbol{\pi}A_2\mathbf{e}$, the equation $R = C_2 + R^2C_0$ has a unique solution $R$ for which $R \geqq 0$, sp $(R) < 1$. This solution, $\lim_{n\to\infty} R(n)$, is the minimal solution to the equation and is irreducible.*

*Proof.* This theorem is proved by repeating almost verbatim the argument given in [14, Thms. 1 and 2, Lemmas 2, 3 and 4].

The irreducibility of the matrix $C$ is needed primarily to derive the equilibrium condition in the explicit form $\boldsymbol{\pi}A_0\mathbf{e} > \boldsymbol{\pi}A_2\mathbf{e}$. The requirement that $C_2$ has no vanishing rows entails the irreducibility of the matrix $R$. For the models discussed in this paper, these conditions are verified. It is easily seen that after a small number of iterations, the matrices $R(n)$ and hence also $R$ are strictly positive.

In [15], the existence of a matrix-geometric invariant vector is established without irreducibility conditions on the matrices arising in the partition of $P$. As in [14], the detailed proofs are given for stochastic matrices, but the translation to the case of infinitesimal generators is elementary. The algorithmic simplifications due to the reducibility of the matrix $R$ are discussed in a forthcoming monograph by the second author. Since in the present case, $R$ is positive, these issues are not germane to the discussion here.

Let $x^* = (x_0, x_1, \cdots, x_{r-1})$, and

$$P^* = \begin{Vmatrix} & A_{02} & \cdots & A_{r-3,1} & A_{r-2,2} \\ A_{01} & A_{11} & \cdots & A_{r-2,1} & A_{r-1,1}+RA_0 \\ A_{10} & A_{20} & \cdots & A_{r-1,0} & \end{Vmatrix}.$$

LEMMA 2. *$P^*$ is an infinitesimal generator.*

*Proof.* Since $P$ is an infinitesimal generator and $RA_0 \geqq 0$, all off-diagonal elements of $P^*$ are nonnegative.

To prove that $P^*\mathbf{e} = \mathbf{0}$, one needs only consider the last $N+1$ rows of $P^*$, since the other rows are identical to rows of $P$. However

$$\begin{aligned} A_{r-1,0}\mathbf{e} + (A_{r-1,1}+RA_0)\mathbf{e} &= -A_2\mathbf{e} + RA_0\mathbf{e} \\ &= -A_2\mathbf{e} + RA_0\mathbf{e} + \sum_{\nu=0}^{\infty} R^\nu(A_2 + RA_1 + R^2A_0)\mathbf{e} \\ &= R(I-R)^{-1}(A_0+A_1+A_2)\mathbf{e} = \mathbf{0}. \end{aligned}$$

LEMMA 3. *Since $A_{r-1,1}+RA_0$ is irreducible, $P^*$ is irreducible.*

*Proof.* The proof is straightforward.

THEOREM 2. *Under the assumption of Lemma 3 and $\boldsymbol{\pi}A_0\mathbf{e} > \boldsymbol{\pi}A_2\mathbf{e}$, let $R \geqq 0$ be the minimal solution of $R = C_2 + R^2C_0$. Let $\mathbf{x}^* = (\mathbf{x}_0^*, \mathbf{x}_1^*, \cdots, x_{r-1}^*)$ be a solution of $\mathbf{x}^*P^* =$*

**0**; *then* $\mathbf{x}^*$ *has components all of the same sign. Furthermore* $\mathbf{x}^*$ *may be normalized by*

$$\sum_{\nu=0}^{r-2} \mathbf{x}_\nu^* \mathbf{e} + \mathbf{x}_{r-1}^* (I-R)^{-1}\mathbf{e} = 1. \tag{4}$$

*The vector* $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \cdots)$ *with*

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}_i^* \quad \text{for } 0 \leqq i \leqq r-1, \\ \mathbf{x}_i &= \mathbf{x}_{r-1}^* R^{i-r+1} \quad \text{for } r-1 \leqq i, \end{aligned} \tag{5}$$

*is the unique, strictly positive steady-state probability vector of the matrix P.*

The proof is now obvious.

*Remark.* Since $R$ is irreducible and $\mathrm{sp}(R) < 1$, $(I-R)^{-1}$ exists and is strictly positive.

COROLLARY 1.

$$RA_0\mathbf{e} = A_2\mathbf{e}. \tag{6}$$

*Proof.*

$$R = C_2 + R^2 C_0 = -A_2 A_1^{-1} - R^2 A_0 A_1^{-1};$$

hence

$$RA_1\mathbf{e} = -A_2\mathbf{e} - R^2 A_0\mathbf{e}$$

and

$$\begin{aligned} -R(A_0+A_2)\mathbf{e} &= -A_2\mathbf{e} - R^2 A_0\mathbf{e}, \\ (I-R)A_2\mathbf{e} &= (I-R)RA_0\mathbf{e}. \end{aligned}$$

Since $I-R$ is nonsingular, formula (6) follows.

*Remark.* In the tandem queue models considered here, $A_0\mathbf{e}$ is the vector of rates of departure from unit I when all servers are busy. $A_2\mathbf{e}$ is the vector of rates of arrival to the queue in front of unit I. This corollary shows that $R$ plays a role similar to a traffic coefficient. In numerical computations, this relation serves usefully as an accuracy check on the evaluation of $R$.

**4. Explicit forms of the equilibrium condition.** For the specific versions $A$, $B$ and $C$ of the buffer model, the equilibrium condition $\boldsymbol{\pi} A_0\mathbf{e} > \boldsymbol{\pi} A_2\mathbf{e}$, may be explicitly written in terms of the parameters of the model. Although the analytic expressions of these explicit forms are complicated, they are are of the general form

$$\lambda < (1-\theta) r\alpha \Psi, \tag{7}$$

where $0 < \Psi < 1$, and $\Psi$ is a function of all the parameters of the model, except for $\lambda$ and $\theta$.

The quantity $(1-\theta)r\alpha$ is the critical input rate of a system consisting only of Unit I with a feedback probability $\theta$. The entire right hand side of (7) may be interpreted as the critical input rate $\lambda^*$ to the system under consideration. The dependence of $\lambda^*$ on the various parameters of the model provides us with a readily accessible means of comparing the effects of buffer size and unblocking rules. It must be borne in mind, however, that queues for which $\lambda$ is close to or equal to $\lambda^*$ will exhibit the typical, frequently undesirable, long-range fluctuations inherent in near-critical queues.

THEOREM 3. *The vector* $\boldsymbol{\pi}$ *and the equilibrium condition* $\boldsymbol{\pi}A_0\mathbf{e} > \boldsymbol{\pi}A_2\mathbf{e}$, *are given by*:

*For Model* A,

$$\pi_0 = \left[\sum_{j=0}^{c-1}\frac{1}{j!}\left(\frac{r\alpha}{\beta}\right)^j + \frac{c^c}{c!}\sum_{j=c}^{M}\left(\frac{r\alpha}{c\beta}\right)^j\right]^{-1},$$

$$\pi_j = \frac{1}{j!}\left(\frac{r\alpha}{\beta}\right)^j\pi_0 \quad \text{for } 1 \leqq j \leqq c, \tag{8}$$

$$\pi_j = \frac{c^c}{c!}\left(\frac{r\alpha}{c\beta}\right)^j\pi_0, \quad \text{for } c \leqq j \leqq M.$$

$$\lambda < (1-\theta)r\alpha\sum_{j=0}^{M-1}\pi_j = (1-\theta)r\alpha(1-\pi_M). \tag{9}$$

*For Model* B ($2 \leqq r^* \leqq r$),

$$\pi_0 = \left[\sum_{j=0}^{c-1}\frac{1}{j!}\left(\frac{r\alpha}{\beta}\right)^j + \frac{c^c}{c!}\sum_{j=c}^{M}\left(\frac{r\alpha}{c\beta}\right)^j + \frac{c^c}{c!}\sum_{j=M+1}^{M+r^*-1}\left(\frac{r\alpha}{c\beta}\right)^j\prod_{\nu=1}^{j-M}\left(1-\frac{\nu}{r}\right)\right]^{-1},$$

$$\pi_j = \frac{1}{j!}\left(\frac{r\alpha}{\beta}\right)^j\pi_0 \quad \text{for } 1 \leqq j \leqq c,$$

$$\pi_j = \frac{c^c}{c!}\left(\frac{r\alpha}{c\beta}\right)^j\pi_0 \quad \text{for } c \leqq j \leqq M, \tag{10}$$

$$\pi_j = \frac{c^c}{c!}\left(\frac{r\alpha}{c\beta}\right)^j\prod_{\nu=1}^{j-M}\left(1-\frac{\nu}{r}\right)\pi_0 \quad \text{for } M+1 \leqq j \leqq M+r^*-1.$$

$$\lambda < (1-\theta)r\alpha\left[\sum_{j=0}^{M-1}\pi_j + \sum_{j=1}^{r^*-1}\left(1-\frac{j}{r}\right)\pi_{M+j-1}\right]$$

$$= (1-\theta)r\alpha\left(1-\pi_{M+r^*-1}-\sum_{j=1}^{r^*-1}\frac{j}{r}\pi_{M+j-1}\right). \tag{11}$$

*For Model* C ($1 \leqq r^* \leqq r$, $c \leqq k^* \leqq M+r^*-2$), *we shall give detailed formulas only for the most useful case where* $c \leqq k^* \leqq M-1$. *The equations for the other cases are entirely similar. We denote by* $\tilde{\pi}_{k^*+1}, \cdots, \tilde{\pi}_{M+r^*-2}$, *the components of* $\boldsymbol{\pi}$ *corresponding to the indices* $j = k^*+1, \cdots, M+r^*-2$. *The explicit formulas for the components of* $\boldsymbol{\pi}$ *are uninspiringly complicated, but their numerical values may readily be computed by solving the linear equations*

$$\pi_j = \frac{r\alpha}{j\beta}\pi_{j-1} \quad \text{for } 1 \leqq j \leqq c,$$

$$\pi_j = \frac{r\alpha}{c\beta}\pi_{j-1} \quad \text{for } c \leqq j \leqq k^*,$$

$$\pi_j = \frac{r\alpha}{c\beta}\pi_{j-1} - \tilde{\pi}_{k^*+1}, \quad \text{for } k^*+1 \leqq j \leqq M, \tag{12}$$

$$\pi_{M+j} = \left(1-\frac{j}{r}\right)\left(\frac{r\alpha}{c\beta}\right)\pi_{M+j-1} - \tilde{\pi}_{k^*+1}, \quad \text{for } 1 \leqq j \leqq r^*-2,$$

$$\pi_{M+r^*-1} = \tilde{\pi}_{M+r^*-2} = \cdots = \tilde{\pi}_{k^*+1} = \frac{c^c}{c!}\left(\frac{r\alpha}{c\beta}\right)^{M+r^*-1}\prod_{\nu=1}^{r^*-1}\left(1-\frac{\nu}{r}\right)\Phi^{-1}\pi_0,$$

*where*

$$\Phi = \sum_{h=1}^{r^*-2} \left(\frac{r\alpha}{c\beta}\right)^h \prod_{\nu=1}^{h} \left(1 - \frac{r^* - \nu}{r}\right) + \prod_{\nu=1}^{r^*-1} \left(1 - \frac{\nu}{r}\right) \sum_{h=0}^{M-k^*-1} \left(\frac{r\alpha}{c\beta}\right)^{h+r^*-1}.$$

*Finally $\pi_0$ is obtained from the normalizing condition $\boldsymbol{\pi}\mathbf{e} = 1$.*

$$\lambda < (1-\theta) r\alpha \left[ \sum_{j=0}^{M-1} \pi_j + \sum_{j=1}^{r^*-1} \left(1 - \frac{j}{r}\right) \pi_{M+j-1} \right]. \tag{13}$$

*Proof.* We shall only sketch the proof for Model C. The equations $\boldsymbol{\pi}A = \mathbf{0}$ may be written as

$$\begin{aligned}
&-r\alpha\pi_0 + \beta\pi_1 = 0,\\
&r\alpha\pi_{j-1} - (r\alpha + j\beta)\pi_j + (j+1)\beta\pi_{j+1} = 0 \quad \text{for } 1 \leqq j \leqq c-1,\\
&r\alpha\pi_{j-1} - (r\alpha + c\beta)\pi_j + c\beta\pi_{j+1} + \delta_{j,k^*} c\beta\tilde{\pi}_{k^*+1} = 0 \quad \text{for } c \leqq j \leqq M-1,\\
&(r-j)\alpha\pi_{M+j-1} - [(r-j-1)\alpha + c\beta]\pi_{M+j} + c\beta\pi_{M+j+1} = 0 \quad \text{for } 0 \leqq j \leqq r^*-3,\\
&(r-r^*+2)\alpha\pi_{M+r^*-3} - [(r-r^*+1)\alpha + c\beta]\pi_{M+r^*-2} = 0 \quad \text{for } r^* \geqq 2\\
&(r-r^*+1)\alpha\pi_{M+r^*-2} - c\beta\pi_{M+r^*-1} = 0,\\
&\pi_{M+r^*-1} = \tilde{\pi}_{M+r^*-2} = \cdots = \tilde{\pi}_{k^*+1}.
\end{aligned} \tag{14}$$

These are clearly equivalent to

$$\begin{aligned}
&r\alpha\pi_{j-1} = \min(j, c)\beta\pi_j \quad \text{for } 1 \leqq j \leqq k^*,\\
&r\alpha\pi_{j-1} = c\beta\pi_j + c\beta\tilde{\pi}_{k^*+1} \quad \text{for } k^*+1 \leqq j \leqq M,\\
&(r-j)\alpha\pi_{M+j-1} = c\beta\pi_{M+j} + c\beta\tilde{\pi}_{k^*+1} \quad \text{for } 1 \leqq j \leqq r^*-2,\\
&(r-r^*+1)\alpha\pi_{M+r^*-2} = c\beta\pi_{M+r^*-1},\\
&\pi_{M+r^*-1} = \tilde{\pi}_{M+r^*-2} - \cdots = \tilde{\pi}_{k^*+1}.
\end{aligned} \tag{15}$$

Equating the expression recursively computed for $\pi_{M+r^*-1}$ with $\tilde{\pi}_{k^*+1}$, we obtain the stated formula relating $\tilde{\pi}_{k^*+1}$ and $\pi_0$.

The inequality $\boldsymbol{\pi}A_0\mathbf{e} > \boldsymbol{\pi}A_2\mathbf{e}$ is equivalent to

$$\begin{aligned}
\lambda < &\sum_{j=1}^{c} (r\alpha\pi_{j-1} - j\beta\theta\pi_j) + \sum_{j=c+1}^{M} (r\alpha\pi_{j-1} - c\beta\theta\pi_j)\\
&+ \sum_{j=1}^{r^*-1} [(r-j)\pi_{M+j-1} - c\beta\theta\pi_{M+j}] - c\beta\theta(M + r^* - k^* - 2),
\end{aligned}$$

and by using (14), we obtain formula (13).

*Remarks.* 1. It is preferable not to write the geometric sums in (8) and (10) in closed forms, so that we do not have to write separate expressions for the case where $r\alpha = c\beta$.

2. For $r = c = 1$, and $\lambda$ chosen, without loss of generality, to be equal to one, we obtain for Model A that the queue will be stable if and only if

$$\sum_{\nu=0}^{M} \left(\frac{\alpha}{\beta}\right)^{\nu} < (1-\theta)\alpha \sum_{\nu=0}^{M-1} \left(\frac{\alpha}{\beta}\right)^{\nu}.$$

This agrees, after elementary manipulations, with the conditions (2) for $\alpha \neq \beta$, and (3) for $\alpha = \beta$, stated in Theorem 2 of A. G. Konheim and M. Reiser [10, p. 334]. A minor correction is, however, needed in the statement of that theorem. Condition (1), i.e.

$(1-\theta)\min(\alpha,\beta)<1$, is claimed to be the equilibrium condition for the system where $M=\infty$. As it is implied by one of the other conditions, depending on whether $\alpha\neq\beta$ or $\alpha=\beta$, its inclusion in the stability condition for finite $M$ is clearly inappropriate.

**5. The system considered at service completions in unit I.** Upon considering the numbers of I- and II-customers immediately after service completions in unit I, we obtain a Markov chain with the states $(i, j)$, where $i\geqq 0$ and $j=1,\cdots,M+r^*-1$. In the interest of notational simplicity, we shall preserve the earlier state space, but note that since service completions during full blocking in unit I are impossible, the states with $j=0$ and the additional states corresponding to full blocking are ephemeral. Our formulas will correctly assign "steady-state probabilities" equal to zero to all such states and it will not be necessary to adjust the dimensions of the matrices which are involved.

THEOREM 4. *The stationary probability vector* $\mathbf{z}=(\mathbf{z}_0,\mathbf{z}_1,\cdots)$ *of the embedded Markov chain at service completions in unit* I *is given by*

$$\begin{aligned}&\mathbf{z}_k=\tau\mathbf{x}_{k+1}A_{k+1,0}\quad \text{for } 0\leqq k\leqq r-1,\\ &\mathbf{z}_k=\tau\mathbf{x}_{k+1}A_0=\tau\mathbf{x}_{r-1}R^{k-r+2}A_0\quad \text{for } k\geqq r-1,\end{aligned}\tag{16}$$

*where* $\tau$ *is given by*

$$\tau=\left[\sum_{i=1}^{r-1}\mathbf{x}_iA_{i,0}\mathbf{e}+\mathbf{x}_{r-1}R(I-R)^{-1}A_0\mathbf{e}\right]^{-1}.\tag{17}$$

*The zero components in the vectors* $\mathbf{z}_k$ *are ignored.*

*Proof.* The formulas (16) are readily obtained by a conditioning argument for the elementary probabilities. The quantity $\tau^{-1}dt$ is clearly the elementary probability that a service completion occurs in $(t, t+dt)$ and the components of $\mathbf{x}_{k+1}A_{k+1,0}dt$ or $\mathbf{x}_{k+1}A_0dt$ are elementary probabilities of transitions of the type $(k+1, j)\to(k, j')$.

COROLLARY 2. *The components with* $j=M+\nu$, $0\leqq\nu\leqq r^*-1$, *of the vector*

$$\mathbf{Z}=\sum_{k=0}^{\infty}\mathbf{z}_k=\tau\left[\sum_{i=1}^{r-1}\mathbf{x}_iA_{i0}+\mathbf{x}_{r-1}R(I-R)^{-1}A_0\right],\tag{18}$$

*yield the stationary probabilities that upon completion of a service in unit* I, $\nu+1$ *servers are blocked in unit* I.

It would satisfy higher standards of rigor to set up explicitly the transition probability matrix of the embedded Markov chain and to verify that $z$ is indeed its invariant vector. In order to avoid introducing a large amount of extra notation, we shall only do this for the case $r=1$. In the process we shall also obtain a different formula for $\mathbf{Z}$, which is also easily implemented and therefore provides us with an accuracy check in numerical computations.

Completely elementary probability arguments yield that for $r=1$, the transition probability matrix of the embedded chain is given by

$$\bar{P}=\begin{vmatrix}\bar{B}_0 & \bar{B}_1 & \bar{B}_2 & \bar{B}_3 & \bar{B}_4 & \cdots\\ \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \bar{A}_3 & \bar{A}_4 & \cdots\\ 0 & \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \bar{A}_3 & \cdots\\ 0 & 0 & \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \cdots\\ 0 & 0 & 0 & \bar{A}_0 & \bar{A}_1 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{vmatrix},\tag{19}$$

where $\bar{A}_n=(-A_1^{-1}A_2)^n(-A_1^{-1}A_0)$, and $\bar{B}_n=(-A_{01}^{-1}A_2)\bar{A}_n$, for $n\geqq 0$.

The steady-state equations $\mathbf{z}\bar{P}=\mathbf{z}$, will be satisfied by the vectors defined in formula (16) provided that

$$\mathbf{x}_0 R^{k+1} A_0 = \mathbf{x}_0 R A_0(-A_{01}^{-1}A_2)\bar{A}_k + \sum_{\nu=1}^{k+1} \mathbf{x}_0 R^{\nu+1} A_0 \bar{A}_{k+1-\nu} \text{ for } k \geqq 0. \tag{20}$$

Since $\mathbf{x}_0 = \mathbf{x}_0 R(-A_0 A_{01}^{-1})$, and using the explicit form of the matrices $\bar{A}_n$, this equation may be equivalently rewritten as

$$\mathbf{x}_0\left[R^{k+1} - (-A_2A_1^{-1})^{k+1} - \sum_{\nu=1}^{k+1} R^{\nu+1}(-A_0A_1^{-1})(-A_2A_1^{-1})^{k+1-\nu}\right]A_0 = \mathbf{0} \text{ for } k \geqq 0. \tag{21}$$

In order to see that the matrix square brackets is zero, write the equation $R=(-A_2A_1^{-1})+R^2(-A_0A_1^{-1})$, $k+1$ times. Multiply the $\nu$th equation on the left by $R^{\nu-1}$ and on the right by $(-A_2A_1^{-1})^{k-\nu+1}$ and sum up.

Finally the expression for $\tau$ is obtained from the normalizing equation $\mathbf{z}\mathbf{e}=1$.

COROLLARY 3. *In the case $r=1$, the vector $\mathbf{Z}$ is also given by*

$$\begin{aligned}\mathbf{Z} &= (\boldsymbol{\pi}A_0\mathbf{e})^{-1}\boldsymbol{\pi}A_0 \\ &\quad - \tau\mathbf{x}_0(A_{01}+A_2)(A_1+A_2)^{-1}A_0[A-(\boldsymbol{\pi}A_0\mathbf{e})^{-1}A_0\Pi A_0]^{-1}(A_1+A_2),\end{aligned} \tag{22}$$

*where $\Pi$ is a matrix with $M+1$ identical rows given by $\boldsymbol{\pi}$.*

*Proof.* Adding the steady-state equations for the matrix $\bar{P}$, we obtain

$$\mathbf{Z}[I+(A_1+A_2)^{-1}A_0] = \mathbf{z}_0(-A_{01}^{-1}A_2)[-(A_1+A_2)^{-1}A_0] - \mathbf{z}_0[-(A_1+A_2)^{-1}A_0]. \tag{23}$$

The matrix $-(A_1+A_2)^{-1}A_0$ is a stochastic matrix, whose first column is zero. All other elements are strictly positive. It has the left invariant vector $\boldsymbol{\pi}A_0$, whose first component is zero and all other components strictly positive. It now follows readily from the theory of finite Markov chains that the matrix

$$I+(A_1+A_2)^{-1}A_0+(\boldsymbol{\pi}A_0\mathbf{e})^{-1}\Pi A_0 = (A_1+A_2)^{-1}[A-(\boldsymbol{\pi}A_0\mathbf{e})^{-1}A_0\Pi A_0], \tag{24}$$

is nonsingular. Adding $(\boldsymbol{\pi}A_0\mathbf{e})^{-1}\mathbf{Z}\Pi A_0 = (\boldsymbol{\pi}A_0\mathbf{e})^{-1}\boldsymbol{\pi}A_0$ to both sides of (23) and replacing $\mathbf{x}_0R(-A_0A_{01}^{-1})$ by $\mathbf{x}_0$, we obtain the stated formula after routine matrix manipulations.

We note that the formula assigns the correct value zero to the first component of $\mathbf{Z}$. Verifying that the result so obtained agrees with $\mathbf{Z}=\tau\mathbf{x}_0R(I-R)^{-1}A_0$ (Corollary 2) provides a partial check on numerical computations.

**6. Remarks on numerical computations.** The solution, presented here, lends itself to a ready numerical implementation. Efficient programming, which takes the high degree of sparsity of the transition probability matrix into account, results in substantial savings in memory storage and execution times. This is particularly worthwhile when the program is to be used to study the design and control aspects of the model. In such studies, one or more parameters of the model need to vary over a range of values, which may require a substantial number of executions of the program. In view of the complicated dependence of the model on each of its parameters, detailed numerical studies appear to be the only way of obtaining the hard qualitative information needed in problems of design and optimization.

The first step, after ascertaining that the queue is stable, is to compute the matrix $R$. This may be done by successive substitutions in the equation $R=-A_2A_1^{-1}-R^2A_0A_1^{-1}$, starting with $R=0$. The relation $RA_0\mathbf{e}=A_2\mathbf{e}$, proved in Lemma 2, serves as an accuracy check.

If $r>1$, the vectors $\mathbf{x}_0, \cdots, \mathbf{x}_{r-1}$, are computed by solving the system of linear equations, discussed in Theorem 2. Since the number of equations in that system may be very large, it is again desirable to take the special structure of its coefficient matrix into account. This may be done as follows. In the system

$$\begin{aligned} &\mathbf{x}_0 A_{01} + \boldsymbol{x}_1 A_{10} = \mathbf{0}, \\ &\mathbf{x}_{\nu-1} A_{\nu-1,2} + \mathbf{x}_\nu A_{\nu,1} + \mathbf{x}_{\nu+1} A_{\nu+1,0} = \mathbf{0} \quad \text{for } 1 \leqq \nu \leqq r-2, \\ &\mathbf{x}_{r-2} A_{r-2,2} + \mathbf{x}_{r-1}(A_{r-1,1} + RA_0) = \mathbf{0}, \end{aligned} \tag{25}$$

the matrices $A_{\nu,2}$, $0 \leqq \nu \leqq r-2$, are clearly nonsingular, so that, using all but the first equation, we can write the vectors $\mathbf{x}_0, \cdots, \mathbf{x}_{r-2}$, as $\mathbf{x}_\nu = \mathbf{x}_{r-1} C_\nu^*$, $0 \leqq \nu \leqq r-2$, where the matrices $C_\nu^*$ are readily computed. The first equation now yields

$$\mathbf{x}_{r-1}(C_0^* A_{01} + C_1^* A_{10}) = \mathbf{0}, \tag{26}$$

which together with the normalizing condition

$$x_{r-1}\left[\sum_{\nu=0}^{r-2} C_\nu^* \mathbf{e} + (I-R)^{-1}\mathbf{e}\right] = 1, \tag{27}$$

uniquely determines the vector $\mathbf{x}_{r-1}$ and hence also the vectors $\mathbf{x}_0, \cdots, \mathbf{x}_{r-2}$.

If there is need to economize on memory storage, as when $r$ and the order of the matrices are large, we can avoid storing the matrices $C_\nu$, as they may be evaluated recursively. This can be done using only three arrays of size $N \times N$ and one linear array of length $N$. In the latter the vector $\sum_{\nu=0}^{r-2} C_\nu^* \mathbf{e}$ is accumulated. This does not significantly increase the processing time, as the systems of equations $\mathbf{x}_\nu A_{\nu 2} = \mathbf{d}$, where $\mathbf{d}$ is a known vector, are particularly easy to solve.

The simplifications, discussed above, are particularly striking when $\theta = 0$ (no feedback) as the matrices $A_{\nu 2}$, $0 \leqq \nu \leqq r-2$, are then scalar matrices.

The remaining computations of the vector $\mathbf{x}$, of various moments and of the marginal queue length densities, as well as the blocking probabilities are now entirely routine.

## REFERENCES


[1] K. L. ARYA (1972), *Study of a network of serial and non-serial servers with phase type service and finite queueing space*, J. Appl. Probability, 9, pp. 198–201.

[2] B. AVI-ITZHAK AND M. YADIN (1965), *A sequence of two servers with no intermediate queue*, Management Sci., 11, pp. 553–564.

[3] A. B. CLARKE (1978), *Tandem queues with blocking*. Abstract 160–35, Bull. Inst. Math. Stat., 7, p. 34.

[4] F. G. FOSTER AND H. G. PERROS (undated), *On the blocking process in queue networks*, Manuscript, Dept. of Statistics, Trinity College, Dublin, Ireland.

[5] I. HANNIBALSSON (1975), *Networks of queues with delayed feedback*, Tech. Rep. 75–10, Dept. of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI.

[6] J. M. HARRISON (1977), *The diffusion approximation for tandem queues in heavy traffic*, Tech. Rept. 45, Dept. of Operations Research, Stanford University, Stanford, CA.

[7] D. K. HILDEBRAND (1967), *Stability of finite queue, tandem server systems*, J. Appl. Probability, 4, pp. 571–583.

[8] F. S. HILLIER AND R. W. BOLING (1971), *Finite queues in series with exponential or Erlang service times—A numerical approach*, Operations Research, 15, pp. 286–303.

[9] G. C. HUNT (1956), *Sequential arrays of waiting lines*, Operations Res., 4, pp. 674–683.

[10] A. G. KONHEIM AND M. REISER (1976), *A queueing model with finite waiting room and blocking*, J. Assoc. Comput. Mach., 23, pp. 328–341.

[11] M. F. NEUTS (1968), *Two queues in series with a finite intermediate waitingroom*, J. Appl. Probability, 5, pp. 123–142.

[12] ——— (1970), *Two servers in series, treated in terms of a Markov renewal branching process*, Advances in Appl. Probability, 2, pp. 110–149.
[13] ——— (1978), *The M/M/1 queue with randomly varying arrival and service rates*, Opsearch, 15, pp. 139–157.
[14] ——— (1978), *Markov chains with applications in queueing theory, which have a matrix-geometric invariant vector*, Advances in Appl. Probability, 10, pp. 185–212.
[15] ——— (1980), *The probabilistic significance of the rate matrix in matrix-geometric invariant vectors*, J. Appl. Probability, to appear.
[16] G. F. NEWELL (1975), *Approximate behavior of tandem queues*, Chaps. I–IV, Research Report, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA.
[17] ——— (1977), *Approximate behavior of tandem queues*, Chaps. V–VIII, Research Report, Institute of Transportation Studies, University of California, Berkeley, CA.
[18] N. U. PRABHU (1966), *Transient behavior of a tandem queue*, Management Sci., 13, pp. 631–639.
[19] M. REISER AND A. G. KONHEIM (1978), *Finite capacity queueing systems with applications in computer modeling*, SIAM J. Comput., 7, pp. 210–229.
[20] O. P. SHARMA (1973), *A model for queues in series*, J. Appl. Probability, 10, pp. 691–696.
[21] T. SUZUKI (1964), *On a tandem queue with blocking*, J. Operations Res. Soc. Japan, 6, pp. 137–157.
[22] V. WALLACE (1969), *The solution of quasi birth and death processes arising from multiple access computer systems*, Ph.D. thesis, Tech. Rept. No. 07742-6-T, Systems Engineering Lab, University of Michigan, Ann Arbor, MI.
[23] B. WONG, W. GRIFFIN, AND R. L. DISNEY (1977), *Two finite M/M/1 queues in tandem: A matrix solution for the steady state*, Opsearch, 14, pp. 1–18.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0014 $01.00/0

# ON CONSTRUCTION OF MATRICES WITH DISTINCT SUBMATRICES*

SHARAD V. KANETKAR† AND MEGHANAD D. WAGH‡

**Abstract.** Given $N$, $M$, $t$ and $s$, a method of generating an $N \times M$ binary matrix such that every nonzero $t \times s$ binary pattern occurs exactly once as its submatrix is presented. This construction is based upon a systematic filling of the matrix with a maximal length recurrent sequence and gives several new solutions yet unreported.

**1. Introduction.** In this paper we consider the problem of construction of an $N \times M$ binary matrix $A$ such that any $t \times s$ nonzero binary pattern occurs exactly once as its submatrix. Similar problems have been attempted earlier by various authors.

Reed and Stewart [5] considered the existence of $A$ given only $t$ and $s$. Gordon [2] later extended their result and showed that given any $t$ and $s$, one can always find $N$ and $M$, $N > t$, $M > s$ such that all $t \times s$ submatrices (in the toroidal sense) in $A$ are distinct. $A$ is then called a perfect map. All the $t \times s$ nonzero binary patterns are not necessarily the submatrices of a perfect map. However, a perfect map with parameters $M = 2^s - 1$ and $N = (2^{st} - 1)/M$ and containing *all* the $t \times s$ nonzero binary patterns was exhibited in [2]. When $N$ and $M$ are relatively prime, a pseudorandom array also gives a perfect map with the same parameters [3].

The toroidal perfect maps of [2], [3] and [5] can be easily converted into nontoroidal ones by repeating the first $t-1$ rows after the last row and the first $s-1$ columns after the last column. In this paper, we will be concerned only with $N \times M$ nontoroidal perfect map $A$ in which every nonzero binary $t \times s$ pattern occurs exactly once as a submatrix. Obviously, the four parameters are then related as

$$(M - s + 1)(N - t + 1) = 2^{st} - 1. \tag{1.1}$$

Banerji [1] has recently described a procedure of designing $A$ when (i) $M = s$ and (ii) $M = 2^s + s - 2$. Note that the required matrix $A$ when $M = 2^s + s - 2$ was also obtained earlier by Gordon [2].

In this paper, we give a criterion for filling up the matrix $A$ with a maximal length recurrent sequence (MLRS) such that $A$ will have the required property. Four schemes have been described which satisfy the criterion and hence generate $A$ for all the earlier known cases and for several new ones. This criterion also enables one to construct $A$ for any $M$, $N$, $s$ and $t$ satisfying (1.1). We have included here the solution to the problem (for all the possible parameter combinations with $st \leqq 15$) obtained by a computer search made easy with the help of the criterion.

**2. Preliminaries.** A linear recurrent sequence $\{x_i\}$ of the elements of $GF(q)$, ($q$: a prime power) of period $q^n - 1$ may be obtained from the recurrence relation

$$x_i = a_1 x_{i-1} + a_2 x_{i-2} + \cdots + a_n x_{i-n} \tag{2.1}$$

over $GF(q)$ with arbitrary nonzero initial condition if the constants $a_1, a_2, \cdots, a_n \in GF(q)$ are chosen such that the polynomial

$$x^n - a_1 x^{n-1} - a_2 x^{n-2} - \cdots - a_n \tag{2.2}$$

is primitive over $GF(q)$. We will use the following property of this maximal length recurrent sequence (MLRS).

* Received by the editors November 13, 1978, and in revised form July 13, 1979.

† Computer Centre, Indian Institute of Technology, Bombay 400 076, India.

‡ Department of Electrical Engineering, Indian Institute of Technology, Bombay 400 076, India. Now at Department of Electrical Engineering, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal H3G 1M8, Canada.

LEMMA 1. *Let $\beta$ be the root of the polynomial* (2.2) *and $i_1, i_2, \cdots, i_n$, any n integers such that $\beta^{i_1}, \beta^{i_2}, \cdots, \beta^{i_n}$ are linearly independent over GF(q). Then the n-tuple $(x_{i+i_1}, x_{i+i_2}, \cdots, x_{i+i_n})$ assumes all the nonzero values exactly once in the range $0 \leqq i \leqq q^n - 2$.*

*Proof.* Solution of (2.1) can be expressed as [6]

$$x_i = \mathrm{Tr}\,(b\beta^i),$$

where Tr denotes the trace function

$$\mathrm{Tr}\,(\alpha) = \alpha + \alpha^q + \alpha^{q^2} + \cdots + \alpha^{q^{n-1}}$$

from $GF(q^n)$ onto $GF(q)$ and $b \in GF(q^n)$ is determined by the initial conditions. Then

$$\begin{bmatrix} x_{i+i_1} \\ x_{i+i_2} \\ \vdots \\ x_{i+i_n} \end{bmatrix} = \begin{bmatrix} \beta^{i_1} & \beta^{i_1 q} \cdots \beta^{i_1 q^{n-1}} \\ \beta^{i_2} & \beta^{i_2 q} \cdots \beta^{i_2 q^{n-1}} \\ \beta^{i_n} & \beta^{i_n q} \cdots \beta^{i_n q^{n-1}} \end{bmatrix} \begin{bmatrix} b\beta^i \\ (b\beta^i)^q \\ (b\beta^i)^{q^{n-1}} \end{bmatrix}.$$

The matrix on the right-hand side is nonsingular over $GF(q)$ as the elements in the first column are linearly independent by assumption. Thus there is a one-one correspondence between the $n$-tuple $(x_{i+i_1}, x_{i+i_2}, \cdots, x_{i+i_n})$ and the quantity $b\beta^i$. But as $\beta$ is the primitive element of $GF(q^n)$, $b\beta^i$ and hence $(x_{i+i_1}, x_{i+i_2}, \cdots, x_{i+i_n})$ takes all the possible $q^n - 1$ nonzero values as $i$ runs over $0 \leqq i \leqq q^n - 2$.

Since we are interested in binary matrices we will restrict ourselves to $q = 2$. However it should be mentioned that the methods developed in this paper can be generalized to the case of matrices with $q$ symbols.

Consider an MLRS $\{x_i\}$ of period $2^{st} - 1$ generated by (2.1) with $n = st$. We now state the central result of this paper.

THEOREM 1. *If A is filled as*

$$A(u, v) = x_{f(u,v)}, \qquad 0 \leqq u \leqq N-1, \quad 0 \leqq v \leqq M-1,$$

*such that*

(C1) *$f$ is linear in $u$ and $v$;*

(C2) *when $u$ and $v$ are restricted to $0 \leqq u \leqq N - t, 0 \leqq v \leqq M - s$, $f(u, v)$ are all distinct modulo $2^{st} - 1$;*

(C3) *$\beta^{f(u,v)}, 0 \leqq u \leqq t-1, 0 \leqq v \leqq s-1$ are all linearly independent over GF(2) where $\beta$ is the root of* (2.2) *with $n = st$;*

*then each binary $t \times s$ pattern occurs as a submatrix of A exactly once.*

*Proof.* Denoting $f(u, v)$, $0 \leqq u \leqq t-1$, $0 \leqq v \leqq s-1$ by $i_1, i_2, \cdots, i_{st}$, it is obvious from (C1) that any $t \times s$ submatrix in $A$ with its left-hand top corner at $(u, v)$ has elements

$$x_{i+i_1}, x_{i+i_2}, \cdots, x_{i+i_{st}} \quad \text{where } i = f(u, v).$$

Further, as $u$, $v$ run over $0 \leqq u \leqq N - t$ and $0 \leqq v \leqq M - s$, (i.e., all possible coordinate values taken by the left hand top corners of $t \times s$ submatrices), $i$ runs over 0 to $2^{st} - 2$ because of (C2) and (1.1). Finally, from (C3), $\beta^{i_1}, \beta^{i_2}, \cdots, \beta^{i_{st}}$ are linearly independent over GF(2) and hence an application of Lemma 1 gives the required result.

**3. Generation of $A$ matrix.** Several schemes to fill up $A$ to satisfy the conditions (C1)–(C3) may be given.

*Scheme* 1.

$$f(u, v) = u + tv, \qquad 0 \leqq u \leqq 2^{st} + t - 3, \quad 0 \leqq v \leqq s - 1$$

generates a matrix $A$ with $N = 2^{st} + t - 2$ and $M = s$. Here, (C1) is obvious. To check (C2), note that for $0 \leqq u \leqq N - t$, $0 \leqq v \leqq M - s = 0$, $f(u, v) = u$ and therefore in this range all $f(u, v)$ are distinct modulo $2^{st} - 1$. Finally, the set $\{\beta^{f(u,v)} | 0 \leqq u \leqq t - 1, 0 \leqq v \leqq s - 1\}$ is $\{1, \beta, \beta^2, \cdots, \beta^{st-1}\}$ elements of which are necessarily linearly independent over $GF(2)$ giving (C3). The generated matrix $A$ will then have the desired properties by Theorem 1. This leads to Banerji's case (i).

The mappings

$$f(u, v) = (M - s + 1)u + v \qquad (H \text{ mapping})$$

$$f(u, v) = u + (N - t + 1)v \qquad (V \text{ mapping})$$

$$0 \leqq u \leqq N - 1, \quad 0 \leqq v \leqq M - 1,$$

obviously satisfy (C1). In the case of $H$ mapping, when $u$ and $v$ are restricted to $0 \leqq u \leqq N - t$, $0 \leq v \leqq M - s$, one gets $0 \leqq f(u, v) \leqq 2^{st} - 2$ by using (1.1). Thus, if in this range $f(u_1, v_1) \equiv f(u_2, v_2) \pmod{2^{st} - 1}$, then $(M - s + 1)(u_1 - u_2) = (v_2 - v_1)$. But, $M - s + 1$ cannot divide $v_2 - v_1$ (as $0 \leqq v_1, v_2 \leqq M - s$) unless $v_2 = v_1$ and in that case $u_1$ also equals $u_2$. Thus $f(u, v)$ are distinct modulo $2^{st} - 1$ in this range showing that (C2) is satisfied. Similarly, $V$ mapping also can be shown to satisfy (C2).

We now present three more schemes based on $H$ and $V$ mappings which satisfy (C3).

*Scheme* 2. When $t = 1$, choosing $H$ mapping, the set $\{\beta^{f(u,v)} | 0 \leqq u \leqq t - 1 = 0, 0 \leqq v \leqq s - 1\}$ is $\{1, \beta, \beta^2, \cdots, \beta^{s-1}\}$. Its elements are linearly independent over $GF(2)$ as $\beta$ is the primitive element of $GF(2^{st})$. Thus (C3) is satisfied and the matrix generated will have the required properties.

*Scheme* 3. When $M = 2s - 1$, using $H$ mapping, $f(u, v) = su + v$. Then the set $\{\beta^{f(u,v)} | 0 \leqq u \leqq t - 1, 0 \leqq v \leqq s - 1\} = \{1, \beta, \beta^2, \cdots, \beta^{st-1}\}$ has elements which are linearly independent over $GF(2)$ as $\beta$ is the primitive element of $GF(2^{st})$. Thus (C3) is satisfied and one gets the matrix with the required properties.

*Scheme* 4. Let $z = (2^{st} - 1)/(2^s - 1)$ and $j$ any integer satisfying $j | (2^s - 1)$ and

$$\frac{2^s - 1}{j} \nmid (2^d - 1), \qquad 0 < d < s. \tag{3.1}$$

When $A$ has dimensions $N = zj + t - 1$ and $M = (2^s - 1)/j + s - 1$, one may use $V$ mapping. Then $f(u, v) = u + zjv$. To check (C3) one should prove the linear independence over $GF(2)$ of the elements of $\{\beta^{u+zjv} | 0 \leqq u \leqq t - 1, 0 \leqq v \leqq s - 1\}$. Note that $\beta^z \in GF(2^s)$ and (3.1) implies that $\beta^{zj}$ does not belong to any subfield of $GF(2^s)$. In other words, $1, \beta^{zj}, \beta^{2zj}, \cdots, \beta^{(s+1)zj}$ are linearly independent over $GF(2)$ because otherwise $\beta^{zj}$ will satisfy a polynomial of degree $\leqq s - 1$ over $GF(2)$ implying $\beta^{zj}$ belongs to a proper subfield of $GF(2^s)$. Further, $1, \beta, \beta^2, \cdots, \beta^{t-1}$ are also linearly independent over $GF(2^s)$ because $\beta$ cannot satisfy a polynomial of degree less than $t$ over $GF(2^s)$. Now if a linear combination of $\beta^{u+zjv}$ is equal to zero, then

$$0 = \sum_{u=0}^{t-1} \sum_{v=0}^{s-1} a_{uv} \beta^{u+zjv}$$

$$= \sum_{u=0}^{t-1} \left( \beta^u \sum_{v=0}^{s-1} a_{uv} \beta^{zjv} \right), \qquad a_{uv} \in GF(2).$$

The result of the inner summation belongs to $GF(2^s)$. But as $\{\beta^u | 0 \leqq u \leqq t-1\}$ are linearly independent over $GF(2^s)$, one has from this

$$0 = \sum_{u=0}^{s-1} a_{uv} \beta^{zjv}$$

which, from the linear independence of $\{\beta^{zjv} | 0 \leqq v \leqq s-1\}$ over $GF(2)$ gives $a_{uv} = 0, 0 \leqq u \leqq t-1, 0 \leqq v \leqq s-1$. Thus (C3) is satisfied and $A$ will have the required property.

$j = 1$ trivially satisfies (3.1) and gives dimensions identical to Banerji's case (ii). Table 1 lists the possible values of $j$ for $1 \leqq s \leqq 18$ satisfying (3.1). Each $j$ gives a matrix with distinct parameters.

*Example.* To illustrate Scheme 4, consider $t = 2$ and $s = 4$. One then has $z = 17$ and by choosing $j = 3$, $N = 52$ and $M = 8$. The required $52 \times 8$ binary matrix $A$ may be obtained by

$$A(u, v) = x_{u+51v}, \qquad 0 \leqq u \leqq 51, \quad 0 \leqq v \leqq 7,$$

where $\{x_i\}$ is obtained from the recurrence relation over $GF(2)$:

$$x_i = x_{i-1} + x_{i-2} + x_{i-7} + x_{i-8}$$

(For a list of primitive polynomials over $GF(2)$, refer to [4]). With the initial conditions $x_0 = x_1 = \cdots = x_6 = 0$, $x_7 = 1$, one gets the MLRS as

$$0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 0\ \cdots$$

TABLE 1
*Allowed values of j for* $1 \leqq s \leqq 18$

| $s$ | allowed $j$ |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1, 3 |
| 5 | 1 |
| 6 | 1, 3, 7 |
| 7 | 1 |
| 8 | 1, 3, 5, 15 |
| 9 | 1, 7 |
| 10 | 1, 3, 11, 31, 93 |
| 11 | 1 |
| 12 | 1, 3, 5, 7, 9, 13, 15, 21, 35, 39, 45, 63, 91, 105, 117, 315 |
| 13 | 1 |
| 14 | 1, 3, 43, 127, 381 |
| 15 | 1, 7, 31, 151, 217 |
| 16 | 1, 3, 5, 15, 17, 51, 85, 255 |
| 17 | 1 |
| 18 | 1, 3, 7, 9, 19, 21, 27, 57, 63, 73, 133, 171, 189, 219, 399, 511, 657, 1197, 1387, 1533, 1971, 4599, 9709, 13797 |

We give below the transpose of the required matrix whose rows, for convenience, have been coded in right justified octal representation.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 6 | 6 | 5 | 0 | 4 | 5 | 7 | 1 | 3 | 0 | 4 | 3 | 1 | 4 | 3 |
| 1 | 4 | 1 | 4 | 0 | 7 | 3 | 0 | 2 | 5 | 4 | 4 | 7 | 1 | 6 | 5 | 3 | 7 |
| 1 | 7 | 3 | 3 | 1 | 7 | 0 | 6 | 5 | 6 | 2 | 0 | 7 | 5 | 6 | 7 | 5 | 0 |
| 0 | 1 | 0 | 0 | 5 | 5 | 7 | 4 | 6 | 7 | 0 | 5 | 6 | 4 | 6 | 2 | 5 | 2 |
| 0 | 2 | 2 | 1 | 2 | 0 | 4 | 6 | 4 | 3 | 7 | 2 | 6 | 4 | 5 | 1 | 7 | 6 |
| 0 | 0 | 0 | 6 | 6 | 5 | 0 | 4 | 5 | 7 | 1 | 3 | 0 | 4 | 3 | 1 | 4 | 3 |
| 1 | 4 | 1 | 4 | 0 | 7 | 3 | 0 | 2 | 5 | 4 | 4 | 7 | 1 | 6 | 5 | 3 | 7 |
| 1 | 7 | 3 | 3 | 1 | 7 | 0 | 6 | 5 | 6 | 2 | 0 | 7 | 5 | 6 | 7 | 5 | 0 |

**4. Solutions for $ts \leqq 15$.** The schemes described in the last section do not provide matrix $A$ for all possible combinations of the four parameters satisfying (1.1). However in the cases not covered under the schemes, it may still be possible to obtain the required $A$ matrix by utilizing the $V$ or $H$ mappings described earlier (which already satisfy (C1) and (C2) and finding a primitive polynomial of degree $st$ such that (C3) is also satisfied. This calls for only a checking of linear independence over $GF(2)$ of $st$ different powers of $\beta$. With the tables of primitive polynomials already available [4], this task can be performed very rapidly with the help of a computer.

We have made a computer search based on this and have obtained solutions in all the cases for $ts \leqq 15$. The results given in Table 2 provide ready design data in these cases. In this table the entries in the column 'mapping' denote either $H$ mapping or $V$ mapping described in § 3. $N-t+1$ takes all values dividing $2^{st}-1$. $M$ can be computed using (1.1). The primitive polynomials used are:

$$P1 : x^6+x+1,$$
$$P2 : x^8+x^5+x^3+x+1,$$
$$P3 : x^{10}+x^3+1,$$
$$P4 : x^{10}+x^4+x^3+x+1,$$
$$P5 : x^{12}+x^6+x^4+x+1,$$
$$P6 : x^{12}+x^{11}+x^9+x^8+x^7+x^5+x^2+x+1,$$
$$P7 : x^{12}+x^{11}+x^{10}+x^8+x^6+x^4+x^3+x+1,$$
$$P8 : x^{12}+x^{11}+x^6+x^4+x^2+x+1,$$
$$P9 : x^{12}+x^{11}+x^9+x^7+x^6+x^5+1,$$
$$P10: x^{14}+x^{13}+x^{11}+x^7+x^6+x^5+x^4+x^3+x^2+x+1,$$
$$P11: x^{15}+x^{12}+x^9+x^8+x^6+x^3+1,$$
$$P12: x^{15}+x^{14}+x^{12}+x^9+x^8+x^6+x^4+x^3+x^2+x+1.$$

In the cases under Schemes 3 or 4, any primitive polynomial of degree $st$ may be used. The cases when either $N-t+1=1$ (or $M-s+1=1$) or $t=1$ (or $s=1$) are not included in the table as they can be directly obtained from Schemes 1 and 2 respectively.

TABLE 2
*Design of A matrix when ts* $\leqq 15$

| $t$ | $s$ | $N-t+1$ | Mapping | Polynomial |
|---|---|---|---|---|
| 2 | 2 | 3 | $H$ | Any (Scheme 4) |
| | | 5 | $V$ | Any (Scheme 4) |
| 2 | 3 | 3 | $H$ | Any (Scheme 4) |
| | | 7 | $H$ | P1 |
| | | 9 | $V$ | Any (Scheme 4) |
| | | 21 | $H$ | Any (Scheme 3) |
| 2 | 4 | 3 | $H$ | Any (Scheme 4) |
| | | 5 | $H$ | P2 |
| | | 15 | $H$ | P2 |
| | | 17 | $V$ | Any (Scheme 4) |
| | | 51 | $V$ | Any (Scheme 4) |
| 3 | 3 | 7 | $H$ | Any (Scheme 4) |
| | | 73 | $V$ | Any (Scheme 4) |
| 2 | 5 | 3 | $H$ | Any (Scheme 4) |
| | | 11 | $V$ | P3 |
| | | 31 | $H$ | P3 |
| | | 33 | $V$ | Any (Scheme 4) |
| | | 93 | $H$ | P4 |
| 2 | 6 | 3 | $H$ | Any (Scheme 4) |
| | | 5 | $V$ | P6 |
| | | 7 | $H$ | P8 |
| | | 9 | $H$ | P7 |
| | | 13 | $H$ | P8 |
| | | 15 | $H$ | P9 |
| | | 21 | $H$ | P7 |
| | | 35 | $H$ | P8 |
| | | 39 | $H$ | P8 |
| | | 45 | $H$ | P6 |
| | | 63 | $H$ | P9 |
| | | 85 | $V$ | Any (Scheme 4) |
| | | 91 | $H$ | P8 |
| | | 105 | $H$ | P8 |
| | | 117 | $H$ | P7 |
| | | 195 | $V$ | Any (Scheme 4) |
| | | 273 | $H$ | P8 |
| | | 315 | $H$ | P9 |
| | | 455 | $V$ | Any (Scheme 4) |
| | | 585 | $H$ | P8 |
| 3 | 4 | 3 | $V$ | Any (Scheme 3) |
| | | 5 | $V$ | P5 |
| | | 7 | $H$ | Any (Scheme 4) |
| | | 9 | $V$ | P5 |
| | | 13 | $V$ | P5 |
| | | 15 | $V$ | P6 |
| | | 21 | $V$ | P5 |
| | | 35 | $V$ | P6 |
| | | 39 | $H$ | P6 |
| | | 45 | $V$ | P7 |
| | | 63 | $H$ | P7 |

TABLE 2 (*Contd.*)

| $t$ | $s$ | $N-t+1$ | Mapping | Polynomial |
|---|---|---|---|---|
| | | 85 | $V$ | P5 |
| | | 91 | $H$ | P8 |
| | | 105 | $V$ | P5 |
| | | 117 | $H$ | P5 |
| | | 195 | $V$ | P5 |
| | | 273 | $V$ | Any (Scheme 4) |
| | | 315 | $V$ | P9 |
| | | 455 | $V$ | P5 |
| | | 585 | $H$ | P5 |
| | | 819 | $V$ | Any (Scheme 4) |
| 2 | 7 | 3 | $H$ | Any (Scheme 4) |
| | | 43 | $H$ | P10 |
| | | 127 | $H$ | P10 |
| | | 129 | $V$ | Any (Scheme 4) |
| | | 381 | $H$ | P10 |
| 3 | 5 | 7 | $H$ | Any (Scheme 4) |
| | | 31 | $H$ | P11 |
| | | 151 | $V$ | P11 |
| | | 217 | $V$ | P11 |
| | | 1057 | $V$ | Any (Scheme 4) |
| | | 4081 | $H$ | P12 |

## REFERENCES


[1] R. B. BANERJI, *The construction of binary matrices with distinct submatrices*, IEEE Trans. Computers, C-27 (1978), pp. 162–164.

[2] B. GORDON, *On the existence of perfect maps*, IEEE Trans. Information Theory, IT-12 (1966), pp. 486–487.

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

[4] W. W. PETERSON AND E. J. WELDON, JR, *Error Correcting Codes*, MIT Press, Cambridge, MA, 1972.

[5] I. S. REED AND R. M. STEWART. *Note on the existence of perfect maps*, IEEE Trans. Information Theory, IT-8 (1962), pp. 10–12.

[6] J. H. VAN LINT, *Coding Theory*, Springer-Verlag, Berlin, 1971.

[7] J. H. VAN LINT, F. J. MACWILLIAMS AND N. J. A. SLOANE, *On pseudo-random arrays*, SIAM J. Appl. Math., 36 (1979), pp. 62–72.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0015 $01.00/0

# RANDOM WALKS ON A 600-CELL*

GÉRARD LETAC† AND LAJOS TAKÁCS‡

**Abstract.** In a series of random walks (random flights) a traveler visits the vertices of a 600-cell (a four-dimensional regular polytope). The traveler starts at a given vertex and in each walk, independently of the others, chooses a vertex at random as the destination. In each walk 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 $n$th walk.

**1. Introduction.** In this paper we shall study random walks (random flights) on a four-dimensional regular polytope, the 600-cell. There are six four-dimensional regular polytopes: the regular simplex, the cross polytope, the measure polytope, the 24-cell, the 600-cell and the 120-cell. See Table 1 where $\{p, q, r\}$ is the Schläfli symbol and $N_0, N_1, N_2, N_3$ are the numbers of vertices, edges, faces and cells of the polytope. The symbol $\{p, q, r\}$ means that the faces of the polytope are $p$-gons, a vertex belongs to $q$ faces and an edge belongs to $r$ cells. We note that $N_0 - N_1 + N_2 - N_3 = 0$. For the theory of regular polytopes we refer to H. S. M. Coxeter [1], D. M. Y. Sommerville [7] and L. Schläfli [4]. The 600-cell has been studied in detail by P. H. Schoute [5], S. L. van Oss [8], [9] and D. M. Y. Sommerville [6].

TABLE 1
*Regular polytopes in four dimensions*

| Polytope | Schläfli symbol | $N_0$ | $N_1$ | $N_2$ | $N_3$ |
|---|---|---|---|---|---|
| Regular simplex | {3, 3, 3} | 5 | 10 | 10 | 5 |
| Cross polytope | {3, 3, 4} | 8 | 24 | 32 | 16 |
| Measure polytope | {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 |

In a previous paper [2] the authors have already studied random walks on the first four polytopes of Table 1. Here we are concerned with random walks on the 600-cell.

We shall use the notations $\mathbf{x} = (x_1, x_2, x_3, x_4)$, $\mathbf{y} = (y_1, y_2, y_3, y_4)$, $\cdots$ for the points of the four-dimensional Euclidean space. We define the norm of $\mathbf{x}$ by $\|\mathbf{x}\| = (x_1^2 + x_2^2 + x_3^2 + x_4^2)^{1/2}$, the distance between $\mathbf{x}$ and $\mathbf{y}$ by $\|\mathbf{x} - \mathbf{y}\|$, and the inner product of $\mathbf{x}$ and $\mathbf{y}$ by $(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4$.

A 600-cell contained in a sphere of radius 2 and center (0, 0, 0, 0) has the following 120 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

$$\tau = \frac{1+\sqrt{5}}{2} = 1.618\,033\,988\,7\cdots. \tag{1}$$

We shall denote the vertices of this 600-cell by $\mathbf{x}_r$ $(r = 0, 1, \cdots, 119)$. This polytope has 720 edges of length $2\tau^{-1} = 2(\tau - 1)$.

First, we suppose that a traveler takes a series of random walks along the edges of the 600-cell. The traveler starts at a given vertex and in each walk, independently of the

* Received by the editors, November 13, 1978.
† Départment de Mathématiques, Université Paul Sabatier, 31400, Toulouse, France.
‡ Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio 44106.

others, moves along an edge to one of the 12 adjacent vertices with probability 1/12. Denote by $p(n)$ the probability that at the end of the $n$th walk the traveler returns to the initial position. We shall prove that

$$120p(n+2) = 1 + \frac{1}{4^n} + \left(\frac{1}{2^n} + \frac{1}{3^n}\right)\left(v_{n+2} + \frac{(-1)^n}{2^n}\right) \tag{2}$$

for $n = 0, 1, 2, \cdots$ where $v_0, v_1, v_2, \cdots$ are the co-called Lucas numbers defined by

$$v_n = \frac{(1+\sqrt{5})^n + (1-\sqrt{5})^n}{2^n} \tag{3}$$

for $n = 0, 1, 2, \cdots$. Starting from $v_0 = 2$ and $v_1 = 1$, we can also determine $v_n$ by the recurrence formula $v_{n+2} = v_{n+1} + v_n$ $(n = 0, 1, 2, \cdots)$. (See E. Lucas [3].)

We shall obtain the above result as a particular case of a more general one. In the more general problem we assume that the traveler takes a series of random flights on the 600-cell. The traveler starts at a given vertex and in each flight chooses a vertex at random as the destination. The successive flights are independent and in each flight the transition probability depends only on the distance between the starting vertex and the end vertex. Define $\mathbf{v}_0$ as the initial position and $\mathbf{v}_n$ $(n = 1, 2, \cdots)$ as the position of the traveler at the end of the $n$th flight. The distance between any two vertices may take only nine possible values: $d_0, d_1, \cdots, d_8$. Let $d_0 < d_1 < \cdots < d_8$. Then $d_0 = 0$ and $d_8 = 2$. Denote by $\sigma_j$ the number of vertices whose distance is $d_j$ from a given vertex. We assume that the probability of the transition $\mathbf{v}_{n-1} \to \mathbf{v}_n$ is $p_j$ if $\|\mathbf{v}_n - \mathbf{v}_{n-1}\| = d_j$ where $p_j \geqq 0$ and

$$\sum_{j=0}^{8} \sigma_j p_j = 1. \tag{4}$$

The numbers $\sigma_j$ $(j = 0, 1, \cdots, 8)$ are given in Table 4.

We are interested in determining $p(n)$, the probability that at the end of the $n$th flight the traveler returns to the initial position. 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. Since the state space contains 120 states, it is not easy to determine the $n$th power of the transition probability matrix. Fortunately, we can solve the problem in a simpler way too. Let us choose a fixed vertex, say, $\mathbf{x}_0 = (0, 0, 0, 2)$, and define $\xi_n = j$ $(j = 0, 1, \cdots, 8)$ if $\|\mathbf{v}_n - \mathbf{x}_0\| = d_j$. It can be shown that $\{\xi_n; n = 0, 1, 2, \cdots\}$ is a homogeneous Markov chain with state space $I = \{0, 1, 2, \cdots, 8\}$ and transition probabilities

$$p_{ij} = \sum_{k=0}^{8} a_{ijk} p_k, \tag{5}$$

where $a_{ijk}$ is equal to the number of subscripts $r = 0, 1, 2, \cdots, 119$ for which $\|\mathbf{x}_r - \mathbf{x}_0\| = d_j$ and $\|\mathbf{x}_r - \mathbf{x}_s\| = d_k$, and $\mathbf{x}_s$ is any vertex for which $\|\mathbf{x}_s - \mathbf{x}_0\| = d_i$. We shall determine the transition probabilities

$$\mathbf{P}\{\xi_n = k | \xi_0 = i\} = p_{ik}^{(n)} \tag{6}$$

for $i \in I$, $j \in I$ and $n = 0, 1, 2, \cdots$. Then $p(n) = p_{00}^{(n)}$ is the probability that the traveler returns to the initial position in $n$ flights.

We shall prove that

$$120p_{ik}^{(n)} = \sigma_k \sum_{j=0}^{8} h_{ij} h_{kj} \lambda_j^n \tag{7}$$

where

$$\lambda_j = \sum_{\nu=0}^{8} p_\nu \lambda_{j\nu} \tag{8}$$

and the matrices $[h_{ij}]$ and $[\lambda_{j\nu}]$ are given in Tables 2 and 3. In Tables 2 and 3, $\tau$ is defined by (1) and $\tau^{-1} = \tau - 1$. In the above formulas $p_0, p_1, \cdots, p_8$ are nonnegative numbers satisfying (4).

TABLE 2
$h_{ij}$

| $i$ \ $j$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 4 | 3 | 3 | 5 | 4 | 6 | 2 | 2 |
| 1 | 1 | −1 | $\tau$ | $-\tau^{-1}$ | 0 | 1 | −1 | $\tau$ | $-\tau^{-1}$ |
| 2 | 1 | 1 | 0 | 0 | −1 | −1 | 0 | 1 | 1 |
| 3 | 1 | −1 | $-\tau^{-1}$ | $\tau$ | 0 | −1 | 1 | $\tau^{-1}$ | $-\tau$ |
| 4 | 1 | 0 | −1 | −1 | 1 | 0 | 0 | 0 | 0 |
| 5 | 1 | −1 | $-\tau^{-1}$ | $\tau$ | 0 | 1 | −1 | $-\tau^{-1}$ | $\tau$ |
| 6 | 1 | 1 | 0 | 0 | −1 | 1 | 0 | −1 | −1 |
| 7 | 1 | −1 | $\tau$ | $-\tau^{-1}$ | 0 | −1 | 1 | $-\tau$ | $\tau^{-1}$ |
| 8 | 1 | 4 | 3 | 3 | 5 | −4 | −6 | −2 | −2 |

TABLE 3
$\lambda_{j\nu}$

| $j$ \ $\nu$ | 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 | $4\tau$ | 0 | $-4\tau^{-1}$ | −10 | $-4\tau^{-1}$ | 0 | $4\tau$ | 1 |
| 3 | 1 | $-4\tau^{-1}$ | 0 | $4\tau$ | −10 | $4\tau$ | 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 | $6\tau$ | 10 | $6\tau^{-1}$ | 0 | $-6\tau^{-1}$ | −10 | $-6\tau$ | −1 |
| 8 | 1 | $-6\tau^{-1}$ | 10 | $-6\tau$ | 0 | $6\tau$ | −10 | $6\tau^{-1}$ | −1 |

In particular, it follows from (7) that

$$120p(n) = \lambda_0^n + 16\lambda_1^n + 9\lambda_2^n + 9\lambda_3^n + 25\lambda_4^n + 16\lambda_5^n + 36\lambda_6^n + 4\lambda_7^n + 4\lambda_8^n \tag{9}$$

for $n = 0, 1, 2, \cdots$. If $p_1 = 1/12$ and $p_j = 0$ for $j \neq 1$, then (9) reduces to (2).

Before proving (7) we would like to mention another generalization of (2). Denote by $D(\mathbf{x}_r, \mathbf{x}_s)$ the smallest number of edges in the paths connecting the vertices $\mathbf{x}_r$ and $\mathbf{x}_s$ of the 600-cell. The possible values of $D(\mathbf{x}_r, \mathbf{x}_s)$ are 0, 1, 2, 3, 4, 5. Let us assume, as an alternative, that in the random flights, the transition $\mathbf{v}_{n-1} \to \mathbf{v}_n$ has probability $q_j$ if $D(\mathbf{v}_{n-1}, \mathbf{v}_n) = j$ where $q_j \geqq 0$ and

$$q_0 + 12q_1 + 32q_2 + 42q_3 + 32q_4 + q_5 = 1. \tag{10}$$

Then, $p(n)$, the probability that the traveler returns to the initial position at the end of the $n$th flight, is given again by (9) where now $p_0 = q_0$, $p_1 = q_1$, $p_2 = p_3 = q_2$, $p_4 = p_5 = q_3$, $p_6 = p_7 = q_4$ and $p_8 = q_5$.

**2. The numbers $a_{ijk}$.** We denote by $\mathbf{x}_r (r = 0, 1, 2, \cdots, 119)$ the 120 vertices of the 600-cell and fix $\mathbf{x}_0 = (0, 0, 0, 2)$. We can divide the 120 vertices into 9 sets $S_j$ $(j = 0, 1, \cdots, 8)$ so that $S_j$ contains the vertices $\mathbf{x}_r$ $(r = 0, 1, 2, \cdots, 119)$ for which $\|\mathbf{x}_r - \mathbf{x}_0\| = d_j$. If $\|\mathbf{x}_r - \mathbf{x}_0\| = d_j$, then $D(\mathbf{x}_r, \mathbf{x}_0) = D_j$ is uniquely determined. If $\|\mathbf{x}_r - \mathbf{x}_0\| = d_j$, then $(\mathbf{x}_r, \mathbf{x}_0) = c_j = (8 - d_j^2)/2$. Table 4 contains $S_j$, $\sigma_j$, $d_j^2$, $D_j$ and $c_j$ for $j = 0, 1, \cdots, 8$. The number of vertices in the set $S_j$ is $\sigma_j$. In Table 4 only some representative vertices are displayed. To obtain all the vertices in the set $S_j$ we need to equip the first three coordinates in each vertex by the signs $\pm$ and form the cyclic permutations of the first three coordinates. We note that if we use a computer to calculate the distances $\|\mathbf{x}_r - \mathbf{x}_0\| = d_j$ or the inner products $(\mathbf{x}_r, \mathbf{x}_0) = c_j$, then we can use the equations $[(d_j^2 + 1)/2] = [4.5 - c_j] = j$ for the determination of the appropriate $S_j$.

TABLE 4

| $j$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $S_j$ | $(0, 0, 0, 2)$ | $(1, 0, \tau^{-1}, \tau)$ | $(1, 1, 1, 1)$<br>$(\tau, \tau^{-1}, 0, 1)$ | $(\tau, 0, 1, \tau^{-1})$ | $(2, 0, 0, 0)$<br>$(\tau, 1, \tau^{-1}, 0)$ |
| $\sigma_j$ | 1 | 12 | 20 | 12 | 30 |
| $d_j^2$ | 0 | $6-2\sqrt{5}$ | 4 | $10-2\sqrt{5}$ | 8 |
| $D_j$ | 0 | 1 | 2 | 2 | 3 |
| $c_j$ | 4 | $2\tau$ | 2 | $2\tau^{-1}$ | 0 |

| $j$ | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| $S_j$ | $(\tau, 0, 1, -\tau^{-1})$ | $(1, 1, 1, -1)$<br>$(\tau, \tau^{-1}, 0, -1)$ | $(1, 0, \tau^{-1}, -\tau)$ | $(0, 0, 0, -2)$ |
| $\sigma_j$ | 12 | 20 | 12 | 1 |
| $d_j^2$ | $6+2\sqrt{5}$ | 12 | $10+2\sqrt{5}$ | 16 |
| $D_j$ | 3 | 4 | 4 | 5 |
| $c_j$ | $-2\tau^{-1}$ | $-2$ | $-2\tau$ | $-4$ |

Since $\|\mathbf{x}_r - \mathbf{x}_s\|^2 = \|\mathbf{x}_r\|^2 + \|\mathbf{x}_s\|^2 - 2(\mathbf{x}_r, \mathbf{x}_s)$ and $\|\mathbf{x}_r\| = \|\mathbf{x}_s\| = 2$, it follows that $\|\mathbf{x}_r - \mathbf{x}_s\| = d_j$ if and only if $(\mathbf{x}_r, \mathbf{x}_s) = c_j$. Thus we can characterize $S_j$ as the set of vertices $\mathbf{x}_r$ $(r = 0, 1, \cdots, 119)$ for which $(\mathbf{x}_r, \mathbf{x}_0) = c_j$. Since $\mathbf{x}_0 = (0, 0, 0, 2)$, therefore it is indeed very easy to sort the vertices $\mathbf{x}_r$ $(r = 0, 1, \cdots, 119)$ into the sets $S_0, S_1, \cdots, S_8$.

We can easily enumerate $a_{ijk}$ if we use the following equivalent definition: $a_{ijk}$ is equal to the number of subscripts $r = 0, 1, 2, \cdots, 119$ for which $(\mathbf{x}_r, \mathbf{x}_0) = c_j$ and $(\mathbf{x}_r, \mathbf{x}_s) = c_k$, and $\mathbf{x}_s$ is any vertex for which $(\mathbf{x}_s, \mathbf{x}_0) = c_i$. Tables 5, 6, 7, 8 contain $a_{ijk}$ for $i, j = 0, 1, \cdots, 8$ and $k = 1, 2, \cdots, 7$. Obviously $a_{ij0} = a_{i,8-j,8} = 1$ if $i = j$ and $a_{ij0} = a_{i,8-j,8} = 0$ if $i \neq j$.

If we choose $\mathbf{x}_r$ such that $\mathbf{x}_{119-r} = -\mathbf{x}_r$ for $r = 0, 1, \cdots, 119$, then it follows immediately from the definition of $a_{ijk}$ that

$$a_{ijk} = a_{8-i,8-j,k} \tag{11}$$

and

$$a_{ijk} = a_{i,8-j,8-k}. \tag{12}$$

Furthermore, we have

$$a_{ijk} = a_{ikj} \tag{13}$$

and

$$\sum_{j=0}^{8} a_{ijk} = \sigma_k \tag{14}$$

for $i = 0, 1, \cdots, 8$.

TABLE 5

$a_{ij1} = a_{i,8-j,7}$

| $i$ \ $j$ | 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 6

$a_{ij2} = a_{i,8-j,6}$

| $i$ \ $j$ | 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 7

$a_{ij3} = a_{i,8-j,5}$

| $i$ \ $j$ | 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 8

$a_{ij4}$

| $i$ \ $j$ | 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 | | | | |

**3. The probabilities $p_{ik}^{(n)}$.** We consider the Markov chain $\{\xi_n;\ n = 0, 1, 2, \cdots\}$ with state space $I = \{0, 1, 2, \cdots, 8\}$ and transition probability matrix

$$\boldsymbol{\pi} = [p_{ij}]_{i,j \in I}, \tag{15}$$

where $p_{ij}$ is defined by (5). If we arrange the $n$step transition probabilities in the form of a matrix, then we get

$$[p_{ik}^{(n)}]_{i,k \in I} = \boldsymbol{\pi}^n \tag{16}$$

for $n = 0, 1, 2, \cdots$. Thus in order to find $p_{ik}^{(n)}$ we need to determine $\boldsymbol{\pi}^n$. If we can form the Jordan decomposition of $\boldsymbol{\pi}$, then $\boldsymbol{\pi}^n$ can easily be obtained. However, for the first sight, it seems hopeless to solve the characteristic equation of a $9 \times 9$ matrix whose elements depend on 8 parameters. Fortunately, several favorable circumstances make it possible to determine the Jordan decomposition of $\boldsymbol{\pi}$.

By (5) we can write that

$$\boldsymbol{\pi} = \sum_{k=0}^{8} p_k \mathbf{A}_k \tag{17}$$

where

$$\mathbf{A}_k = [a_{ijk}]_{i,j\in I} \tag{18}$$

for $k = 0, 1, \cdots, 8$.

First, let us consider the matrix $\mathbf{A}_1$. The elements of $\mathbf{A}_1$ satisfy the symmetry relation $a_{ij1} = a_{8-i,8-j,1}$ and thus by using the method described in the Appendix we can reduce the problem of finding the Jordan decomposition of $\mathbf{A}_1$ to the problem of finding the Jordan decompositions of a $4\times 4$ matrix and a $5\times 5$ matrix. Thus we obtain that

$$\mathbf{A}_1\mathbf{H} = \mathbf{H}\mathbf{\Lambda}_1, \tag{19}$$

where the elements of the matrix $\mathbf{H} = [h_{ij}]_{i,j\in I}$ are given in Table 2 and $\mathbf{\Lambda}_1$ is one of the matrices

$$\mathbf{\Lambda}_\nu = [\delta_{ij}\lambda_{j\nu}]_{i,j\in I}. \tag{20}$$

In (20) $\delta_{ij} = 1$ for $i = j$, $\delta_{ij} = 0$ for $i \neq j$ and $\lambda_{j\nu}$ $(j \in I, \nu \in I)$ are given in Table 3. In (19) the matrix $\mathbf{H}$ is nonsingular and thus $\lambda_{j1}$ $(j = 0, 1, \cdots, 9)$ are the eigenvalues of $\mathbf{A}_1$. The nine eigenvalues of $\mathbf{A}_1$ are distinct.

Luckily $\mathbf{H}$ has a simple inverse. We observe that if $\mathbf{H}'$ is the transpose of $\mathbf{H}$, then

$$\mathbf{H}\mathbf{H}' = 120[\delta_{ij}\sigma_j^{-1}]_{i,j\in I}, \tag{21}$$

that is, $\mathbf{H}\mathbf{H}'$ is a diagonal matrix and the diagonal elements are $120\sigma_k^{-1}$ $(k \in I)$ where $\sigma_k$ $(k \in I)$ are given in Table 4.

Quite surprisingly, it turns out that

$$\mathbf{A}_k\mathbf{H} = \mathbf{H}\mathbf{\Lambda}_k \tag{22}$$

holds not only for $k = 1$ but for every $k = 0, 1, 2, \cdots, 8$. Since $\mathbf{A}_0$ is the identity matrix, (22) obviously holds for $k = 0$. If (22) is true for $k = 0, 1, 2, 3, 4$, then by symmetry it is true for $k = 5, 6, 7, 8$ too. Thus it remains to check (22) for $k = 2, 3, 4$. In each case (22) is indeed correct. It would be interesting to infer (22) from the structure of the symmetry group of the 600-cell. This would save some calculations. By (22) it follows that the eigenvalues of $\mathbf{A}_k$ are $\lambda_{jk}$ $(j = 0, 1, \cdots, 8)$.

Finally, by (17) and (22) it follows that

$$\boldsymbol{\pi}\mathbf{H} = \mathbf{H}\mathbf{\Lambda}, \tag{23}$$

where

$$\mathbf{\Lambda} = [\delta_{ij}\lambda_j]_{i,j\in I} \tag{24}$$

is a diagonal matrix with diagonal elements $\lambda_j$ $(j = 0, 1, \cdots, 8)$ given by (8). Accordingly, the eigenvalues of $\boldsymbol{\pi}$ are $\lambda_j$ $(j = 0, 1, \cdots, 8)$ defined by (8). By (21) and (23) we get

$$120\boldsymbol{\pi} = \mathbf{H}[\delta_{ij}\lambda_j]\mathbf{H}'[\delta_{jk}\sigma_k]. \tag{25}$$

Thus we arrived at the Jordan decomposition of $\boldsymbol{\pi}$ and by (25) we have

$$120\boldsymbol{\pi}^n = \mathbf{H}[\delta_{ij}\lambda_j^n]\mathbf{H}'[\delta_{ij}\sigma_j] \tag{26}$$

for all $n = 0, 1, 2, \cdots$. Hence (7) follows.

**Appendix.** Let us suppose that the elements of the matrix

$$\mathbf{A} = [a_{ij}]_{i,j\in I}, \tag{A.1}$$

where $I = \{0, 1, \cdots, m\}$, satisfy the symmetry relation

$$a_{ij} = a_{m-i,m-j} \tag{A.2}$$

for $i, j \in I$. Define

$$b_{ij} = \begin{cases} a_{ij} + a_{i,m-j} & \text{for } 0 \leqq i \leqq \frac{m}{2}, \quad 0 \leqq j < \frac{m}{2}, \\ a_{ij} & \text{for } 0 \leqq i \leqq \frac{m}{2}, \quad j = \frac{m}{2} \ (m = \text{even}) \end{cases} \tag{A.3}$$

and

$$c_{ij} = a_{ij} - a_{i,m-j} \quad \text{for } \frac{m}{2} < i \leqq m, \qquad \frac{m}{2} < j \leqq m \tag{A.4}$$

and write $I_1 = \{i : 0 \leqq i \leqq m/2\}$ and $I_2 = \{i : m/2 < i \leqq m\}$.

If there exist two nonsingular matrices $[\beta_{ij}]_{i,j \in I_1}$, and $[\gamma_{ij}]_{i,j \in I_2}$ such that

$$[b_{ij}][\beta_{jk}] = [\beta_{ij}][\delta_{jk}\lambda_k] \qquad (i, j, k \in I_1) \tag{A.5}$$

and

$$[c_{ij}][\gamma_{jk}] = [\gamma_{ij}][\delta_{jk}\lambda_k] \qquad (i, j, k \in I_2), \tag{A.6}$$

then there exists a nonsingular matrix $[\alpha_{ij}]_{i,j \in I}$ such that

$$[a_{ij}][\alpha_{jk}] = [\alpha_{ij}][\delta_{jk}\lambda_k] \qquad (i, j, k \in I) \tag{A.7}$$

and we have

$$\alpha_{ij} = \begin{cases} \beta_{ij} & \text{if } i \in I_1, \quad j \in I_1, \\ \gamma_{ij} & \text{if } i \in I_2, \quad j \in I_2, \\ \beta_{m-i,j} & \text{if } i \in I_2, \quad j \in I_1, \\ -\gamma_{m-i,j} & \text{if } i \in I_1, \quad j \in I_2, \end{cases} \tag{A.8}$$

where $\gamma_{ij} = 0$ if $i = m/2$ and $m$ is an even integer.

## REFERENCES


[1] H. S. M. COXETER, *Regular Polytopes*, second edition, Macmillan, New York, 1963.

[2] G. LETAC AND L. TAKÁCS, *Random walks on an m-dimensional cube*, J. Reine Angew. Math., to appear.

[3] E. LUCAS, *Sur l'emploi du calcul symbolique, dans la théorie des séries récurrentes*, Nouvelle Correspondance Mathématique, 2 (1876), pp. 201–206, 214.

[4] L. SCHLÄFLI, *Theorie der vielfachen Kontinuität*, Neue Denkschriften der allgemeinen schweizerischen Gesellschaft für die gesamten Naturwissenschaften, Band 38, Zürich, 1901; *Gesammelte Mathematische Abhandlungen*, Band I, Verlag Birkhäuser, Basel, 1950, pp. 167–387.

[5] P. H. SCHOUTE, *Regelmässige Schnitte und Projektionen des Hundertzwanzigzelles und Sechshundertzelles im vierdimensionalen Raume*, I–II, Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam, Eerste Sectie 2, (7) (1894), 20 pp., and Ibid. 9, (4) (1907), 32 pp.

[6] D. M. Y. SOMMERVILLE, *Description of a projection model of the 600-cell in space of four dimensions*, Proc. Roy. Soc. Edinburgh, 34 (1913–1914), pp. 253–258.

[7] ———, *An Introduction to the Geometry of N Dimensions*, Dover, New York, 1958.

[8] S. L. VAN OSS, *Das regelmässige Sechshundertzell und seine selbsdeckenden Bewegungen*, Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam, Eerste Sectie 7, (1) (1900), 18 pp.

[9] ———, *Die regelmässigen vierdimensionalen Polytope höherer Art*, Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam, Eerste Sectie 12, (1) (1915), 13 pp.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 1, March 1980


0196-5212/80/0101-0016 $01.00/0

# ASYMPTOTIC EQUILIBRIA IN A CLASS OF $N$-PERSON SYMMETRIC GAMES*

HARVEY DIAMOND†

**Abstract.** In a class of symmetric $N$-person games we consider the behavior, for large $N$, of the symmetric equilibrium. In the game each player independently chooses one of $M$ alternatives; the payoff is a decreasing function of the number of players choosing the same alternative. Asymptotic approximations for the equilibrium strategy and payoff are obtained. A number of examples are treated in detail. We also consider the maximin strategy of a single player against $N-1$ minimizing opponents. It is shown, for a certain subclass of payoff functions, that this maximin payoff is asymptotic to the symmetric equilibrium payoff and the maximin strategy is related asymptotically to the symmetric equilibrium strategy.

**1. Introduction.** In this paper we investigate the symmetric equilibrium of the following noncooperative symmetric $N$-person game: Let $N$ players independently choose one of $M$ alternatives. The payoff to each of $n$ players who choose alternative $i$ is denoted by $S_i(n)$. Initially, we require only that the functions $S_i(n)$ be nonincreasing but not constant. Certain asymptotic requirements on the $S_i(n)$ will be added later.

As suggested by the title, our primary concern is describing the behavior and properties of the symmetric equilibrium for large $N$. The asymptotic procedures and results developed in § 3–5 apply to a fairly large class of payoff functions whose members are characterized by an asymptotic condition and, roughly speaking, have the following property: If the function $f(x)$, $x \geqq 0$, is in the class then $E[f(X)] \sim f(Np)$ as $N \to \infty$ where $X$ is a binomial random variable with distribution $B(N, p)$ and $E[\cdot]$ denotes expectation. This property allows the asymptotic expansion of the expected payoffs which appear in the equations for the equilibrium. Formal definitions and theorems appear in § 3.

After obtaining some general first order results in § 3, we develop in §§ 4 and 5 asymptotic expansions for the symmetric equilibrium probability distribution and payoff in a number of special cases, chosen both for their individual interest and to illustrate collectively the wide range of asymptotic behavior possible within the class of treatable payoff functions. These cases include $S_i(n) \sim a_i/n^{\alpha}$ and $S_i(n) \sim a_i/n^{\alpha_i}$.

We will also consider, for large $N$, the behavior of the optimal strategies and payoff when our game is played by a single player versus an $N-1$ player coalition seeking to minimize his payoff. The optimal payoff in this game gives the single player a lower bound on what he can obtain in the non-cooperative game; his corresponding optimal strategy guarantees at least this lower bound. In § 3 we obtain results which relate, for large $N$, the optimal strategies and payoff in the coalitional game with those of the symmetric equilibrium in the noncooperative game. In particular, we show that the payoffs are asymptotically equal and, in the examples treated later, show that the symmetric equilibrium strategy is asymptotically optimal for the single player in the coalitional game.

In § 4 we consider a dynamic version of our game in which the alternatives are presented and chosen sequentially. For large $N$ we show that the sequential nature of the game has no effect to first order on the equilibria. Finally, § 7 contains some general remarks about the results obtained and some further questions concerning extensions of the asymptotic analysis.

* Received by the editors February 26, 1979.
† Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506.

**2. Equations and structure of the equilibrium.** For the purposes of this paper, the term symmetric equilibrium (often simply equilibrium) refers to a common randomized strategy for the players having the property that no player can improve his expected payoff by unilaterally changing his strategy.

The existence of a symmetric equilibrium in symmetric games was proved by Nash [2] in an early paper on equilibria in $N$-person games. We will show uniqueness for our particular game later in this section.

We introduce some notation first. Let

$p_i$ = probability of choosing alternative $i$ under the equilibrium strategy.

$A$ = the set of active alternatives, i.e. those $i$ for which $p_i > 0$.

$I$ = the set of inactive alternatives having $p_i = 0$.

$C$ = the expected payoff to each player under the equilibrium strategy.

For a particular player, let

$n_i$ = the random variable giving the number of his opponents who choose alternative $i$ under the equilibrium strategy. Observe that $n_i$ has a binomial distribution with parameters $(N-1, p_i)$ or notationally, $n_i \sim B(N-1, p_i)$.

If the random variable $X \sim B(N-1, p)$, define the functions

$$L_i(p) = E[S_i(1+X)] = \sum_{k=0}^{N-1} \binom{N-1}{k} S_i(1+k) p^k (1-p)^{N-1-k}, \qquad i = 1, \cdots, M.$$

The equations for the equilibrium strategy $\{p_i\}_1^M$ and payoff $C$ are then

$$C \geqq L_i(p_i); \qquad C > L_j(p_j) \Rightarrow p_j = 0; \qquad \sum_{i=1}^{M} p_i = 1.$$

We note that if $i \in A$ then $C = L_i(p_i)$ and if $j \in I$ then $L_j(p_j) = S_j(1)$. The equations for the equilibrium may then be rewritten as

$$C = L_i(p_i), \quad i \in A; \qquad C \geqq S_j(1), \quad j \in I; \qquad \sum_{i \in A} p_i = 1. \tag{2.1}$$

To prove the uniqueness of the equilibrium we need the following:

PROPOSITION 1. *For $p \in (0, 1)$, $L_i(p)$ is monotone decreasing.*

*Proof.* We easily calculate

$$dL_i/dp = (N-1) \sum_{k=0}^{N-2} \binom{N-2}{k} [S_i(k+2) - S_i(k+1)] p^k (1-p)^{N-2-k}.$$

The functions $S_i(k)$ were assumed nonincreasing but not constant so each term in the sum is nonpositive and at least one is negative.

THEOREM 1. *There is a unique symmetric equilibrium.*

*Proof.* If $\{p_i\}$ and $\{p_i'\}$ are two distinct symmetric equilibria there exist indices $j$ and $k$ such that $p_j > p_j'$ and $p_k < p_k'$. Evidently $p_j > 0$ and $p_k' > 0$ so that $L_j(p_j) \geqq L_k(p_k)$ and $L_k(p_k') \geqq L_j(p_j')$. Applying Proposition 1 however, $L_j(p_j) < L_j(p_j') \leqq L_k(p_k') < L_k(p_k)$ and a contradiction is obtained.

Using Proposition 1 and Theorem 1 we can prove the following useful result:

PROPOSITION 2. *For any probability distribution $\{p_i'\}$ we have*

$$\min_{i \in A'} L_i(p_i') \leqq C \leqq \max_i L_i(p_i'), \quad \text{where } A' = (i: p_i' > 0)$$

*Proof.* If $\{p_i'\}$ is the equilibrium strategy then the conclusion obviously holds. Otherwise there exist indices $j$ and $k$ such that $p_j > p_j'$ and $p_k < p_k'$. Then

$\min_{i\in A'} L_i(p_i') \leqq L_k(p_k') < L_k(p_k) \leqq C = L_j(p_j) < L_j(p_j') \leqq \max_i L_i(p_i')$. This proves the proposition.

Proposition 2 shows that, roughly speaking, any strategy and payoff which produces a good approximation to (2.1) necessarily provides a good approximation to the equilibrium payoff.

Propositions 1 and 2 render the structure of the equilibrium relatively transparent. By "structure" is meant the classification according to player of those pure strategies which have positive probability and those which have zero probability in the equilibrium being considered. In the computation of equilibria there are many possible structures to search through in general. For the symmetric equilibrium of our game however, it is easy to see at most $M$ different possibilities for the set $A$ exist: Renumber the alternatives so that $S_i(1)$ is a nonincreasing sequence. Then (2.1) and Proposition 1 together imply that the sets $A$ and $I$ are of the forms $A = \{1, 2, \cdots, k\}$ and $I = \{k+1, k+2, \cdots, M\}$ (of course $I$ is empty if $k = M$). A simple algorithm for computing the solution would be to successively try $k = 1, 2, \cdots$ at each stage solving the equations $L_i(p_i') = C', \sum_{i=1}^{k} p_i' = 1, i = 1, \cdots, k$.

In producing successive approximations while solving these equations we use Proposition 2 to discard alternatives which cannot be active by virtue of their value of $S_i(1)$ falling below the lower bound for $C$; and cease computations at the present value of $k$ and increment by 1 if $S_{k+1}(1) > \max_{i\in A'} L_i(p_i') \geqq C'$ occurs. The structure of the equilibrium is determined when either $k = M$ or $S_{k+1}(1) \leqq \min_{i\in A'} L_i(p_i')$ occurs.

While it is certainly possible to refine the preceding algorithm to a certain extent we will not delve into the details here as the computational aspects of the solution are not our primary interest. Similarly we will not consider the numerical solution of the nonlinear equations in (2.1). The following sections deal with the asymptotic behavior of the $p_i$ and $C$ under various additional assumptions on the $S_i(n)$.

**3. Some general asymptotic results.** Our motivations for investigating asymptotic behavior and developing asymptotic approximations for the equilibrium are several:

(a) The numerical solution becomes more difficult for large $N$;

(b) The asymptotic solution becomes more accurate for large $N$;

(c) The asymptotic solution exhibits analytically the dependence of the solution on $N$ and the parameters of the problem. This dependence is often of a simple character capable of explicit description via elementary functions.

(d) The asymptotic solution has a number of interesting and intuitively appealing properties.

Notationally, it will often be necessary to explicitly point out the dependence of the solution on $N$ by writing $p_i(N)$ and $C(N)$. Asymptotic approximations will generally be denoted by an asterisk viz. $p_i^*(N)$, $C^*(N)$. For the remainder of the paper we will assume that $S_i(n) \to 0$ as $n \to \infty$ for each $i$.

A simple asymptotic result is the following:

THEOREM 2. *If for each $i$, $S_i(n) \to 0$ as $n \to \infty$ then $C(N) \to 0$, $Np_i(N) \to \infty$ and $Np_i(N)q_i(N) \to \infty$ as $N \to \infty$ for each $i$, where $q_i = 1 - p_i$.*

*Proof.* We first show $C(N) \to 0$. For each $m \geqq 1$, $n \geqq 1$ we have

$$S_i(n) \leqq S_i(1)2^{m-n} + S_i(m). \tag{3.1}$$

By definition, $L_i(p) = E[S_i(1+X)]$ where $X \sim B(N-1, p)$. Using (3.1) gives

$$L_i(p) \leqq S_i(1)2^{m-1}E[2^{-X}] + S_i(m) = S_i(1)2^{m-1}(1-p/2)^{N-1} + S_i(m).$$

Now for each $N$ there is some index $j$ for which $p_j(N) \geqq 1/M$. Then

$$C(N) = L_j(p_j) \leqq S_j(1)2^{m-1}[1-1/(2M)]^{N-1} + S_j(m).$$

Letting $N \to \infty$, we have

$$\limsup_N C(N) \leqq \max_i S_i(m)$$

and by choosing $m$ large enough we can make the right hand side as small as we like. Thus $C(N) \to 0$.

For the second part of the proof, suppose that for some $i$, $Np_i(N)$ doesn't go to infinity. Then there exists a constant $K$ such that $Np_i(N) < K$ for infinitely many $N$. For such an $N$, sufficiently large

$$C(N) \geqq L_i(p_i) \geqq S_i(1)[1-p_i(N)]^{N-1} > S_i(1) \exp(-2K)$$

which contradicts $C(N) \to 0$. Thus $Np_i(N) \to \infty$. It then follows that $Nq_i(N) \to \infty$ and finally $Np_i(N)q_i(N) \to \infty$. This completes the proof of Theorem 2.

It follows from Theorem 2, in particular the fact that $C(N) \to 0$, that for $N$ sufficiently large all alternatives are active. The equations for the equilibrium are then

$$C = E[S_i(1+n_i)], \quad \text{where } n_i \sim B(N-1, p_i); \qquad \sum_1^M p_i = 1. \tag{3.2}$$

The asymptotic evaluation of $E[S_i(1+n_i)]$, on which the bulk of our results depend, will be accomplished using [1]. The relevant results are presented here for reference:

DEFINITION. The locally bounded function $f(x)$, $x \in [0, \infty)$, is said to have *essentially zero asymptotic relative variation* (EZARV) if the function $g(x, a)$ defined by

$$g(x, a) \triangleq \sup_{|y| \leqq a} \left| \frac{f(x+y\sqrt{x})}{f(x)} - 1 \right|$$

satisfies $g(x, a) \to 0$ as $x \to \infty$ for any fixed $a > 0$.

PROPOSITION 3. *If $f(x)$ is differentiable and $f'(x)/f(x) = o(1/\sqrt{x})$ then $f(x)$ has EZARV.*

THEOREM 3. *Let $k \geqq 0$ be even and suppose $f^{(k)}(x)$ exists, is locally bounded and has EZARV. Let $X \sim B(N, p(N))$ where $Np(1-p) \to \infty$. Then*

$$E[f(X)] = \sum_{i=0}^{k} \frac{f^{(i)}(\mu)}{i!} s_i + \frac{f^{(k)}(\mu)}{k!} s_k o(1), \quad \text{where } s_k = E[(X-\mu)^k], \qquad \mu = Np. \tag{3.3}$$

For the remainder of the paper, all functions $S_i(n)$ we consider will be assumed to have EZARV unless otherwise noted.

It is our intention to replace the exact equations for the equilibrium (3.2) with the *asymptotic equations* obtained by replacing $E[S_i(1+n_i)]$ with the appropriate asymptotic expansion from (3.3). Hopefully these asymptotic equations will have a solution which is at the same time easily calculated and asymptotic to the exact equilibrium solution. Notationally, we will use $L_i^*(p)$ to denote an asymptotic expansion of $L_i(p)$ valid for $p = p(N)$ satisfying the hypothesis of Theorem 3. In particular of course, we can write $L_i(p_i) \sim L_i^*(p_i)$. We then seek solutions of the *asymptotic equations*

$$C^*(N) \sim L_i^*(p_i^*), \quad i = 1, \cdots, N; \qquad \sum_1^M p_i^* = 1; \qquad Np_i^*(1-p_i^*) \to \infty. \tag{3.4a}$$

We will use Proposition 2 in our examples to show that $C(N) \sim C^*(N)$ holds. Admittedly, it is slightly unclear exactly what $L_i(p) \sim L_i^*(p)$ and (3.4a) mean. Our examples will clarify this point better than a clumsy formal definition.

Of special interest to us will be the *first order asymptotic equations* obtained by using for $L_i^*(p)$ a one term expansion of $L_i(p)$ from (3.3). As defined in § 2, if $X \sim B(N-1, p)$ then $L_i(p) = E[S_i(1+X)]$. By hypothesis $S_i(1+x)$ has EZARV and using (3.3) with $\mu = (N-1)p$ gives $L_i(p) = E[S_i(1+X)] = S_i(q+Np)[1+o(1)]$. Because of the EZARV condition, dropping the $q$ produces a relative change of $o(1)$ so that $L_i(p) = S_i(Np)[1+o(1)]$. We then take $L_i^*(p) = S_i(Np)$. The first order asymptotic equations are then

$$C^* = S_i(Np_i^*)[1+o(1)]; \qquad \sum_1^M p_i^* = 1. \tag{3.4}$$

Certainly, $C^* = C$, $p_i^* = p_i$ is one solution to (3.4). We would hope to be able to guarantee that any solution of (3.4) satisfies $C^* = C[1+o(1)]$, $p_i^* = p_i[1+o(1)]$. An example presented later shows that the latter conjecture does not necessarily hold. As for the former, suppose $p_i^*$ is a solution to (3.4). Then

$$L_i(p_i^*) = S_i(Np_i^*)[1+o(1)] = C^*[1+o(1)].$$

By Proposition 2, $C$ must fall within the range of the $L_i(p_i^*)$; we must have $C = C^*[1+o(1)]$ or $C \sim C^*$.

In addition to the symmetric equilibrium, it is also of interest to consider, for an individual player, the strategy which maximizes his minimum expected payoff under the least favorable actions by his $N-1$ opponents. This problem may be cast as a zero-sum two-person game in which the single player (which we call the *individual*) attempts to maximize his payoff against the efforts of the coalition formed by his $N-1$ opponents (called the *director*). We denote this game by $G$. The next proposition shows that under certain additional restrictions on the $S_i$, the individual can guarantee himself a payoff which is asymptotic to $C(N)$; further, the payoff $C(N)$ can be asymptotically guaranteed by using a strategy expressed in terms of the asymptotic equilibrium strategy $p_i^*$ obtained from (3.4).

PROPOSITION 4. *Suppose $S_i(x)$ is continuous, piecewise differentiable and for $x$ sufficiently large, $S_i'(x)$ is monotone decreasing in magnitude. Let $V(N)$ denote the value of the game $G$ and let $\{p_i^*\}$ be any solution of* (3.4). *Then*

$$V(N) = C(N)[1+o(1)] \tag{3.5}$$

*and any strategy $\{\bar{p}_i\}$ for the individual which satisfies*

$$\bar{p}_i \sim \frac{[S_i'(Np_i^*)]^{-1}}{\sum_{j=1}^M [S_j'(Np_j^*)]^{-1}}; \qquad \sum_1^M \bar{p}_i = 1 \tag{3.6}$$

*guarantees a payoff $V(N)[1+o(1)]$.*

*Proof.* It is clear that the director can hold the individual to a payoff $C(N)$ by having his $N-1$ agents independently play the equilibrium strategy $\{p_i\}$. Thus $V(N) \leqq C(N)$. On the other hand, if the individual plays some strategy $\{\bar{p}_i\}$ he is guaranteed a payoff of at least

$$Z \triangleq \min_{n_i} \sum_1^M \bar{p}_i S_i(n_i+1) \quad \text{subject to} \quad \sum_1^M n_i = N-1. \tag{3.7}$$

If we further restrict $n_i \geqq \sqrt{N}$, the individual's minimum payoff will rise to at most $Z[1+o(1)]$. This is because, in any optimal solution of (3.7), $(M-1)\sqrt{N}$ players can be reallocated from the maximum of the $n_i$ to the other $M-1$ alternatives with the

EZARV condition guranteeing a payoff increase of relative size at most $[1+o(1)]$. Replacing $N-1$ by $N$ in (3.7) and dropping the integrality constraints on the $n_i$ can only decrease the payoff. We consider then

$$Z' \triangleq \min_{x_i} \sum_1^M \bar{p}_i S_i(x_i) \quad \text{subject to} \quad x_i \geqq \sqrt{N}, \qquad \sum_1^M x_i = N,$$

and have $Z' \leqq Z[1+o(1)]$.

Now choose $\bar{p}_i$ so that

$$\bar{p}_i S_i'(Np_i^*) = y; \qquad \sum_1^M \bar{p}_i = 1. \tag{3.8}$$

(Note that (3.8) implies equality in (3.6).)

Given this choice of $\bar{p}_i$ by the individual, we claim that $x_i = Np_i^*$, if feasible, (i.e. $Np_i^* \geqq \sqrt{N}$) gives the minimum payoff of $Z'$. This is because (3.8) shows that $x_i = Np_i^*$ satisfies the Langrange multiplier optimality condition; and the monotonicity condition on $S_i'$ along with the constraint $x_i \geqq \sqrt{N}$ shows this choice to be uniquely optimal. In that case, since (3.4) is satisfied we have

$$Z' = \sum_1^M \bar{p}_i S_i(Np_i^*) = \sum_1^M \bar{p}_i C[1+o(1)] = C[1+o(1)].$$

If $x_i = Np_i^*$ is not feasible then $Z' \geqq C[1+o(1)]$ must hold. In any case, it is then true that $Z \geqq C[1+o(1)]$ and finally $V(N) \geqq Z \geqq C[1+o(1)]$. It is easy to show that asymptotic optimality of the $\bar{p}_i$ is retained if equality in (3.6) is replaced with the asymptotic condition shown there; the reader may supply the argument. This completes the proof of Proposition 4.

Proposition 4 is an interesting result, for it says that the equilibrium non-cooperative payoff is asymptotically the minimum a player will obtain if he uses the $\bar{p}_i$ of (3.6). If it happened that the asymptotic equilibrium or in fact the exact equilibrium $\{p_i\}$ was not asymptotically optimal for the game $G$ (in that it guarantees a payoff $C(N)[1+o(1)]$) then serious doubt would be cast on its use, for the strategy of (3.6) would then be preferable, at least for large $N$. On the other hand (3.6) does give an asymptotically optimal strategy for $G$ in terms of solutions to the first order asymptotic equations ($\{p_i\}$ of course being one of them) and so the asymptotic equilibrium strategies are worthwhile for study in any case. It happens however, that in all the examples treated later, the asymptotic equilibrium does turn out to be asymptotically optimal for $G$. While this suggests a general result along these lines is possible, we have been unable to prove one. It would also be of interest to investigate in some examples how good the equilibrium strategy $\{p_i\}$ is for the game $G$ and how close $V(N)$ and $C(N)$ are when $N$ is taken as finite and fixed as opposed to the limiting case we discuss here. We do not provide such examples in this paper.

**4. Examples with $S_i$ algebraically decreasing.** The payoff functions we will consider in this section will be assumed most generally to have an asymptotic expansion of the form

$$S(x) \sim \sum_i a_i / x^{\alpha_i}, \quad \text{where } \{\alpha_i\} \text{ is an increasing sequence; } \alpha_i > 0. \tag{4.1}$$

Our object will be to replace the equations for the equilibrium (3.2) with their asymptotic expansions obtained by using Theorem 3 on $E[S_i(1+n_i)]$ and finally to calculate an expansion for the $p_i$.

PROPOSITION 5. *Let* $f(x)=(1+x)^{-\alpha}$. *Then for each even* $k$ *the conclusion of Theorem* 3 *holds* ((3.3)). *In particular* $E[f(X)]\sim 1/(1+\mu)^{\alpha}$.

*Proof.* For every integral $k\geqq 0$, $f^{(k+1)}(x)/f^{(k)}(x)=O(1/x)$, so Proposition 3 shows that $f^{(k)}(x)$ has EZARV and Theorem 3 may be applied when $k$ is even.

PROPOSITION 6. *If* $S(x)$ *satisfies* (4.1) *and* $X$ *is as in Theorem* 3 *then*

$$E[S(1+X)]\sim\sum_i a_i E[1/(1+X)^{\alpha_i}]. \tag{4.2}$$

*Proof.* Taking the first $k$ terms in (4.1) we have, from the definition of an asymptotic expansion $S(x)=\sum_{i=1}^{k} a_i/x^{\alpha_i}+f(x)$ where $f(x)=o(x^{-\alpha_k})$. Now for $x\geqq 1$, $f(x)$ is clearly bounded, say by $K$ and for any $\varepsilon>0$ we can choose an $m(\varepsilon)$ to satisfy $|f(x)|<\varepsilon/x^{-\alpha_k}$ for $x>m$. We then have $|f(x)|<K2^{m-x}+\varepsilon/x^{\alpha_k}$ and taking expectations

$$E[|f(1+X)|]<K2^{m-1}(1-p/2)^N+\varepsilon(1+Np)^{-\alpha_k}[1+o(1)]<2\varepsilon(Np)^{-\alpha_k}$$

for $N$ sufficiently large, where we used Proposition 5 to asymptotically evaluate $E[(1+X)^{-\alpha_K}]$. The last inequality however means precisely that

$$E[|f(1+X)|]=o[(Np)^{-\alpha_k}],$$

whence it follows that for each $k$

$$E[S(1+X)]=\sum_{i=1}^{k} a_i E[1/(1+X)^{\alpha_i}]+E[1/(1+X)^{\alpha_K}]o(1)$$

which is the definition of (4.2).

Using Proposition 5 we calculate as an example and for our future reference the three term expansion of $E[1/(1+X)^{a+1}]$ where $X\sim B(N-1,p)$ and $a>-1$:

$$E[1/(1+X)^{a+1}]=\frac{1}{(Np)^{a+1}}\left[1+\frac{(a^2+a)q}{2Np}+\frac{q([3a^4+14a^3+21a^2+10a]-[3a^4+10a^3+9a^2+2a]p)}{24(Np)^2}+O[1/(Np)^3]\right]. \tag{4.3}$$

The first two terms are easy to verify. The exponent $a+1$ was chosen because the case $a=0$ can be computed exactly as the first term in (4.3) with exponentially small error and (4.3) is easily seen as confirming this through three terms. As usual, $q=1-p$. We proceed now to treat some important examples.

A special case of particular interest is $S_i(x)=s_i/x$. Using elementary techniques, we can compute the relevant expectations explicitly in closed form:

$$L_i(p)=s_i\frac{1-[1-p(N)]^N}{Np(N)} \tag{4.4}$$

where we write $p(N)$ to remind ourselves that $p$ depends on $N$.

In (4.4) the term $[1-p(N)]^N$ is exponentially small, i.e. is $o[(Np)^{-k}]$ for any $k$. This is because $Np\to\infty$. Thus (4.4) has the asymptotic expansion

$$L_i(p)=s_i/Np+\text{e.s.t.}\quad(\text{e.s.t}=\text{exponentially small terms}).$$

We now replace the exact equations for the equilibrium, (3.2), by their asymptotic expansion. Solutions to these asymptotic equations will be denoted with an asterisk, $p_i^*$, $C^*$. We have

$$C^*=s_i/Np_i^*(N)+\text{e.s.t.};\qquad \sum_1^M p_i^*=1 \tag{4.5}$$

with solution

$$p_i^* = s_i / \sum_1^M s_j; \qquad C^* = \sum_1^M s_i / N. \tag{4.6}$$

Since $p_i^*$, $C^*$ satisfy the exact equations with exponentially small error we would hope that (4.6) is in error by a similar magnitude. We show this next.

PROPOSITION 7. *If $S_i(x) = s_i/x$ then the error in* (4.6) *is exponentially small with respect to powers of $1/N$, i.e. $|p_i - p_i^*| = o(N^{-k})$; $|C - C^*| = o(N^{-k})$ for every integer $k$.*

*Proof.* It is easy to show that $C = C^* + \text{e.s.t.}$ We have $L_i(p_i^*) = s_i/Np_i^* + \text{e.s.t.} = C^* + \text{e.s.t.}$ and $C = C^* + \text{e.s.t.}$ follows from Proposition 2. With this result established we can write, for any $k$, $C^* + o(N^{-k}) = s_i/Np_i + o[(Np_i)^{-k}]$. Assuming for some $i$ that $p_i > p_i^* + o(N^{2-k})$ then produces a contradiction and the proposition then follows.

The asymptotic solution (4.6) has intuitive appeal: One chooses alternative $i$ according to the relative weight of $s_i$. The common payoff to the players is the total amount of money divided by $N$. Because some alternatives may not be chosen by anyone (an event with exponentially small probability) the value of $C$ is actually slightly less than $C^*$. Now consider the game $G$ of individual vs. director. Proposition 4 shows that each player can asymptotically guarantee himself a fair share of one $N$th of all the money. On the other hand from (3.6), we have $S_i'(Np_i^*) = -s_i/(Np_i^*)^2 = -C^*/Np_i^* = -(C^*)^2/s_i$ whence asymptotically optimal for the game $G$ is the asymptotic equilibrium strategy $p_i^*$. Thus we have the very satisfying result that the intuitively appealing relative weight strategy asymptotically guarantees each player his fair share of the spoils even with hostile, collusive opponents.

A more general class of payoff functions, which we consider next, is

$$S_i(x) \sim \sum_{k=0} \frac{s_{ik}}{x^{\alpha+k}}, \quad \text{where } s_{i0} \neq 0. \tag{4.7}$$

The payoff functions thus decrease at the same algebraic rate and are $O(1)$ with respect to each other.

The first order asymptotic equations are

$$C^* = [s_{i0}/(Np_i^*)^\alpha][1 + o(1)]$$

with solution

$$p_i^* = (s_{i0})^{1/\alpha} \Big/ \sum_1^M (s_{j0})^{1/\alpha}; \qquad C^* = \left[\sum_1^M (s_{i0})^{1/\alpha}/N\right]^\alpha. \tag{4.8}$$

It is not difficult to show that $p_i^*$ and $C^*$ are in fact asymptotic solutions for the equilibrium; we will prove this more generally for the expansion discussed below. We observe from (4.8) that if $\alpha$ is large then we have roughly $p_i^* \approx 1/M$, $C^* \approx [\prod_1^M s_{i0}]^{1/M}(M/N)^\alpha$. A later example with payoff functions exponentially decreasing will have solutions asymptotically exhibiting this behavior.

We can treat the payoff functions in (4.7) in greater asymptotic detail by replacing $E[S_i(1+X)]$ with its asymptotic expansion calculated using Propositions 5 and 6 and substituting into the equations (3.2) for the equilibrium. Asymptotic solutions of these equations will then hopefully give us higher order approximations to the exact solution.

The expansion of $E[S_i(1+X)]$ (where $X \sim B(N-1, p)$) is

$$E[S_i(1+X)] \sim \sum_{j=0} \frac{a_{ij}(p)}{(Np)^{\alpha+j}} \triangleq L_i^*(p) \tag{4.9}$$

where $a_{ij}(p)$ is a polynomial of degree $j$ in $p$, with coefficients depending on $\alpha$ and $s_{ik}$ for $k \leqq j$. Truncating the expansion (4.9) after $k$ terms results in an error which is $0[1/(Np)^{\alpha+k+1}]$. We then seek solutions of the asymptotic equations for the equilibrium

$$C^* \sim L^*(p_i^*), \qquad i = 1, \cdots, M; \qquad \sum_1^M p_i^* = 1. \tag{4.10}$$

To solve (4.10) we assume asymptotic expansions for $p_i^*$ and $C^*$ of the forms $p_i^* \sim \sum_{j=0} b_{ij}/N^j$ and $C^* \sim \sum_{j=0} c_j/N^{\alpha+j}$. Then we reexpand (4.10) in inverse powers of $N$ and equate coefficients. The result is a sequence of equations for the coefficients. The lowest order equations (resulting from equating coefficients of $1/N^\alpha$ in $C^* \sim L^*(p_i^*)$ and $1/N^0$ in $\sum_1^M p_i^* = 1$) provide the first order solution given by (4.8) with $b_{i0}$ in place of $p_i^*$ and $c_0$ in place of $C^*$. Successive sets of unknown coefficients can then be explicitly computed in a recursive fashion as the solutions of simple linear equations. We explain further directly.

In the reexpansion of (4.10) using the assumed expansion for $p_i^*$, the coefficient $b_{ij}$ first appears in the coefficient of $1/N^{\alpha+j}$ as a result of the expansion

$$(Np_i^*)^{-\alpha} \sim \left[1 - \alpha \sum_{k=0} \frac{b_{ik}}{b_{i0} N^k} + \cdots\right] \Big/ (N^\alpha b_{i0}^\alpha).$$

Thus $b_{ij}$ first appears linearly. Also appearing in the coefficient of $1/N^{\alpha+j}$ may be $\{b_{ik}\}$ for $k < j$. Recalling that (4.8) gives the solution for $b_{i0}$ and $c_0$, we inductively assume that $b_{ik}$, $i = 1, \cdots, M$; $k = 1, \cdots, j-1$ and $c_k$, $k = 1, \cdots, j-1$ have been computed. Equating the coefficients of $1/N^{\alpha+j}$ in $C^* \sim L^*(p_i^*)$ and $1/N^j$ in $\sum_1^M p_i^* = 1$ gives $M+1$ equations for the $m+1$ unknowns $b_{ij}$, $i = 1, \cdots, M$ and $c_j$ of the form

$$c_j/c_0 = f_i(b_{ik}, k < j)/b_{i0} - \alpha b_{ij}/b_{i0}, \qquad i = 1, \cdots, M; \qquad \sum_{i=1}^M b_{ij} = 0, \tag{4.11}$$

where $f_i$ is a multinomial in its arguments. The unknown coefficients can then be trivially solved for, e.g. sum the set of $M$ equations over $i$ using $\sum_1^M b_{ij} = 0$ to solve for $c_j$, then substitute back and solve for the $b_{ij}$. We have therefore shown that the coefficients in the asymptotic expansions of $p_i^*$ and $C^*$ can be recursively solved for in a very simple manner. The explicit equations for $j = 1$ are:

$$c_1/c_0 = \frac{(1-b_{i0})(\alpha-1)\alpha + 2s_{i1}/s_{i0}}{2b_{i0}} - \frac{\alpha b_{i1}}{b_{i0}}; \qquad \sum_{i=1}^M b_{i1} = 0.$$

Computations for higher order terms, while elementary, are also messy. In practice of course, the work is simplified in dealing with the particular parameters of the problem at hand. It would seem, in general, that a two-term expansion ought to be adequate for large enough $N$. One would want the second term in the event that one's expenses in playing the game are of the order of the equilibrium payoff so that the net payoff might be the same order as the second term in the expansion of $C^*$.

We prove now that $p_i^*$ and $C^*$ are in fact asymptotic expansions of $p_i$ and $C$.

THEOREM 4. *Let $S_i(x)$ be given by (4.7) and suppose $C^*, p_i^*$ satisfy $C^* + O(N^{-\alpha-k}) = L_i(p_i^*)$, $i = 1, \cdots, M$ and $k \geqq 1$ a fixed integer;*

$$\sum_1^M p_i^* = 1; \qquad Np_i^*(1-p_i^*) \to \infty.$$

*Then $C = C^* + O(N^{-\alpha-k})$ and $p_i = p_i^* + O(N^{-k})$, $i = 1, \cdots, M$.*

*Proof.* By Proposition 2, we must have $C = C^* + O(N^{-\alpha-k})$. If $p_i = p_i^* + O(N^{-k})$ does not hold then given any $R > 0$ there are infinitely many $N$ such that for some index $j(N)$, $p_j(N) > p_j^* + RN^{-k}$. Then $L_j(p_j) < L_j(p_j^* + RN^{-k})$. Using (4.9) to compare the asymptotic expansion of $L_j(p_j^*)$ and $L_j(p_j^* + RN^{-k})$ we see that for $N$ sufficiently large (possibly depending on $R$)

$$L_j(p_j^* + RN^{-k}) < L_j(p_j^*) - cRN^{-\alpha-k} = C^* + O(N^{-\alpha-k}) - cRN^{-\alpha-k}$$

and $c > 0$ is a constant independent of $N$ and $R$. Finally,

$$C^* + O(N^{-\alpha-k}) = C = L_j(p_j) < L_j(p_j^* + RN^{-k}) < C^* + O(N^{-\alpha-k}) - cRN^{-\alpha-k}$$

and for $R$ sufficiently large a contradiction is obtained.

Consider now the game $G$ for the class of payoff functions in (4.7). In order to apply (3.6) we first consider $S_i(x) = s_i/x^\alpha$. The first order equilibrium for the non-cooperative game is still given by (4.8), with $s_i$ in place of $s_{i0}$. Applying (3.6), we have $S_i'(Np_i^*) = -\alpha s_i/(Np_i^*)^{\alpha+1} = -\alpha C^*/(Np_i^*)$ so that $\bar{p}_i = p_i^*$ is asymptotically optimal for $G$. Consider now the general case of (4.7). Recalling the proof of Proposition 4, if the individual uses $p_i^*$ (or any other strategy for that matter) the optimal allocation by the director can be assumed at least equal to $\sqrt{N}$ for each alternative for purposes of computing the asymptotic payoff to the individual. On the other hand such asymptotic payoff can then be asymptotically obtained by replacing $S_i(x)$ with $s_{i0}/x^\alpha$ and in this latter case the individual receives at least $C[1 + o(1)]$ if he uses $p_i^*$. Thus $p_i^*$ is still optimal asymptotically in the general case.

PROPOSITION 8. *For payoff functions of the form* (4.7), *the asymptotic equilibrium of* (4.8) *is asymptotically optimal for the game G and guarantees a payoff* $C[1 + o(1)]$.

The next class of payoff functions we consider is $S_i(x) = s_i/x^{\alpha_i} + \text{e.s.t.}$ We assume the $\alpha_i$ are not equal and obey the ordering $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_M$. Determining an asymptotic expansion for the $p_i$ in this case is a much more difficult task. We are unable even to say what form the expansion has other than $p_i \sim \sum_j b_{ij}/N^{\beta_{ij}}$ and similarly for $C$. These difficulties are demonstrated in calculating a two-term expansion below.

Recalling from (4.3) the two-term expansion of $E[1/(1+n_i)^{\alpha_i}]$:

$$E[s_i/(1+n_i)^{\alpha_i}] = \frac{s_i}{(Np_i)^{\alpha_i}}\left[1 + \frac{\alpha_i(\alpha_i - 1)q_i}{2Np_i} + O[1/(Np_i)^2]\right]. \tag{4.12}$$

The equilibrium equations are $C = E[s_i/(1+n_i)^{\alpha_i}] = L_i(p_i)$; $\sum_1^M p_i = 1$. It is easily seen that only $p_1$ can be bounded away from zero for large $N$ so that $p_1 \to 1$; $p_i \to 0$, $i \neq 1$.

If $p_1 \to 1$ then $C = (s_1/N^{\alpha_1})[1 + o(1)]$. To satisfy the equations $C = L_i(p_i)$ for $i \neq 1$ requires that

$$p_i = \frac{b_i}{N^{\beta_i}}[1 + r_i(N)] \quad \text{where } b_i = (s_i/s_1)^{1/\alpha_i}, \quad \beta_i = 1 - \alpha_1/\alpha_i, \quad r_i(N) = o(1) \tag{4.13}$$

and (4.13) holds for all $i$. Note that $0 < \beta_i < 1$ for $i \neq 1$.

Observe next that the constraint $\sum_1^M p_i = 1$ can only be satisfied if $-b_2/N^{\beta_2}$ appears as the second term in the expansion of $p_1$. In that case we can evaluate $C$ to two terms as

$$C \sim \frac{s_1}{N^{\alpha_1}}[1 + \alpha_1 b_2/N^{\beta_2}]. \tag{4.14}$$

The second terms in the expansions of $p_i$ for $i > 1$ are now obtained by satisfying (4.14).

Substituting (4.13) into (4.12) and retaining the two largest terms:

$$L_i(p_i) \sim \frac{s_1}{N^{\alpha_1}}\left[\frac{1}{[1+r_i(N)]^{\alpha_i}} + \frac{\alpha_i(\alpha_i-1)}{2b_i N^{\alpha_1/\alpha_i}}\right]. \tag{4.15}$$

The $r_i(N)$ are then chosen to give the same two-term expansion as in (4.14).

If $\alpha_1/\alpha_i > \beta_2$ then we must pick

$$r_i(N) \sim -b_2\alpha_1/(\alpha_i N^{\beta_2}); \qquad i: \alpha_1/\alpha_i > \beta_2, \tag{4.16}$$

and if $\alpha_1/\alpha_i < \beta_2$ then we must pick $r_i(N)$ to cancel the other term in the brackets in (4.15), i.e.

$$r_i(N) \sim \frac{\alpha_i - 1}{2b_i N^{\alpha_i/\alpha_i}}; \qquad i: \alpha_i/\alpha_i < \beta_2. \tag{4.17}$$

(The case $\alpha_1/\alpha_i = \beta_2$ may be supplied by the reader.)

The two-term expansion for the equilibrium strategy and payoff is then given by (4.13), (4.16) and (4.17); and (4.14) respectively.

The further development of the asymptotic expansions is apparently a rather messy exercise though no further conceptual difficulties should arise. We have seen that even the form of the expansions depends on the particular values of the $\alpha_i$.

Consider now the game $G$ with payoff functions of the form $S_i = s_i/x^{\alpha_i}$ + e.s.t. From (4.14), $C = (s_1/N^{\alpha_1})[1+o(1)] = S_1(N)[1+o(1)]$. Thus for this game it is a trivial observation that any strategy for the individual satisfying $\bar{p}_1 \to 1$ guarantees him a payoff $C[1+o(1)]$ and thus is asymptotically optimal. We note though that the strategy of (3.6) is not asymptotic to the equilibrium strategy (though both satisfy $\bar{p}_1 \to 1$ and are thus both asymptotically optimal for $G$). Indeed, (3.6) provides $\bar{p}_i \sim (\alpha_1/\alpha_i)p_i^*$ as the asymptotically optimal solution. It is not immediately clear which strategy for the individual provides a better guaranteed payoff to second order—we have not investigated this question.

**5. Two more examples.** Below we find the first order asymptotic equilibrium strategy for two examples which exhibit interesting behaviors. In the first example, although the payoff functions are asymptotically equal (i.e. have limiting ratio of 1) the first order equilibrium strategy does not share the same property. In the second example although the payoff functions are not asymptotically equal the first order equilibrium strategy has the $p_i$ asymptotically equal to $1/M$.

The first example has payoff functions

$$S_i(x) = \frac{1}{\log(x+1)} + \frac{a_i}{[\log(x+1)]^2}. \tag{5.1}$$

It is easily verified that $[\log(1+x)]^k$ has EZARV for any $k$. The analysis below can be equally well carried out if (5.1) is only an asymptotic relationship rather than equality. Indeed it is easy to show as in Proposition 6 that one could in that case simply take the expectation of both sides to obtain an asymptotic expression for $E[S_i(1+X)]$.

The first order asymptotic equations are

$$C^*[1+o(1)] = 1/\log(Np_i^*); \qquad \sum_1^M p_i^* = 1. \tag{5.2}$$

We observe that (5.2) is satisfied by *any* constant distribution $\{p_i^*\}$ $(p_i^* \neq 0)$. Thus the first order equilibrium strategy is not determined by (5.2) although Proposition 2 does imply that $C = (1/\log(N))[1+o(1)]$. (Or more simply, since one of the $p_i$ must be

greater than $1/M$ infinitely often we must have $C \sim 1/\log(N)$.) This degeneracy can only be resolved by including the second order terms in the expansions of $E[S_i(1+n_i)]$.

We have, using Theorem 3,

$$E[1/\log(2+n_i)] = 1/\log[(N-1)p_i+2] + o(1/Np_i) = 1/\log(Np_i) + o(1/Np_i).$$

The requirement $C \sim 1/\log(N)$ leads to $\log(p_i)/\log(N) = o(1)$ so that we may write

$$\frac{1}{\log(Np_i)} = \frac{1}{\log(N)[1+\log(p_i)/\log(N)]} \sim \frac{1}{\log(N)} - \frac{\log(p_i)}{[\log(N)]^2}.$$

A two-term expansion of $E[S_i(1+n_i)] = C$ using the above gives asymptotic equations:

$$C^* = \frac{1}{\log(N)} + \frac{a_i - \log(p_i^*)}{[\log(N)]^2}[1+o(1)]. \tag{5.3}$$

If we put $C^* = 1/\log(N) + c/[\log(N)]^2$ and then choose $p_i^*$ to satisfy

$$c = a_i - \log(p_i^*); \qquad \sum_1^M p_i^* = 1, \tag{5.4}$$

then (5.3) will be satisfied. Our usual technique will show that in fact $C \sim C^*$, $p_i \sim p_i^*$ is satisfied by the equilibrium. (That is we assume $p_i = p_i^* + o(1)$ doesn't hold. Then we must have $p_i > p_i^* + \varepsilon$, $p_j < p_j^* - \varepsilon$ for infinitely many $N$, where $i = i(N)$, $j = j(N)$. This will produce a contradiction in (5.3).)

From (5.4) our asymptotic equilibrium is then

$$C = \frac{1}{\log(N)} + \frac{c}{[\log(N)]^2}[1+o(1)]; \qquad p_i = \frac{\exp(a_i)}{\sum_i \exp(a_i)}[1+o(1)] \tag{5.5}$$

where $c = \log[\sum_{i=1}^M \exp(a_i)]$.

We see then that in general, although the payoff functions are asymptotically equal, the $p_i$ will not be.

Considering the game $G$ for these payoff functions, we find the situation rather uninteresting to first order as was the case with the previous example. Any constant distribution used by the individual will asymptotically guarantee him a payoff of $C[1+o(1)]$. An investigation into the possible second order optimality of the equilibrium might prove worthwhile but we will not attempt that here. We note that (3.6) admits the asymptotic equilibrium as a solution.

Our next example has payoff functions which are exponentially decreasing. We will see that such behavior in the payoff functions leads to an equilibrium strategy which is asymptotically equally distributed, provided the payoff functions are of the same order of magnitude.

We consider $S_i(x) = s_i \exp(-x^a)$. If $a < \frac{1}{2}$, as we shall assume, then $S_i(x)$ has EZARV by Proposition 3. The first order asymptotic equations are

$$C^* = s_i \exp[-(Np_i^*)^a][1+o(1)]; \qquad \sum_1^M p_i^* = 1. \tag{5.6}$$

We look for a solution in the form

$$C^* = c \exp[-(N/M)^a]; \qquad p_i^* = [1+b_i/N^a]/M.$$

Then (5.6) will be satisfied if

$$c = s_i \exp[-ab_i/M^a]; \qquad \sum_1^M b_i = 0$$

which in turn implies

$$c = \left[\prod_{i=1}^{M} s_i\right]^{1/M}; \qquad b_i = \frac{M^a}{a}\log\left(\frac{s_i}{c}\right).$$

Again, it is not difficult to prove that $p_i^*$ gives $p_i$ correct to second order and that $C$ is asymptotic to $C^*$. Thus $p_i$ is asymptotically equidistributed among the alternatives. Applying (3.6) to this example shows that $\bar{p}_i = 1/M$ is asymptotically optimal for the game $G$.

**6. Application to a dynamic game.** As a final application of our asymptotic analysis we consider the following game: Again we have $N$ players, $M$ alternatives and the same hypotheses on the payoff functions, but in this game the alternatives are chosen sequentially in the order $1, 2, \cdots, M$. The number of players remaining at any stage is known to each remaining participant.

We consider first the exact calculation of the symmetric equilibrium. This is easily accomplished, in principle, using backward induction. Suppose we are at the stage of the game where only two alternatives remain with payoff functions $S_{M-1}$ and $S_M$ and $j$ players are left in the game. This is clearly our original game with two alternatives, with $j$ taking the place formerly occupied by $N$. We can therefore calculate the equilibrium strategy and payoff as a function of $j$. Denote the payoff by $C_2(j)$. Taking one step backwards, consider the stage of the game where three alternatives remain, along with $j$ players. It is not difficult to see that the equilibrium strategy at this stage, which consists of a probability of choosing alternative $M-2$ as opposed to passing, may be found by considering a two alternative game with payoff functions $S_{M-2}$ and $C_2$ and $j$ players. Solving this game gives the equilibrium payoff function $C_3(j)$ and the equilibrium probability of choosing alternative $M-2$ as a function of $j$. The procedure continues until the $M$ stage strategy is determined.

We inquire next as to the asymptotic behavior of the equilibrium. It turns out that, roughly speaking, the added information of the sequential game is not of any value to first order. Consider first the sequential game played as individual versus director. It is clear that the sequential nature of the game gives the director no new information: his allocations may as well be determined prior to the start of the game. The value of the game to the individual must be larger for the sequential game: after all, he can always use his maximin nonsequential strategy. These observations hold for all $N$. For large $N$ however, the director can hold the individual to an asymptotic payoff of $C(N)$ by using the deterministic strategy of allocating $[Np_i^*]$ players to alternative $i$, where $[\ ]$ is the greatest integer function and $p_i^*$ is any solution to the first order asymptotic equations (3.4). On the other hand, under the conclusions of Proposition 4, namely (3.5) and (3.6), which hold in particular for our examples, the individual can always guarantee himself $C(N)[1+o(1)]$ with a non-sequential strategy.

In practice of course, a sequential strategy is advisable. Asymptotically, one would expect that some improvement could be gained if the individual simply updates his asymptotically optimal strategy at each stage, taking into account the number of players actually remaining. Against an optimal director however, the relative payoff increase is at most $o(1)$.

As for the noncooperative symmetric equilibrium in the sequential game, it can probably be shown that if each player uses the asymptotic equilibrium strategy in a sequential manner (updating $N$ after each stage) the remaining player will be held to a payoff $C[1+o(1)]$. Indeed it is probably true that if $N-1$ players simply ignore the sequential nature of the game and play the asymptotic equilibrium, the remaining

player will still only obtain $C[1+o(1)]$. We will not try to prove these conjectures here for any general class of payoff functions. However, one can perform an explicit calculation in the examples treated; we briefly consider $S_i = s_i/n^\alpha$, $i = 1, \cdots, M$.

Let $P_i(j)$ be the equilibrium probability of choosing the alternative when $i$ stages remain; thus $1-P_i$ is the probability of passing when $i$ stages remain. Using (4.8) we have

$$P_2(j) \sim (s_{M-1})^{1/\alpha} \Big/ \sum_{M-1}^{M} (s_k)^{1/\alpha}; \qquad C_2(j) \sim \left[\sum_{M-1}^{M} (s_k)^{1/\alpha}\right]^\alpha \Big/ (j)^\alpha. \tag{6.1}$$

Thus, when two stages remain and $j$ is large, $P_2$ and $C_2$ approach the asymptotic equilibrium values; but this was obvious to begin with.

Consider now three stages remaining. $P_3$ and $C_3$ are calculated, as described above, using the payoff functions $S_{M-2}$ and $C_2$. We showed however that the asymptotic behavior of $P_3$ and $C_3$ depends only on the asymptotic behavior of $S_{M-2}$ and $C_2$, which we have calculated above. (The asymptotic specification of (4.7) lead to (4.8).) Note that the payoff function $C_2(j)$ is of the form (4.7). We easily calculate

$$P_3(j) \sim (s_{M-2})^{1/\alpha} \Big/ \sum_{M-2}^{M} (s_k)^{1/\alpha}; \qquad C_3(j) \sim \left[\sum_{M-2}^{M} (s_k)^{1/\alpha}\right]^\alpha \Big/ (j)^\alpha. \tag{6.2}$$

Again, the results in (6.2) are the same as those for the non-sequential equilibrium. Continuing the induction backwards to alternative 1, we will obtain

$$P_M(j) \sim (s_1)^{1/\alpha} \Big/ \sum_{1}^{M} (s_k)^{1/\alpha}; \qquad C_M(j) \sim \left[\sum_{1}^{M} (s_k)^{1/\alpha}\right]^\alpha \Big/ (j)^\alpha. \tag{6.3}$$

Suppose now that $N-1$ players use the asymptotic sequential strategy above; that is, $N-1$ players use $P_M = (s_1)^{1/\alpha}/\sum_1^M (s_k)^{1/\alpha}$; $P_{M-1} = (s_2)^{1/\alpha}/\sum_2^M (s_k)^{1/\alpha}$; $\cdots$; $P_2 = (s_{M-1})^{1/\alpha}/\sum_{M-1}^M (s_k)^{1/\alpha}$. Then the remaining player, by the computations above, is held to $C_M(j)[1+o(1)] = C(j)[1+o(1)]$ where $C$ is the nonsequential symmetric equilibrium payoff. But that strategy is actually nonsequential so the alternatives may be chosen at the beginning; the resulting unconditioned probabilities are precisely the asymptotic equilibrium probabilities of (4.8). Thus we have obtained:

For $S_i \sim s_i/n^\alpha$, a single player is asymptotically held to the nonsequential equilibrium payoff if the remaining $N-1$ players use the asymptotic equilibrium strategy of the nonsequential game.

Of interest for these sequential games would be first, a higher order asymptotic calculation of the effects of the sequential assumption; and secondly some numerical examples comparing sequential and non-sequential, exact and asymptotic, aspects of these games.

**7. Further questions.** We have, along the way, suggested some further areas of investigation in connection with the class of games considered in this paper. In this section we briefly consider some issues peripheral to the analysis carried out. We make some unsatisfactory remarks concerning the accuracy and utility of our asymptotic results; and then we present some alternative asymptotic embeddings which could be useful in situations where our asymptotic results do not apply.

Estimating the error in an asymptotic approximation is always a difficult task. The general rule of thumb is to estimate the error as the size of the first neglected term in the asymptotic expansion and this could be employed for our results, where more than a first order approximation has been obtained. An alternative is to investigate the accuracy with which the equations $C = L_i(p_i)$ are satisfied. More precisely, the first

neglected term in the expansion of $C$ alone may be a better indicator of the operational value of the asymptotic strategies. Even cruder is to take $p_i = O(1/M)$ and estimate relative error as $1/Np_i$, in which case $M/N$ is seen to be the important gauge of asymptotic accuracy. Based on nothing at all, we will suggest that $M/N \approx \frac{1}{10}$ gives reasonable approximations by the asymptotic results.

The asymptotic equilibrium may of course always be profitably used as a first approximation for a numerical solution. If one is interested in computing the optimal strategy in the game $G$ discussed previously, then again the asymptotic equilibrium is a good starting point, by itself or thru the use of Proposition 4. Some numerical work would be helpful in clarifying the relationships between the exact equilibrium, the asymptotic equilibrium, and the optimal strategy and payoff in the game $G$.

Summing up our thoughts on the utility of the asymptotic analysis in this paper, we feel we have developed, in analytic form, some interesting qualitative properties of optimal strategies in the games considered. For large $M/N$ our results are probably accurate but are probably best and most safely used as part of a more detailed operational and numerical analysis of the particular situation at hand.

When we apply asymptotic analysis to a finite game as in this paper, we embed the game in a sequence of games which exhibits limiting behavior. It may not always be clear however what the most "natural" such sequence is for a particular game which, after all, is described in terms of a finite set of parameters. In what follows we discuss a few alternative sequences (other than that considered previously) which may be more suitable in certain instances and some associated questions.

(a) It is entirely possible that the $S_i(n)$ will not be given to us in analytic form (e.g. $s_i/n$, etc.) but rather will simply be, for each $i$, a set of $N$ given numbers for which no "natural" analytic extension presents itself. How then, if say $N/M$ is large, shall we (or can we) apply the asymptotic results of this paper? For instance, might we try and fit $S_i(n)$ with a finite polynomial in $1/n$ and then apply the results? What would constitute a sufficiently good fit for the asymptotic results to apply? There are actually two parts to this last question: How good is the approximation; and is $N$ large enough that the solution depends only on the asymptotic behavior of the analytic extension approximating $S_i(n)$?

(b) Consider a game where say $M = 10$ and $N = 30$. Then the accuracy of our asymptotic results are suspect since $N/M = 3$ is not very large. On the other hand, one feels that 30 is a "large enough" number for some sort of asymptotic behavior to manifest itself. Clearly, what is called for here is an asymptotic sequence in which $M$ depends on $N$, say $M = N/3$ and $N \to \infty$. But how does one (and can one) embed the $S_i(n)$ for $i = 1, \cdots, 10$ in an infinite sequence of $S_i(n)$ on which asymptotic analysis can be profitably carried out?

(c) Most generally, one may want to consider a sequence in which everything depends on $N$; in particular, the payoff functions may be functions $S_i(n;N)$ of $N$. Consider for instance this scenario: $N$ commuters await a train with $M$ cars, car $i$ containing $k_i$ seats. Assume each commuter independently chooses one car to enter and his payoff is one if he gets a seat, zero otherwise (or equivalently the payoff is the probability of getting a seat). This leads to $S_i(n) = \min(1, k_i/n)$. Typically, while $N/M$ is large we also expect that $k_i = O(N/M)$ will hold, i.e. that the total number of seats is less than $N$, but $O(N)$ rather than $o(N)$. It is clear that this requires the scaling of the $k_i$ with respect to $N$, say $k_i = r_iN/M$ and our asymptotic sequence of payoff functions is $S_i(n; N) = \min(1, r_iN/(Mn))$ where $r_i$ is fixed by the given values (e.g. if $N = 200$, $M = 10$, $k_1 = 12$ then $r_1 = .6$). clearly the asymptotic results of this paper will not apply; the analysis is more difficult.

The problem of asymptotic embedding poses some interesting mathematical questions. Beyond this particular game, one may ask whether some classes of "large" problems might perhaps be investigated by embedding them in some appropriate sequence possessing asymptotic behavior and amenable to analytic techniques of analysis.

## REFERENCES

[1] H. R. DIAMOND, *Asymptotic expectations for a class of functions of a binomial random variable*, Unpublished manuscript.

[2] J. NASH, *Non-cooperative games*, Ann. of Math., 54 (1951), pp. 286–290.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0001/$01.00/0

# ORTHOGONAL POLYNOMIALS IN TWO VARIABLES OF $q$-HAHN AND $q$-JACOBI TYPE*

CHARLES F. DUNKL†

**Abstract.** Two families of orthogonal polynomials in two discrete variables are constructed for a weight function of $q$-hypergeometric type. The polynomials are expressed in terms of $q$-Hahn polynomials. The connection coefficients between the two families involve Askey and Wilson's ${}_4\phi_3$-polynomials, which are certain balanced, terminating basic hypergeometric series. The method is to consider functions on the lattice of subspaces of a finite vector space which are invariant under the subgroup of the corresponding general linear group which fixes a pair of nested subspaces.

By limiting methods, corresponding results are obtained for Andrews and Askey's little $q$-Jacobi polynomials, which are orthogonal on a countable compact set. The classical Hahn version of the theory, where the underlying group is the symmetric group, has been worked out by the author in a previous paper.

**Introduction.** Results on orthogonal $q$-polynomials are often suggested by the theory of the corresponding ordinary polynomials. The theories of the $q$-Hahn and classical Hahn polynomials have some close parallels; a group-theoretical explanation of this is the correspondence between the Hecke algebras of a finite general linear group with respect to parabolic subgroups, and of the associated Weyl group (a symmetric group) with respect to Young subgroups. To put it another way, $q$-Hahn and classical Hahn polynomials appear in the analysis of functions on the lattice of subspaces of a finite vector space, and the lattice of subsets of a finite set, respectively.

By analyzing the action of the symmetric group on pairs of disjoint sets, the author [8] discovered interesting connections among Hahn polynomials, a simple first-order difference equation, both in two variables, and a family of orthogonal polynomials expressed as balanced ${}_4F_3$-functions (closely related to Racah's $6-j$ symbols). Askey and Wilson [3] found a corresponding $q$-family for these (terminating, balanced ${}_4\phi_3$-functions), and it seemed that a similar setting for them could be found on the finite general linear group. Indeed, that is done in this paper; it turns out there is a neat interplay between $q$-Hahn polynomials in two variables and Askey and Wilson's ${}_4\phi_3$-type. Similar results hold for a family of $q$-Jacobi polynomials in two variables, by a limiting argument.

We give a short exposition on the relation of the symmetric group to the finite general linear group. Suppose $G$ is the group of nonsingular $N\times N$ matrices over a finite field of order $q$; then $G$ can be expressed as $BWB$, where $B$ is the subgroup of upper-triangular matrices and $W$ is the group of $N\times N$ permutation matrices (the Weyl group, isomorphic to $S_N$, the symmetric group on $N$ objects). The group $W$ is generated by $R=\{w_1, w_2, \cdots, w_{N-1}\}$, where $w_j$ corresponds to the transposition $(j, j+1)$. If $W_J$ is the group generated by $J\subset R$ then $BW_JB$ is a parabolic subgroup of $G$. Deleting one point from $R$, say $J_n := R\backslash\{(n, n+1)\}$, produces a maximal parabolic subgroup $BW_{J_n}B$, the subgroup of $G$ fixing the subspace spanned by the first $n$ basis vectors. There is a linear isomorphism between the functions on $G$ bi-invariant under $BW_{J_n}B$ (these form the Hecke algebra) and the functions on $W\cong S_N$ bi-invariant under $W_{J_n}\cong S_n\times S_{N-n}$ (see Curtis, Iwahori and Kilmoyer [5]). These two spaces are spanned by corresponding families of $q$-Hahn and classical Hahn polynomials (see Dunkl [6] and [7]).

* Received by the editors July 19, 1979.

† Department of Mathematics, University of Virginia, Charlottesville, Virginia 22903. This research was supported in part by the National Science Foundation under Grant MCS-76-07022 A01.

The results on two variables mentioned above came from functions on $S_N$ which are $S_M \times S_{N-M}$-invariant on one side, and $S_a \times S_b \times S_c$-invariant on the other ($a+b+c=N$). For a $q$-analogue we look at the subgroup $H = BW_JB$ of $G$ with $J = R\setminus\{(a, a+1), (a+b, a+b+1)\} = J_a \cap J_{a+b}$ (thus $W_J \cong S_a \times S_b \times S_c$). Then $H$ is the set of nonsingular matrices of the form

$$\begin{array}{c} \\ a \\ b \\ c \end{array}\begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \end{array}.$$

In [8] the action of larger subgroups like $W_{J_a} \cong S_a \times S_{b+c}$ was used to construct orthogonal bases (by the splitting into inequivalent representations). The ${}_4F_3$-functions appeared in the connection coefficients between the bases coming from $S_a \times S_{b+c}$ and $S_{a+b} \times S_c$ respectively. Actually a third basis was obtained, as well, for the subgroup $S_{a+c} \times S_b$ but this is not of the form $W_J$ in the present set-up. Here there is a difference between the $q$- and ordinary theories.

Thus the approach will be to analyze the $H$-invariant functions with respect to the representations of $BW_{J_a}B$ and $BW_{J_{a+b}}B$. The elements of these subgroups look like

$$\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} * & * & * \\ * & * & * \\ 0 & 0 & * \end{bmatrix} \quad \text{respectively.}$$

We will be guided by the sequence of development and calculation used in [8], so the aim here is to present the actual results, with brief indications of the calculations. (Thus theorems and formulas will be stated and proved here, but the reasons why their existence or form was conjectured are drawn from [8] and will not be further elaborated.)

We begin in § 1 with some basic facts about $q$-calculations and other details. Then in § 2, we do the analysis on the groups mentioned above, set up the difference equation satisfied by functions which are invariant and in irreducible submodules, and determine two orthogonal bases for the space of solutions, expressed in $q$-Hahn polynomials. In § 3 the transformation from one base to another is worked out, in terms of ${}_4\phi_3$-polynomials. In § 4 the spaces of invariant functions from § 3 are transferred to the collection of $M$-dimensional subspaces of the underlying vector space as domain. This produces two general families of $q$-Hahn polynomials in two variables, orthogonal for the same weight function, and which are again related by the ${}_4\phi_3$-polynomials. Finally in § 5 a limiting argument leads to two families of "little" $q$-Jacobi polynomials (the term is from Andrews and Askey [1]); these are orthogonal on a countably infinite compact set, and likewise the transformation from one base to another involves the ${}_4\phi_3$. Also in this section we show how some transformations for one-variable $q$-Jacobi polynomials can be derived from two-variable relations.

**1. Basic facts and notation.** We list here some notation, sums, and transformations important in $q$-calculations. Fix a number $q > 1$ (mostly it will be a power of a prime number). The $q$-analogue of the shifted factorial is

$$(a; q^{-1})_n := (1-a)(1-aq^{-1}) \cdots (1-q^{1-n}) \quad \text{for } n = 1, 2, \cdots,$$

$$(a; q^{-1})_0 := 1.$$

The symbol $(a;q)_n := \prod_{j=1}^{n}(1-aq^{j-1})$ may also appear. The two are related by

$$(q^a;q^{-1})_n = (-1)^n q^{an-\binom{n}{2}}(q^{-a+n-1};q^{-1})_n$$
$$= (-1)^n q^{an-\binom{n}{2}}(q^{-a};q)_n.$$

For convenience we will write $(a)_n$ for $(a;q^{-1})_n$ throughout. Since $q>1$, the symbol $(a)_\infty := \lim_{n\to\infty}(a)_n$ is also meaningful; in fact it is an analytic function in $a$. The statement $\lim_{q\to 1}(q^a;q)_n/(1-q)^n = a(a+1)\cdots(a+n-1)$ shows the relation to ordinary shifted factorials.

The basic hypergeometric series for parameters $p=0, 1, 2, \cdots, \alpha_1, \cdots, \alpha_{p+1}, \beta_1, \cdots, \beta_p$ is defined by

$${}_{p+1}\phi_p\left(\begin{matrix}\alpha_1, \alpha_2, \cdots, \alpha_{p+1}\\ \beta_1, \cdots, \beta_p\end{matrix}; q^{-1}, x\right) := \sum_{n=0}^{\infty}\frac{(\alpha_1)_n\cdots(\alpha_{p+1})_n}{(\beta_1)_n\cdots(\beta_p)_n(q^{-1})_n}x^n.$$

The series terminates if one of the $\alpha_i$ equals $q^m$, $m=0, 1, \cdots$. The series is said to be balanced if $x=q^{-1}$ and $\alpha_1\alpha_2\cdots\alpha_{p+1}q^{-1} = \beta_1\beta_2\cdots\beta_p$.

The $q$-binomial coefficient is also useful; for $k=0, 1, 2, \cdots$, a real number $x$, set

$$\binom{x}{k}_q := \frac{(q^x)_k}{(q^k)_k}.$$

Heine's sum for ${}_2\phi_1(\begin{smallmatrix}q^m,\alpha\\ \beta\end{smallmatrix}; q^{-1}, q^{-1})$ will be needed in forward and backward form, namely

$$\sum_{j=0}^{c}\binom{a}{j}_q\binom{b}{c-j}_q q^{(a-j)(c-j)} = \binom{a+b}{c}_q, \tag{1.1}$$

$$\sum_{j=0}^{c}\binom{a}{j}_q\binom{b}{c-j}_q q^{j(b-c+j)} = \binom{a+b}{c}_q \tag{1.2}$$

(see Bailey [4, p. 67].

The $q$-binomial sum is

$${}_1\phi_0(\alpha;q^{-1},x) = \frac{(\alpha x)_\infty}{(x)_\infty} \quad \text{for } |x|<1 \tag{1.3}$$

(or if $\alpha = q^m$, $m=0, 1, 2, \cdots$, any $x$), [4, p. 66].

There is a transformation for balanced ${}_4\phi_3$-functions: for a positive integer $n$, and real numbers $\alpha, \beta, \gamma, \delta, \varepsilon, \theta$ such that $q^{n-1}\alpha\beta\theta = \delta\varepsilon\gamma$,

$$\begin{aligned}&{}_4\phi_3\left(\begin{matrix}q^n, \alpha, \beta, \theta\\ \delta, \varepsilon, \gamma,\end{matrix}; q^{-1}, q^{-1}\right)\\ &= \theta^n\frac{(\delta/\theta)_n(\varepsilon/\theta)_n}{(\delta)_n(\varepsilon)_n}\cdot{}_4\phi_3\left(\begin{matrix}q^n, \gamma/\alpha, \gamma/\beta, \theta\\ q^{n-1}\theta/\delta, q^{n-1}\theta/\varepsilon, \gamma\end{matrix}; q^{-1}, q^{-1}\right).\end{aligned} \tag{1.4}$$

Bailey's proof [4, p. 56] of the ${}_4F_3$-version can be easily adapted to (1.4): consider the coefficient of $z^n$ in

$${}_2\phi_1\left(\begin{matrix}\alpha,\beta\\ \gamma\end{matrix}; q^{-1}, z\right){}_2\phi_1\left(\begin{matrix}q^{n-1}/\delta, q^{n-1}/\varepsilon\\ q^{n-1}/\theta\end{matrix}; q^{-1}, \frac{\alpha\beta}{\gamma}z\right)$$
$$= {}_2\phi_1\left(\begin{matrix}\gamma/\alpha, \gamma/\beta\\ \gamma\end{matrix}; q^{-1}, \frac{\alpha\beta z}{\gamma}\right){}_2\phi_1\left(\begin{matrix}\delta/\theta, \varepsilon/\theta\\ q^{n-1}/\theta\end{matrix}; q^{-1}, z\right)$$

which equality results from [4, p. 68]

$$ {}_2\phi_1\left(\begin{matrix}\alpha,\beta\\ \gamma\end{matrix};\ q^{-1}, z\right)=\frac{(\alpha\beta z/\gamma)_\infty}{(z)_\infty}\,{}_2\phi_1\left(\begin{matrix}\gamma/\alpha,\gamma/\beta\\ \gamma\end{matrix};\ q^{-1}, \frac{\alpha\beta}{\gamma}z\right). $$

If $X$ is a finite set, then $L(X)$ denotes the space of complex functions on $X$ with inner product $\langle f, g\rangle := \sum_{x\in X} f(x)\overline{g(x)}$ and norm $\|f\| = \langle f, f\rangle^{1/2}$, $(f, g \in L(X))$. If a finite group $G$ acts on $X$ (there is a map $G\times X\to X$, $(g, x)\mapsto gx$ with $(g_1g_2)x = g_1(g_2x)$ and $ex = x$, $e$ being the identity in $G$), then $L(X)$ is a $G$-module; equivalently there is a unitary representation of $G$ on $L(X)$ (also called a permutation representation), the action being $\lambda(g)f(x) = f(g^{-1}x)$, $(f\in L(X),\ g\in G,\ x\in X)$, called left translation. A submodule of $L(X)$ is a linear subspace which is invariant under $\{\lambda(g)\colon g\in G\}$. For real numbers $a, b$ define $a\wedge b := \min(a, b)$, $a\vee b := \max(a, b)$.

**2. The general linear group and a $q$-difference equation.** Fix a prime power $q$ and let $GF(q)$ be the field with $q$ elements. For $N = 1, 2, \cdots$ let $GF(q)^N$ be the column vector space of dimension $N$ over $GF(q)$, and let $G = GL(N, q)$ denote the general linear group of this space, that is, the group of nonsingular $N\times N$ matrices over $GF(q)$ acting on the space of column vectors.

Let $\Omega$ denote the collection of all linear subspaces of $GF(q)^N$ and for $\xi\in\Omega$ let $\dim\xi$ denote the dimension over $GF(q)$. The group $G$ acts on $\Omega$ with orbits $\Omega_M := \{\omega\in\Omega\colon \dim\omega = M\}$, $M = 0, 1, \cdots, N$. Thus a collection $\{L(\Omega_M)\}$ of $G$-modules is obtained. A basis for $L(\Omega)$ is constructed by using the incidence structure of $\Omega$: indeed for $\omega\in\Omega$ define $\hat\omega\in L(\Omega)$ by $\hat\omega(\xi) = 1$ if $\omega\subset\xi$, else $= 0$. Then $L(\Omega) = \sum_{r=0}^{N}\oplus P_r$ (algebraic direct sum) where $P_r := \operatorname{span}\{\hat\omega\colon \dim\omega = r\}$. Observe that $\lambda(g)\hat\omega = (g\omega)\hat{}$, $(g\in G, \omega\in\Omega)$. To get irreducible submodules we define an operator $d$ on $L(\Omega)$ such that $dP_r\subset P_{r-1}$ and $d$ commutes with $\lambda$, namely $d\hat\xi = \sum\{\hat\eta\colon \eta\subset\xi,\ \dim\eta = \dim\xi - 1\}$ for $\xi\in\Omega$, extended by linearity.

Then $V_r := P_r\cap\ker d$ is irreducible for $0\leqq r\leqq N/2$ (a result of Steinberg [12], or see [7]). It is our aim to find functions in $V_r$ which are invariant under the parabolic subgroup of $GL(N, q)$ arising from the deletion of the generators $w_a$, $w_{a+b}$ from the Weyl group (recall the introduction).

Specifically, fix nonnegative integers $a, b, c$ with $a+b+c = N$ and fix $\zeta_1, \zeta_2\in\Omega$ with $\zeta_1\subset\zeta_2$, $\dim\zeta_1 = a$, $\dim\zeta_2 = a+b$, and let $H_{abc} := \{g\in G\colon g\zeta_1 = \zeta_1,\ g\zeta_2 = \zeta_2\}$. Further let $H_{a,b+c} := \{g\in G\colon g\zeta_1 = \zeta_1\}$ and $H_{a+b,c} := \{g\in G\colon g\zeta_2 = \zeta_2\}$, the maximal parabolic subgroups of $G$ containing $H_{abc}$, corresponding to $J_a$ and $J_{a+b}$ respectively.

We describe the obvious basis for the $H_{abc}$-invariant functions in $L(\Omega)$, their values, and the effect of $d$ on the basis elements.

PROPOSITION 2.1. *The $H_{abc}$-invariant functions in $P_r$, $0\leqq r\leqq N$ have the basis $\{g^r_{xy}\colon (x, y)\in D_r\}$ where $g^r_{xy} := \sum\{\hat\omega\colon \dim\omega = r,\ \dim(\omega\cap\zeta_1) = x,\ \dim(\omega\cap\zeta_2) = r - y\}$ and $D_r := \{(x, y)\colon x, y$ integers, $0\leqq x\leqq a$, $0\leqq y\leqq c$, $r - b\leqq x + y\leqq r\}$.*

PROPOSITION 2.2. *For $(x, y)\in D_r$,*

$$ g^r_{xy}(\xi) = \binom{u_1}{x}_q\binom{u_2}{r-x-y}_q\binom{u_3}{y}_q q^{(u_1-x)(r-x-y)+y(u_1+u_2-r+y)}, $$

*where $u_1 = \dim(\xi\cap\zeta_1)$, $u_2 = \dim(\xi\cap\zeta_2) - u_1$, $u_3 = \dim\xi - u_1 - u_2$, $(\xi\in\Omega)$. The value is zero unless $x\leqq u_1$, $y\leqq u_3$, $r - x - y\leqq u_2$. Further if $\dim\xi = r$ then $g^r_{xy}(\xi) = 0$ unless $x = u_1$, $y = u_3$ in which case $g^r_{xy}(\xi) = 1$.*

*Proof.* For a given $\xi$, count the number of subspaces $\omega\subset\xi$ of dimension $r$, with $\dim(\omega\cap\zeta_1) = x$, $\dim(\omega\cap\zeta_2) = r - y$ in the usual way (first form $\omega\cap\zeta_1$ in $\xi\cap\zeta$, then extend to $\omega\cap\zeta_2$, then to $\omega$, see [7, Proposition 3.1]). □

In the following, $g^r_{xy}$ with $(x, y) \notin D_r$ is defined to be the zero function.

PROPOSITION 2.3. $dg^r_{xy} = (1/(q-1))\quad ((q^{a+b-r+y+1} - q^{a-x})g^{r-1}_{xy} + (q^{a-x+1} - 1)$ $g^{r-1}_{x-1,y} + q^{a+b-r+1}\,(q^c - q^{y-1})\,g^{r-1}_{x,y-1})$.

*Proof.* To find the coefficients, take a particular $(r-1)$-dimensional $\omega$ associated with $g^{r-1}_{xy}$ and count the ways it can be enlarged to an $r$-dimensional space in $g^r_{xy}$, $g^r_{x+1,y}$ or $g^r_{x,y+1}$. Indeed adjoin a vector from $\zeta_2 \backslash (\zeta_1 + \omega)$, $\zeta_1 \backslash \omega$, or $GF(q)^N \backslash (\zeta_2 + \omega)$ with $q^{a+b} - q^{a-r-x-y-1}$, $q^a - q^x$, or $q^N - q^{a+b+y}$ possibilities and divide by the overcount (from generating the same space in different ways) $q^{r-y} - q^{r-y-1}$, $q^{x+1} - q^x$, or $q^r - q^{r-1}$ respectively. This produces the coefficients of $g^{r-1}_{xy}$ in the expressions for $dg^r_{xy}$, $dg^r_{x+1,y}$, or $dg^r_{x,y+1}$. □

As in [8] we set up a correspondence between $H_{abc}$-invariant elements of $V_r$ and a space of functions on $D_r$.

PROPOSITION 2.4. *Let $f$ be a function on $D_r$, then $\sum_{(x,y)\in D_r} f(x, y)\, g^r_{xy} \in V_r$ if and only if $f$ satisfies the q-difference equation*

$$\begin{aligned}(q^{a-x} - 1)f(x+1, y) + q^{a-x}(q^{b-r+1+x+y} - 1)f(x, y)\\ + (q^c - q^y)\, q^{a+b-r+1} f(x, y+1) = 0\end{aligned} \tag{2.1}$$

*for $0 \leqq x \leqq a$, $0 \leqq y \leqq c$, $r - b - 1 \leqq x + y \leqq r - 1$, and $f(a+1, y)$, $f(x, c+1)$, $f(x, r-b-1-x)$ taken as zero.*

Let $W_r$ be the linear space of solutions of (2.1). From the correspondence [5] of Hecke algebras of $GL(N; q)$ modulo parabolic subgroups to those of the Weyl group and from Proposition 2.3 of [8] we know that

$$\dim_R W_r = r \wedge a \wedge b \wedge c \wedge (a+b-r) \wedge (b+c-r) \wedge (a+c-r) \wedge (a+b+c-2r) + 1$$

if $r \leqq (a+b) \wedge (a+c) \wedge (b+c) \wedge (N/2)$, and $W_r = \{0\}$ otherwise.

As in [8] we look for a formula relating the values of $f \in W_r$ to its boundary values $f(j, r-j)$, $0 \vee (r-c) \leqq j \leqq r \wedge a$. The following is obtained; it can be inductively verified using the values $r - x - y = 0, 1, 2, \cdots, (r-b) \vee 0$, and by aid of the identity

$$\binom{A-1}{B}_q + q^{A-B}\binom{A-1}{B-1}_q = \binom{A}{B}_q.$$

The formula is found by changing the $S_N$-formula ((2.2) in [8]) to a $q$-type expression, multiplying by a power of $q$ (depending on $x$, $y$, $j$) found by solving a difference equation in $x$ and $y$.

PROPOSITION 2.5. *Let $f \in W_r$, then*

$$\begin{aligned}f(x, y) = \sum_{j=0\vee(r-c)}^{a\wedge r} f(j, r-j)\binom{r-x-y}{j-x}_q \frac{(q^{a-x})_{j-x}(q^{c-y})_{r-j-y}}{(q^b)_{r-x-y}}\\ \cdot (-1)^{r-x-y} q^{(j-a-1)(j-x)+b(r-j-y)-\binom{r-x-y}{2}}\end{aligned}$$

*If $b \geqq r$, the values of $f(j, r-j)$ can be chosen arbitrarily.*

The correspondence of $W_r$ to functions on $\Omega$ prompts us to look for a group-related inner product for $W_r$ which uses the values $f(x, y)$ ($f \in W_r$) directly. The inner product induced from $L(\Omega_r)$ is the desired one, since the $\{g^r_{xy}\}$ are an orthogonal basis for $L(\Omega_r)$ (see Proposition 2.2). Further $\|g^r_{xy}\|^2$ is the number of $\omega \in \Omega_r$ with $\dim(\omega \cap \zeta_1) = x$, $\dim(\omega \cap \zeta_2) = r - y$, namely

$$m^r_{xy} := \binom{a}{x}_q \binom{b}{r-x-y}_q \binom{c}{y}_q q^{(a-x)(r-x-y)+y(a+b-r+y)}. \tag{2.2}$$

Hence we define an inner product and norm for $W_r$: for $f_1, f_2 \in W_r$,

$$\langle f_1, f_2\rangle := \sum_{(x,y)\in D_r} m'_{xy} f_1(x,y)\overline{f_2(x,y)},$$
$$\|f_1\| = \langle f_1, f_1\rangle^{1/2}. \tag{2.3}$$

By the group invariance we know there must be two orthogonal bases for $W_r$ corresponding to the orthogonal splitting of $V_r$ into (inequivalent) irreducible submodules for $H_{a,b+c}$ or $H_{a+b,c}$ respectively (each such submodule contains at most one $H_{abc}$-invariant by the branching theorem, see [8, (2.3)].

From the results for $S_N$ we expect solutions to (2.1) expressed in terms of $q$-Hahn polynomials. We collect some results (from [7]) for reference, in bases $q^{-1}$ and $q$. For integers $a, b, c, m \leqq a \wedge b \wedge c \wedge (a+b-c)$, a (form of) $q$-Hahn polynomial is

$$E_m(a,b,c,x;q^{-1}) := (q^a)_m q^{\binom{m}{2}} \sum_{j=0}^{m} \binom{m}{j}_q$$
$$\cdot \frac{(q^{-b+m-1})_j}{(q^a)_j} q^{j(a+b-m+1)} (q^x)_j (q^{c-x})_{m-j} q^{(x-j)(m-j)} \tag{2.4}$$
$$= q^{\binom{m}{2}} (q^a)_m (q^c)_m {}_3\phi_2\left(\begin{matrix} q^m, q^{a+b-m+1}, q^x \\ q^a, q^c \end{matrix}; q^{-1}, q^{-1}\right);$$

orthogonality:

$$\sum_{x=0\vee(c-b)}^{a\wedge c} \binom{a}{x}_q \binom{b}{c-x}_q q^{(a-x)(c-x)}$$
$$\cdot E_m(a,b,c,x;q^{-1}) E_n(a,b,c,x;q^{-1}) \tag{2.5}$$
$$= \delta_{mn} (q^a)_m (q^b)_m (q^m)_m (q^{a+b-m+1})_m \binom{a+b-2m}{c-m}_q q^{m(a+c-1)}$$

(discovered by Andrews and Askey);

(2.6) special value: $E_m(a,b,c,c;q^{-1}) = (-1)^m q^{am} (q^b)_m (q^c)_m$;

special degree:

$$E_m(a,b,m,x;q^{-1}) = (q^a)_m (q^m)_m q^{\binom{m}{2}}$$
$$\cdot \frac{(q^{-b+m-1})_x}{(q^a)_x} q^{x(a+b-m+1)} \quad \text{for } x = 0, 1, \cdots, m; \tag{2.7}$$

difference equation:

$$(q^{a-x}-1)E_m(a,b,c+1,x+1;q^{-1}) + (q^{a+b-c}-q^{a-x})E_m(a,b,c+1,x;q^{-1})$$
$$= (q^{a+b-c}-q^m)E_m(a,b,c,x;q^{-1}). \tag{2.8}$$

The functions $E_m(a,b,c,x;q)$ are defined by replacing $q^{-1}$ by $q$. Further calculations give the following (in base $q^{-1}$ symbols for convenience):

$$E_m(a,b,c,x;q) = (q^a)_m q^{-m(a+c)+\binom{m}{2}}$$
$$\cdot \sum_{j=0}^{m} \binom{m}{j}_q \frac{(q^{-b+m-1})_j}{(q^a)_j} (q^x)_j (q^{c-x})_{m-j} q^{j(c-x-m+j)} \tag{2.9}$$
$$= (-1)^m q^{-m(a+b+c)+\binom{m}{2}} E_m(b,a,c,c-x;q^{-1});$$

orthogonality:

$$\sum_{x=0\vee(c-b)}^{a\wedge c}\binom{a}{x}_q\binom{b}{c-x}_q q^{x(b-c+x)}E_m(a,b,c,x;q)E_n(a,b,c,x;q)$$
$$=\delta_{mn}(q^a)_m(q^b)_m(q^m)_m(q^{a+b-m+1})_m$$
$$\cdot\binom{a+b-2m}{c-m}_q q^{-m(2a+b+c-m+2)}; \tag{2.10}$$

$$E_m(a,b,c,c;q)=(-1)^m q^{-m(a+b+c-m+1)}(q^b)_m(q^c)_m; \tag{2.11}$$

special degree:

$$E_m(a,b,m,x;q)=(q^a)_m(q^m)_m q^{-m(a+m)+\binom{m}{2}}\frac{(q^{-b+m-1})_x}{(q^a)_x}, \tag{2.12}$$
$$x=0,1,2,\cdots,m.$$

Keeping in mind the weight function and the $S_N$ case, we try for a solution of the form $f(x,y)=g(y)E_m(a,b,r-y,x;q^{-1})$. Using (2.8) we see that $g$ must satisfy $g(y+1)=((1-q^{m-a-b+r-1-y})/(1-q^{c-y}))g(y)$, and (2.12) shows that $g(y)=E_{r-m}(c,a+b-2m,r-m,y;q)$ works. As in [8] one can show that nontrivial solutions are possible exactly for $0\vee(r-c)\leqq m\leqq a\wedge b\wedge r\wedge(a+b-r)$. These will be denoted by $\phi_{rm}(x,y):=E_{r-m}(c,a+b-2m,r-m,y;q)E_m(a,b,r-y,x;q^{-1})$.

An aside remark: we have decided to use this mixed-base expression, although by (2.9) it would be possible to express the $E_{r-m}$ in base $q^{-1}$. However the present expression parallels the $S_N$-version as well as indicates the $(a,c)$ symmetry more neatly.

The orthogonality relations for $\{\phi_{rm}\}$ (from (2.3), (2.5), (2.10)) are

$$\langle\phi_{rm},\phi_{rn}\rangle=\delta_{mn}(q^m)_m(q^{r-m})_{r-m}(q^a)_m(q^b)_m$$
$$\cdot(q^{a+b-m+1})_m(q^c)_{r-m}(q^{a+b-2m})_{r-m}(q^{a+b+c-m-r+1})_{r-m}$$
$$\cdot q^{m(2a+b+2c+3r-2m+1)-r(a+b+2c+2)}. \tag{2.13}$$

Observe that the difference equation (2.1) and the weight $m^r_{xy}$ in (2.2) are essentially invariant (multiplied by scalars) under the interchange of $a$ with $c$, $x$ with $y$, and $q$ with $q^{-1}$. By applying this symmetry to $\phi_{rm}$ we obtain the other orthogonal family of solutions of (2.1),

$$\psi_{rk}(x,y):=E_{r-k}(a,b+c-2k,r-k,x;q^{-1})E_k(c,b,r-x,y;q),$$
$$\text{with } 0\vee(r-a)\leqq k\leqq b\wedge c\wedge r\wedge(b+c-r).$$

The orthogonality formula for these is

$$\langle\psi_{rk},\psi_{rn}\rangle=\delta_{kn}(q^k)_k(q^{r-k})_{r-k}(q^b)_k(q^c)_k(q^{b+c-k+1})_k$$
$$\cdot(q^a)_{r-k}(q^{b+c-2k})_{r-k}(q^{a+b+c-k-r+1})_{r-k}$$
$$\cdot q^{-k(b+2c+3r-2k+1)+r(a+r-1)}. \tag{2.14}$$

**3. Expansions and connection coefficients.** We will use orthogonality and Proposition 2.5 to develop a manageable expression for the expansion of an arbitrary element of $W_r$ in terms of $\{\phi_{rm}\}$. Then we will apply it to $\psi_{rk}$ and get the connection coefficients in terms of balanced ${}_4\phi_3$-series.

Suppose $f \in W_r$ and it is desired to express $f = \sum_m \alpha_m \phi_{rm}$. The orthogonality relations holding along lines of constant $y$ show that there is a formula

$$\alpha_m = \sum_x \binom{a}{x}_q \binom{b}{r-x-y}_q q^{(a-x)(r-x-y)} f(x, y) E_m(a, b, r-y, x; q^{-1})$$

times an expression in $m$, $y$, $a$, $b$, $c$, $r$ (requiring $y \leqq r-m$). This is a double sum (one more for $E_m$) but can be reduced to a single sum.

THEOREM 3.1. *Let $f \in W_r$, then $f = \sum_m \alpha_m \phi_{rm}$ with*

$$\alpha_m = (-1)^r \frac{q^{r(a+b+c+1)-m(a+c+2r-m)}}{(q^c)_{r-m}(q^m)_m(q^b)_m(q^{a+b-m+1})_m(q^{a+b-2m})_{r-m}(q^{r-m})_{r-m}}$$
$$\sum_{j=0\vee(r-c)}^{a\wedge r} f(j, r-j) \frac{(q^c)_{r-j}(q^m)_j(q^{a+b-m+1})_j}{(q^{-1})_j} (-1)^j q^{-j(a+b+1)+\binom{j}{2}}$$

*for $0 \vee (r-c) \leqq m \leqq a \wedge b \wedge r \wedge (a+b-r)$.*

*Proof.* This is a straightforward $q$-adaptation of the proof of Theorem 4.1 in [8]. In this case we arrive at the sum

$$\sum_x (-1)^x \frac{E_m(a, b, r-y, x; q^{-1})}{(q^x)_x (q^{j-x})_{j-x}} \cdot q^{-x(j-1)+\binom{x}{2}}.$$

The finite $q$-binomial sum (1.3)

$$\sum_{n=0}^{N} \frac{(q^N)_n}{(q^{-1})_n} x^n = (q^N x)_N$$

leads to the identity

$$\sum_{x=0}^{j} (-1)^x \frac{q^{-x(j-1)+\binom{x}{2}}}{(q^x)_x (q^{j-x})_{j-x}} \sum_{i=0}^{M} d_i (q^x)_i = (-1)^j q^{-\binom{j}{2}} d_j$$

(arbitrary constants $d_i$, positive integer $M$), a finite $q$-difference formula for polynomials in $q^x$. This applies to $E_m$ in its ${}_3\phi_2$ form (see (2.4)). □

We apply this formula to $f = \psi_{rk}$.

THEOREM 3.2. *For $0 \vee (r-a) \leqq k \leqq b \wedge c \wedge r \wedge (b+c-r)$, $\psi_{rk} = \sum_m \alpha_{km} \phi_{rm}$, where*

$$\alpha_{km} = (-1)^{r+k+m} \frac{q^A (q^a)_{r-k}(q^r)_m(q^b)_k(q^{a+b+c-r+1})_m}{(q^m)_m(q^{a+b-m+1})_m(q^{a+b-2m})_{r-m}(q^a)_m}$$
$$\cdot {}_4\phi_3\left(\begin{matrix} q^m, q^{a+b-m+1}, q^k, q^{b+c-k+1} \\ q^r, q^b, q^{a+b+c-r+1} \end{matrix}; q^{-1}, q^{-1}\right)$$

*and $A = (r-k-m)(a+b+c+2r+\frac{1}{2}) + \frac{3}{2}(m^2+k^2-r^2) + ak$.*

*Proof.* We apply Theorem 3.1 to $f = \psi_{rk}$, using the values

$$\psi_{rk}(j, r-j) = (q^a)_{r-k}(q^{r-k})_{r-k}(q^{-b-c+r+k-1})_j((q^a)_j)^{-1}(q^b)_k$$
$$\cdot (q^{r-j})_k (-1)^k q^{j(a+b+c-r+1)+\binom{k}{2}-k(b+c+2r-k)+\binom{r}{2}}$$

(from (2.7), (2.11)). This results in $\alpha_{km} = S$ times an expression not involving $j$, where

$$S = \sum_{j=0\vee(r-c)}^{m\wedge(r-k)} \frac{(q^c)_{r-j}(q^m)_j(q^{a+b-m+1})_j(q^{-b-c+r+k-1})_j(q^{r-j})_k}{(q^{-1})_j(q^a)_j} (-1)^j q^{j(c-r)+\binom{j}{2}}.$$

Note that $(q^{r-j})_k = (q^r)_k (q^{r-k})_j/(q^r)_j$. Further replace $(q^c)_{r-j}$ by $q^{-j(c-r+1)-\binom{j}{2}}(-1)^j (q^c)_r/(q^{-c+r-1})_j$, which is valid if $c \geqq r$; but if $c < r$, consider instead $c+\varepsilon$ with $0 < \varepsilon < \frac{1}{2}$ and take the limit of the expression as $\varepsilon \to 0$. Then

$$S = (q^c)_r (q^r)_k \, {}_4\phi_3\left(\begin{matrix} q^m, q^{a+b-m+1}, q^{r-k}, q^{-b-c+r+k-1} \\ q^r, q^a, q^{-c+r-1} \end{matrix}; \; q^{-1}, q^{-1}\right)$$

which is balanced. The problem with $c < r$, and the lack of symmetry in $(a, c)$, $(m, k)$ can be overcome with the transformation (1.4). Put $n = m$, $\gamma = q^r$, $\theta = q^{a+b-m+1}$ to obtain

$$S = (-1)^m q^{-mb+\binom{m}{2}} \frac{(q^c)_{r-m}(q^b)_m (q^{a+b+c-r+1})_m (q^r)_k}{(q^a)_m}$$

$$\cdot \, {}_4\phi_3\left(\begin{matrix} q^m, q^{a+b-m+1}, q^k, q^{b+c-k+1} \\ q^r, q^b, q^{a+b+c-r+1} \end{matrix}; \; q^{-1}, q^{-1}\right).$$

The restriction on $c$ can now be ignored. □

We have thus found an application for the balanced ${}_4\phi_3$-polynomials of Askey and Wilson [3]. Note that the ${}_4\phi_3$ in Theorem 3.2 can be considered as a polynomial of degree $k$ in $q^m + q^{a+b-m+1}$, or as a polynomial of degree $m$ in $q^k + q^{b+c-k+1}$.

COROLLARY 3.3. *The values of $\alpha_{km}$ can be found in closed form (no sum) for $k = 0$ if $a \geqq r$, $k = r - a$ if $r > a$, $k = r$ if $b \geqq r$, $k = b$ if $b < r$. The values of the ${}_4\phi_3$ in Theorem* 3.2 *are* 1 *if $k = 0$;*

$$q^{m(r-a)}(q^a)_m (q^{a+b-r})_m/((q^r)_m (q^b)_m) \quad \text{if } k = r-a,$$

$$q^{m(c+1)}(q^{a+b-r})_m (q^{r-c-1})_m/((q^r)_m (q^{a+b+c-r+1})_m) \quad \text{if } k = b,$$

$$q^{m(b+c-r+1)}(q^a)_m (q^{-c+r-1})_m/((q^b)_m (q^{a+b+c-r+1})_m) \quad \text{if } k = r.$$

*Proof.* In each of the latter three cases the ${}_4\phi_3$ reduces to a balanced ${}_3\phi_2$ which is done by Jackson's sum ([4, p. 68]; put $\alpha = \gamma$ in (1.4)). □

COROLLARY 3.4. *For $0 \vee (r-c) \leqq m \leqq a \wedge b \wedge r \wedge (a+b-r)$, $\phi_{rm} = \sum_k \beta_{mk} \psi_{rk}$, where*

$$\beta_{mk} = (-1)^{r+m+k} \frac{q^B (q^r)_k (q^b)_m (q^c)_{r-m} (q^{a+b+c-r+1})_k}{(q^k)_k (q^c)_k (q^{b+c-k+1})_k (q^{b+c-2k})_{r-k}}$$

$$\cdot \, {}_4\phi_3\left(\begin{matrix} q^m, q^{a+b-m+1}, q^k, q^{b+c-k+1} \\ q^r, q^b, q^{a+b+c-r+1} \end{matrix}; \; q^{-1}, q^{-1}\right)$$

*and $B = (k+m-r)(a+c+r+\frac{1}{2}) + \frac{1}{2}(r^2 - m^2 - k^2) - ak$.*

*Proof.* Clearly $\beta_{mk}\|\psi_{rk}\|^2 = \langle \phi_{rm}, \psi_{rk} \rangle = \alpha_{km}\|\phi_{rm}\|^2$. Use the norms from (2.13) and (2.14). □

The orthogonality of the ${}_4\phi_3$-polynomials is implied by their being connection coefficients for two different orthogonal bases for the same space.

PROPOSITION 3.5. $\sum_m \|\phi_{rm}\|^2 \alpha_{km}\alpha_{lm} = \delta_{kl}\|\psi_{rk}\|^2$ *with* $0 \vee (r-c) \leqq m \leqq a \wedge b \wedge r \wedge (a+b-r)$, *and* $0 \vee (r-a) \leqq k, l \leqq b \wedge c \wedge r \wedge (b+c-r)$.

*Proof.* Both sides have the value $\langle \psi_{rk}, \psi_{rl} \rangle$. □

**4. General $q$-Hahn polynomials in two variables.** Recall from § 2 that $W_r$ corresponds to a space of functions on $\Omega$, under the map $f \mapsto \sum_{(x,y)\in D_r} f(x, y) g^r_{xy}$, ($f \in W_r$). The values of $g^r_{xy}(\xi)$, $\xi \in \Omega$, depend on the integers $u_1$, $u_2$, $u_3$ where $u_1 = \dim(\xi \cap \zeta_1)$, $u_2 = \dim(\xi \cap \zeta_2) - u_1$, $u_3 = \dim \xi - u_1 - u_2$ (see Proposition 2.2).

THEOREM 4.1. *The functions $\phi_{rm}$, $\psi_{rk}$ in $W_r$ correspond to the functions*

$$\hat{\phi}_{rm}(u_1, u_2, u_3) := q^{(r-m)(u_1+u_2+u_3-r)} E_{r-m}(c, a+b-2m, u_1+u_2+u_3-m, u_3; q) \cdot E_m(a, b, u_1+u_2, u_1; q^{-1}),$$

$$\hat{\psi}_{rk}(u_1, u_2, u_3) := q^{k(u_1+u_2+u_3-r)} E_{r-k}(a, b+c-2k, u_1+u_2+u_3-k, u_1; q^{-1}) \cdot E_k(c, b, u_2+u_3, u_3; q),$$

*respectively.*

*Proof.* From Proposition 2.2 we see that

$$\hat{\phi}_{rm}(u_1, u_2, u_3) = \sum_{(x,y)\in D_r} \phi_{rm}(x,y) \cdot \binom{u_1}{x}_q \binom{u_2}{r-x-y}_q \binom{u_3}{y}_q q^{(u_1-x)(r-x-y)+y(u_1+u_2-r+y)}.$$

This sum can be done by the following projection formulas for the $E_m$-functions, which are the $q$-analogues of Gasper's formula (2.5) in [9]:

$$\sum_x \binom{d-u}{c-x}_q \binom{u}{x}_q q^{(c-x)(u-x)} E_m(a, b, c, x; q^{-1}) = \binom{d-m}{c-m}_q E_m(a, b, d, u; q^{-1}) \tag{4.1}$$

(proved by using (2.4), then

$$(q^x)_j (q^{c-x})_{m-j} q^{(x-j)(m-j)} \binom{d-u}{c-x}_q \binom{u}{x}_q q^{(c-x)(u-x)}$$

$$= (q^u)_j (q^{d-u})_{m-j} q^{(m-j)(u-j)} \binom{u-j}{x-j}_q \binom{d-u-m+j}{c-m-x+j}_q q^{(u-x)(c-x-m+j)},$$

which can be summed over by $x$ by Heine's sum (1.1));

$$\sum_x \binom{d-u}{c-x}_q \binom{u}{x}_q q^{x(d-u-c+x)} E_m(a, b, c, x; q) = q^{m(d-c)} \binom{d-m}{c-m}_q E_m(a, b, d, u; q) \tag{4.2}$$

(proved by using (2.9) and proceeding similarly, summing over $x$ by (1.2)).

Similarly $\hat{\psi}_{rk}(u_1, u_2, u_3)$ can be evaluated, summing over $y$ first, using (4.2), then over $x$ using (4.1). □

COROLLARY 4.2. *$\hat{\psi}_{rk} = \sum_m \alpha_{km} \hat{\phi}_{rm}$ and $\hat{\phi}_{rm} = \sum_k \beta_{mk} \hat{\psi}_{rk}$.*

*Proof.* These relations carry over linearly from Theorem 3.2 and Corollary 2.4. □

The functions $\hat{\phi}_{rm}$ and $\hat{\psi}_{rk}$ can also be obtained as special cases of results in [7, Thms. 4.19 and 5.6]. These express intertwining functions for a larger class of subgroups and representations than studied here. The advantage of the present development is that it is selfcontained, and less technical and involved than that of [7].

Each $\Omega_M$ is a homogeneous space for $GL(N, q)$, corresponding to the subgroup $BW_{J_M}B$ (see the Introduction). The restriction of $V_r$ to $\Omega_M$ is isomorphic to $V_r$ if $r \leqq M \wedge (N-M)$, else zero (Schur's lemma). We have produced two orthogonal bases for the $H_{abc}$-invariant functions in $L(\Omega_M)$, namely the sets $\{\hat{\phi}_{rm}: 0 \leqq r \leqq M \wedge (N-M), 0 \vee (r-a) \leqq m \leqq b \wedge c \wedge r \wedge (b+c-r)\}$ and $\{\hat{\psi}_{rk}: 0 \leqq r \leqq M \wedge (N-M), 0 \vee (r-c) \leqq k \leqq a \wedge b \wedge r \wedge (a+b-r)\}$. The orthogonality for different values of $r$ is a result of the fact that $V_r$ is not equivalent to $V_s$ if $r \neq s$.

To get actual orthogonal polynomials in two variables, we set $u_2 = M - u_1 - u_3$ then $\hat{\phi}_{rm}$ is a polynomial in $q^{u_1}$ and $q^{-u_3}$ of total degree $r$, and degree $m$ in $q^{u_1}$ (the result of the Gram–Schmidt process applied to $1, q^{-u_3}, q^{u_1}, q^{-2u_3}, q^{u_1-u_3}, q^{2u_1}, \cdots$). On the other hand, $\hat{\psi}_{rk}$ is a polynomial in $q^{u_1}$ and $q^{-u_3}$ of total degree $r$, and degree $k$ in $q^{-u_3}$; the orthogonalization of $1, q^{u_1}, q^{-u_3}, q^{2u_1}, q^{u_1-u_3}, q^{-2u_3}, \cdots$, yields $\hat{\psi}_{00}, \hat{\psi}_{10}, \hat{\psi}_{11}, \hat{\psi}_{20}, \hat{\psi}_{21}, \hat{\psi}_{22}, \cdots$.

The weights for $L(\Omega_M)$ in terms of the coordinates $(u_1, u_3)$ are given by

$$\binom{a}{u_1}_q \binom{b}{M-u_1-u_3}_q \binom{c}{u_3}_q q^{(a-u_1)(M-u_1-u_3)+u_3(a+b-M+u_3)}, \tag{4.3}$$

the number of $M$-dimensional subspaces $\omega$ satisfying $\dim(\omega \cap \zeta_1) = u_1$, $\dim(\omega \cap \zeta_2) = M - u_3$.

To extend the validity of the orthogonality and connection results, we note that if $a$, $b$, $c \geqq M$ then the indices $r$, $m$, $k$ are constrained only by $0 \leqq m$, $k \leqq r \leqq M$. All expressions dealt with are rational in $q^a, q^b, q^c$ with singularities at most at $1, q, q^2, \cdots, q^{M-1}$. Thus the formulas in Corollary 4.2 are valid for any $q^a, q^b, q^c$ with none in the set $\{1, q, q^2, \cdots, q^{M-1}\}$, the summations extending over $0 \leqq m \leqq r$, or $0 \leqq k \leqq r$ as appropriate. Also $q$ can be any real number greater than 1.

**5. $q$-Jacobi polynomials in two variables.** In this section we find two families of $q$-polynomials in two variables, both orthogonal with respect to a measure supported by a countably infinite compact set. The transformation from one family to another involves the ${}_4\phi_3$-polynomials.

Ordinary Jacobi polynomials can be obtained as limits of Hahn polynomials. For the $q$-version, fix $a$, $b$, $m$, $x$, $q > 1$ and observe

$$\lim_{M\to\infty} q^{-mM} E_m(a, b, M, M-x; q^{-1}) = (-1)^m (q^a)_m {}_2\phi_1\left(\begin{matrix} q^m, q^{a+b-m+1} \\ q^a \end{matrix}; \; q^{-1}, q^{-x-1}\right),$$

a polynomial of degree $m$ in $q^{-x}$. Andrews and Askey [1] called these little $q$-Jacobi polynomials and determined their orthogonality relations.

DEFINITION 5.1. For parameters $\alpha$ $(\neq q, q^2, q^3, \cdots)$, $\beta$, $m = 0, 1, 2, \cdots$ the little *$q$-Jacobi polynomial* of degree $m$ in $q^{-x}$, base $q^{-1}$ is

$$p_m(q^{-x}; \alpha, \beta | q^{-1}) := {}_2\phi_1\left(\begin{matrix} q^m, \alpha\beta q^{-m-1} \\ \alpha q^{-1} \end{matrix}; \; q^{-1}, q^{-x-1}\right).$$

These are orthogonal with respect to the weight $((\beta q^{-1})_x/(q^{-1})_x)\,(\alpha q^{-1})^x$ at $x = 0, 1, 2, \cdots$, which is positive and summable if either $0 < \alpha < q$, $\beta < q$ (infinite support), or $\beta = q^{N+1}$ some $N = 0, 1, 2, \cdots$ and $\alpha < 0$ (finite support, the degree $m \leqq N$, these are called $q$-Krawtchouk polynomials by Stanton [11], and $p_m(q^{-x}; \alpha, q^{N+1} | q^{-1}) = ((q^N)_m/(\alpha q^{-m})_m) K_m(q^{-x}; -1/(\alpha q^{N+1}), N; q)$ in his notation).

The orthogonality relation [1, (3.8)] is

$$\sum_{x=0}^{\infty} \frac{(\beta q^{-1})_x}{(q^{-1})_x} \alpha^x q^{-x} p_n(q^{-x}; \alpha, \beta | q^{-1}) p_m(q^{-x}; \alpha, \beta | q^{-1})$$
$$= \delta_{mn} \frac{(\alpha\beta q^{-2})_\infty}{(\alpha q^{-1})_\infty} \frac{(q^{-1})_m (\beta q^{-1})_m (1 - \alpha\beta q^{-m-1})}{(\alpha q^{-1})_m (\alpha\beta q^{-2})_m (1 - \alpha\beta q^{-2m-1})} \alpha^m q^{-m}. \tag{5.1}$$

Also

$$p_m(1; \alpha, \beta | q^{-1}) = (-1)^m \alpha^m q^{-m-\binom{m}{2}} (\beta q^{-1})_m / (\alpha q^{-1})_m. \tag{5.2}$$

We restate the above limit

$$\lim_{M\to\infty} q^{-mM}E_m(a, b, M, M-x; q^{-1}) = (-1)^m(q^a)_m p_m(q^{-x}; q^{a+1}, q^{b+1}|q^{-1}). \tag{5.3}$$

The corresponding limit for $E_m(a, b, c, x; q)$ is

$$\lim_{M\to\infty} E_m(a, b, M, x; q) = q^{-m(a+b)+\binom{m}{2}}(q^b)_m p_m(q^{-x}; q^{b+1}, q^{a+1}|q^{-1}), \tag{5.4}$$

(convert $E_m$ to a multiple of $E_m(b, a, M, M-x; q^{-1})$ by (2.9), then take limit).

The formula (5.1) can be obtained by taking the limit of the $E_m$ orthogonality (2.5) with the dominated convergence theorem for sums.

We now apply the same limiting process in the polynomials $\hat{\phi}_{rm}$ and $\hat{\psi}_{rk}$ from § 4. To do this we require $q^a$, $q^b$, $q^c$ to avoid the values 1, $q$, $q^2, \cdots$; let $u_1 = M-x$ and $u_3 = y$ (the constraints $u_1$, $u_3 \geqq 0$ and $u_1+u_3 \leqq M$ become $0 \leqq y \leqq x \leqq M$). Also we will switch to the $Q_k$ notation for $q$-Hahn polynomials to avoid base-$q$ logarithms: indeed

$$Q_k(q^x; \alpha, \beta, N|q^{-1}) := {}_3\phi_2\left(\begin{matrix} q^k, \alpha\beta q^{-k-1}, q^x \\ \alpha q^{-1}, q^N \end{matrix}; q^{-1}, q^{-1}\right) \tag{5.5}$$

related to $E_k$ by $E_k(a, b, c, x; q^{-1}) = q^{\binom{k}{2}}(q^a)_k(q^c)_k\, Q_k(q^x; q^{a+1}, q^{b+1}, c|q^{-1})$.

PROPOSITION 5.2. *For* $0 \leqq m, k \leqq r$

$$\lim_{M\to\infty} q^{-rM}\hat{\phi}_{rm}(M-x, x-y, y) = (-1)^m(q^a)_m(q^{a+b-2m})_{r-m}$$
$$\cdot q^{-r(a+b+c+r)+\binom{r}{2}+m(a+b+c-2m+2r+1)+\binom{m}{2}}$$
$$\cdot \Phi_{rm}(x, y; q^{a+1}, q^{b+1}, q^{c+1}|q^{-1})$$

*and*

$$\lim_{M\to\infty} q^{-rM}\hat{\psi}_{rk}(M-x, x-y, y) = (-1)^r(q^a)_{r-k}(q^b)_k$$
$$\cdot q^{-k(b+c+2r-2k+1)}\Psi_{rk}(x, y; q^{a+1}, q^{b+1}, q^{c+1}|q^{-1}),$$

*where*

$$\Phi_{rm}(x, y; \alpha, \beta, \gamma|q^{-1}) := p_{r-m}(q^{-y}; \alpha\beta q^{-2m-1}, \gamma|q^{-1})$$
$$\cdot p_m(q^{y-x}; \alpha, \beta|q^{-1})q^{-my}$$

*and*

$$\Psi_{rk}(x, y; \alpha, \beta, \gamma|q^{-1}) := p_{r-k}(q^{k-x}; \alpha, \beta\gamma q^{-2k-1}|q^{-1})$$
$$\cdot q^{-kx}(q^x)_k Q_k(q^{x-y}; \beta, \gamma, x|q^{-1}).$$

*Proof.* This is a straightforward calculation using (5.3), (5.4) and (5.5). □

The functions $\Phi_{rm}$, $\Psi_{rk}$ are defined if $\alpha$ and $\beta$ avoid the values $q$, $q^2, \cdots, q^r$. Observe that both $\Phi_{rm}$, $\Psi_{rk}$ are polynomials in $q^{-x}$, $q^{-y}$ of total degree $r$, and $\Phi_{rm}$ is of degree $m$ in $q^{-x}$ while $\Psi_{rk}$ is of degree $k$ in $q^{-y}$.

We will find the weight function for orthogonality by taking the limit of (4.3) divided by $\begin{pmatrix} a+b+c \\ M \end{pmatrix}_q$, for which $\{\hat{\phi}_{rm}\}$ and $\{\hat{\psi}_{rk}\}$ are orthogonal sets:

for $q>1$, $0 \leqq y \leqq x$,

$$\lim_{M\to\infty} \begin{pmatrix} a \\ M-x \end{pmatrix}_q \begin{pmatrix} b \\ x-y \end{pmatrix}_q \begin{pmatrix} c \\ y \end{pmatrix}_q q^{(a-M+x)(x-y)+y(a+b-M+y)} \Big/ \begin{pmatrix} a+b+c \\ M \end{pmatrix}_q \tag{5.6}$$

$$= \frac{(q^b)_{x-y}(q^c)_y}{(q^{-1})_{x-y}(q^{-1})_y} q^{ax+by} \frac{(q^a)_\infty}{(q^{a+b+c})_\infty}.$$

These values sum to one if $b$ and $c$ are nonnegative integers, or if $|q^a|<1$ and $|q^{a+b}|<1$. To prove (5.6) note that

$$\begin{pmatrix} a \\ M-x \end{pmatrix}_q \Big/ \begin{pmatrix} a+b+c \\ M \end{pmatrix}_q = (-1)^x q^{Mx-\binom{x}{2}} (q^{-M+x-1})_x (q^a)_{M-x} / (q^{a+b+c})_M.$$

DEFINITION 5.3. For numbers $\alpha$, $\beta$, $\gamma$ and integers $x$, $y$ with $0 \leqq y \leqq x$ let

$$w_{\alpha\beta\gamma}(x, y) := \frac{(\beta q^{-1})_{x-y}(\gamma q^{-1})_y}{(q^{-1})_{x-y}(q^{-1})_y} \alpha^x \beta^y q^{-xy}.$$

This is a *positive weight function* if 1) $0<\alpha<q$, $0<\beta<q$, $\gamma<q$; or 2) $0<\alpha<q$, $-q<\beta<0$, $\gamma=q^{c+1}$; or 3) $\alpha<0$, $\beta=q^{b+1}$, $\gamma=q^{c+1}$; for nonnegative integers $b$, $c$.

We now have two families of orthogonal polynomials for this weight function. The orthogonality relations are calculated from (5.1) and (2.5).

THEOREM 5.4. *The family* $\{\Phi_{rm}(x, y; \alpha, \beta, \gamma|q^{-1})\}$ *is a complete set of orthogonal polynomials for the weight* $w_{\alpha\beta\gamma}$, *resulting from the Gram–Schmidt process applied to* $1, q^{-y}, q^{-x}, q^{-2y}, q^{-x-y}, q^{-2x}, \cdots$, *and satisfies the relation*

$$\sum_{0\leqq y\leqq x}\sum w_{\alpha\beta\gamma}(x, y)\Phi_{rm}(x, y; \alpha, \beta, \gamma|q^{-1})\Phi_{sn}(x, y; \alpha, \beta, \gamma|q^{-1})$$

$$= \delta_{rs}\delta_{mn}(-1)^r q^{-3r-\binom{r}{2}+m(m+1-r)} \frac{(\alpha\beta\gamma q^{-r-3})_\infty(1-\alpha\beta\gamma q^{-r-2})}{(\alpha q^{-1})_\infty(1-\alpha\beta\gamma q^{-2r-2})}$$

$$\cdot \alpha^r \beta^{r-m} \frac{(q^m)_m(q^{r-m})_{r-m}(\beta q^{-1})_m(\gamma q^{-1})_{r-m}(\alpha\beta q^{-m-1})_m}{(\alpha q^{-1})_m(\alpha\beta\gamma q^{-r-2})_m(\alpha\beta q^{-2m-2})_{r-m}}.$$

*For* $w_{\alpha\beta\gamma}$ *in case* 1), *the family is infinite and* $0 \leqq m \leqq r$; *for case* 2), *the family is infinite and* $0 \vee (r-c) \leqq m \leqq r$; *for case* 3), *the family is finite and* $0 \vee (r-c) \leqq m \leqq b \wedge r$.

THEOREM 5.5. *The family* $\{\Psi_{rk}(x, y; \alpha, \beta, \gamma|q^{-1})\}$ *is a complete set of orthogonal polynomials for the weight* $w_{\alpha\beta\gamma}$, *resulting from the orthogonalization of* $1, q^{-y}, q^{-x}, q^{-2x}, q^{-x-y}, q^{-2y}, \cdots$, *and satisfies*

$$\sum_{0\leqq y\leqq x}\sum w_{\alpha\beta\gamma}(x, y)\Psi_{rk}(x, y; \alpha, \beta, \gamma|q^{-1})\Psi_{sn}(x, y; \alpha, \beta, \gamma|q^{-1})$$

$$= \delta_{rs}\delta_{kn}(-1)^r q^{-2r-\binom{r}{2}+k(r-2k)} \frac{(\alpha\beta\gamma q^{-r-3})_\infty(1-\alpha\beta\gamma q^{-r-2})}{(\alpha q^{-1})_\infty(1-\alpha\beta\gamma q^{-2r-2})}$$

$$\cdot \alpha^r \beta^k \frac{(q^k)_k(q^{r-k})_{r-k}(\gamma q^{-1})_k(\beta\gamma q^{-k-1})_k(\beta\gamma q^{-2k-2})_{r-k}}{(\beta q^{-1})_k(\alpha q^{-1})_{r-k}(\alpha\beta\gamma q^{-r-2})_k}.$$

*For $w_{\alpha\beta\gamma}$ in case* 1), *the family is infinite and* $0 \leqq k \leqq r$; *for case* 2), *the family is infinite and* $0 \leqq k \leqq c \wedge r$; *for case* 3), *the family is finite and* $0 \leqq k \leqq b \wedge c \wedge r \wedge (b+c-r)$.

The connection coefficients for $\{\hat{\phi}_{rm}\}$ and $\{\hat{\psi}_{rk}\}$ in Corollary 4.2 can be transferred to the sets $\{\Phi_{rm}\}$ and $\{\Psi_{rk}\}$. In Corollary 4.2 multiply both sides by $q^{-rM}$ and take the limit as $M \to \infty$ using Proposition 5.2; then replace $q^a$, $q^b$, $q^c$ by $\alpha q^{-1}$, $\beta q^{-1}$, $\gamma q^{-1}$ respectively, obtaining

$$\Psi_{rk}(x, y; \alpha, \beta, \gamma | q^{-1}) = \sum_{m=0}^{r} A_{km} C_{km} \Phi_{rm}(x, y; \alpha, \beta, \gamma | q^{-1}), \tag{5.7}$$

$$\Phi_{rm}(x, y; \alpha, \beta, \gamma | q^{-1}) = \sum_{k=0}^{r} B_{mk} C_{km} \Psi_{rk}(x, y; \alpha, \beta, \gamma | q^{-1}), \tag{5.8}$$

where

$$A_{km} = (-1)^k q^{-\binom{k}{2}} \frac{(q^r)_m (\alpha\beta\gamma q^{-r-2})_m}{(q^m)_m (\alpha\beta q^{-m-1})_m},$$

$$B_{mk} = (-1)^k \beta^{r-m-k} q^{-r-m(r-m-1)-k(r-k-1)+\binom{k}{2}} \cdot \frac{(q^r)_k (\beta q^{-1})_m (\beta q^{-1})_k (\alpha q^{-1})_{r-k} (\gamma q^{-1})_{r-m} (\alpha\beta\gamma q^{-r-2})_k}{(q^k)_k (\alpha q^{-1})_m (\gamma q^{-1})_k (\beta\gamma q^{-k-1})_k (\beta\gamma q^{-2k-2})_{r-k} (\alpha\beta q^{-2m-2})_{r-m}}$$

and

$$C_{km} = {}_4\phi_3\left(\begin{matrix} q^m, \alpha\beta q^{-m-1}, q^k, \beta\gamma q^{-k-1} \\ q^r, \beta q^{-1}, \alpha\beta\gamma q^{-r-2} \end{matrix}; q^{-1}, q^{-1}\right).$$

Some transformations for $q$-Jacobi polynomials may be derived from (5.7) and (5.8). We give a few examples:

1) Multiply both sides of (5.7) by $(\beta q^{-1})_{x-y} (\alpha q^{-1})^{x-y} / (q^{-1})_{x-y}$ and sum over $x = y, y+1, y+2, \cdots$ obtaining $A_{k0}\Phi_{r0}\, (\alpha\beta q^{-2})_\infty / (\alpha q^{-1})_\infty$ on the right side. The case $k = 0$ yields

$$p_r(q^{-y}; \alpha\beta q^{-1}, \gamma | q^{-1}) = \frac{(\alpha q^{-1})_\infty}{(\alpha\beta q^{-2})_\infty} \sum_{x=y}^{\infty} \frac{(\beta q^{-1})_{x-y}}{(q^{-1})_{x-y}} (\alpha q^{-1})^{x-y} p_r(q^{-x}; \alpha, \beta\gamma q^{-1} | q^{-1}), \tag{5.9}$$

$$r = 0, 1, 2, \cdots.$$

2) Multiply both sides of (5.8) by $((\beta q^{-1})_{x-y} (\gamma q^{-1})_y / (q^{-1})_{x-y} (q^{-1})_y)\, (\beta q^{-1})^{y-x}$ and sum over $y = 0, 1, \cdots, x$; obtaining $B_{m0}\Psi_{r0}(\gamma\beta q^{-2})_x (\beta q^{-1})^x / (q^{-1})_x$ on the right side. The resulting expressions look better if we use the values $p_r(1; \alpha, \beta | q^{-1})$ from (5.2), indeed:

$$\frac{p_r(q^{-x}; \alpha, \beta\gamma q^{-1} | q^{-1})}{p_r(1; \alpha, \beta\gamma q^{-1} | q^{-1})} = \frac{(q^{-1})_x (\beta q^{-1})^x}{(\gamma\beta q^{-2})_x} \sum_{y=0}^{x} \frac{(\beta q^{-1})_{x-y} (\gamma q^{-1})_y}{(q^{-1})_{x-y} (q^{-1})_y} \cdot (\beta q^{-1})^{y-x} \frac{p_r(q^{-y}; \alpha\beta q^{-1}, \gamma | q^{-1})}{p_r(1; \alpha\beta q^{-1}, \gamma | q^{-1})} \tag{5.10}$$

for $m = 0$;

$$\frac{p_r(q^{-x};\alpha,\beta\gamma q^{-1}|q^{-1})}{p_r(1;\alpha,\beta\gamma q^{-1}|q^{-1})} = \frac{(q^{-1})_x(\beta q^{-1})^x}{(\gamma\beta q^{-2})_x}\sum_{y=0}^{x}\frac{(\beta q^{-1})_{x-y}(\gamma q^{-1})_y}{(q^{-1})_{x-y}(q^{-1})_y}\cdot(\beta q^{-1})^{y-x}q^{-ry}\frac{p_r(q^{y-x};\alpha,\beta|q^{-1})}{p_r(1;\alpha,\beta|q^{-1})} \tag{5.11}$$

for $m = r$.

Another possibility is to put $y = 0$ in the expansion $\Psi_{rr} = \sum_m A_{rm}C_{rm}\Phi_{rm}$ ($C_{rm}$ can be summed, see Corollary 3.3), obtaining the expansion of $q^{-rx}(q^x)_r$ as a series in $\{p_m(q^{-x};\alpha,\beta|q^{-1}): 0 \leqq m \leqq r\}$.

The transformations (5.9), (5.10), (5.11) are $q$-analogues of transformations of Jacobi polynomials found by Askey and Fitch [2] (see also the survey by Gasper [10]). Of course the same trick can be applied to the $q$-Hahn polynomials in Corollary 4.2, resulting in the $q$-analogues of Gasper's projection formulas (2.2), (2.3), (2.4), [9, p. 179] (the actual calculations will be left as exercises).

The author hopes that the reader has seen how some simple ideas from group representation theory, such as restriction to subgroups, and orthogonality of inequivalent representations, and some counting of subspaces of a finite vector space can lead to far-reaching results on $q$-polynomials in two variables. These explain, motivate, and suggest proofs of various transformation formulas of $q$-Hahn and $q$-Jacobi polynomials, and the orthogonality of balanced ${}_4\phi_3$-polynomials.

## REFERENCES


[1] G. ANDREWS AND R. ASKEY, *Enumeration of partitions: The role of Eulerian series and q-orthogonal polynomials,* Higher Combinatorics, M. Aigner, ed., Reidel, Dordrecht, Holland, 1977, pp. 3–26.

[2] R. ASKEY AND J. FITCH, *Integral representations for Jacobi polynomials and some applications,* J. Math. Anal. Appl., 26 (1969), pp. 411–437.

[3] R. ASKEY AND J. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients or* 6-j *symbols,* SIAM J. Math. Anal., 10 (1979), pp. 1008–1016.

[4] W. BAILEY, *Generalized Hypergeometric Series,* Cambridge University Press, London, 1935.

[5] C. CURTIS, N. IWAHORI AND R. KILMOYER, *Hecke algebras and characters of parabolic type of finite groups with* (B, N)*-pairs,* Inst. Hautes Études Sci. Publ. Math., 40 (1971), pp. 81–116.

[6] C. DUNKL, *An addition theorem for Hahn polynomials: The spherical functions,* SIAM J. Math. Anal., 9 (1978), pp. 627–637.

[7] ———, *An addition theorem for some q-Hahn polynomials,* Monatsh. Math., 85 (1977), pp. 5–37.

[8] ———, *A difference equation and Hahn polynomials in two variables,* Pacific J. Math., to appear.

[9] G. GASPER, *Projection formulas for orthogonal polynomials of a discrete variable,* J. Math. Anal. Appl., 45 (1974), pp. 176–198.

[10] ———, *Positivity and special functions,* Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 375–433.

[11] D. STANTON, *Some q-Krawtchouk polynomials on Chevalley groups,* to appear.

[12] R. STEINBERG, *A geometric approach to the representations of the full linear group over a Galois field,* Trans. Amer. Math. Soc., 71 (1951), pp. 274–282.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0002 $01.00/0

# RANDOM COLORINGS OF A LATTICE OF SQUARES IN THE PLANE*

E. N. GILBERT†

**Abstract.** Square cells, tesselating the plane in a lattice arrangement, will be colored black or white by a random process. The coloring tries to imitate the appearance of cells with statistically independent colors, with black and white equally likely. Here only a relatively small initial set of cells is colored independently; the remaining colors are then determined by solving a linear recurrence equation. In this way one obtains colorings which, for some value of $n$, have independent colors in every set of $n$ cells. The value of $n$, which depends on the recurrence equation used, can be deduced from divisibility properties of certain polynomials.

**1. Introduction.** A *lattice of squares* is a tesselation of the plane into unit square cells. Each cell may be specified by its midpoint $P=(\xi, \eta)$ with integer Cartesian coordinates $\xi$, $\eta$. A *coloring* of a lattice of squares is obtained by painting each square cell black or white. Or, with numbers 0 and 1 representing white and black a coloring is a binary function $a(P)=a(\xi, \eta)$.

Here the colorings will be generated by random processes. The quantity of main interest is the *n-gram probability* $p(P_1, \cdots, P_n, a_1, \cdots a_n)$, the joint probability that $a(P_k)=a_k$ for $k=1$ to $n$. For instance if cells are colored independently at random with black and white equally likely, then

$$p(P_1, \cdots, P_n, a_1, \cdots, a_n)=2^{-n} \tag{1}$$

holds for every $n$ and every choice of $n$ distinct cells $P_1, \cdots, P_n$ and their colors $a_1, \cdots, a_n$. Other colorings, to be described, satisfy (1) for a fixed $n$ and all $P_1, \cdots, P_n, a_1 \cdots, a_n$. These are said to have the *n-th order randomness* property. Need for these colorings arose in testing B. Julesz's statistical theory of visual texture discrimination (see [4], [5], [6], [7]). Each test required two colorings that had the same $n$-gram probabilities for some small value of $n$, say 3 or 4, but which differed in some other way that was easily noticed by eye. Rosenblatt and Slepian [8] produced (one-dimensional) colorings of cells arranged in a straight line, again with randomness properties designed for Julesz's experiments.

The plane colorings with a $n$th order randomness property can also be regarded as two-dimensional analogs of the pseudorandom sequences that often appear in coding theory or cryptography [1], [2], [3]. In either case a small fraction of the cells is colored independently at random (perhaps to serve as a key) and then a recurrence equation extends the coloring to the rest of the cells.

One-dimensional linear pseudorandom sequences are necessarily periodic; hence they cannot have $n$th order randomness with $n \geqq 2$, although they have a large period if properly designed. The plane colorings will be aperiodic. The main problems will be to determine the highest order of randomness possessed by a given recurrence and to find recurrences that achieve $n$th order randomness without using a large number $K$ of terms (of course, $K<n$).

**2. Floaters.** The coloring function will be a solution of a linear recurrence

$$\sum_{k=1}^{K} a(\xi+\xi_k, \eta+\eta_k)=b \tag{2}$$

which holds for all integers $\xi$ and $\eta$. Equation (2) and all subsequent equations for

* Received by the editors May 30, 1979.
† Bell Laboratories, Murray Hill, New Jersey 07974.

coloring functions $a(\xi, \eta)$ are taken modulo 2. In (2), $b = 0$ or 1. If $b = 0$, the coloring is called an *even coloring*; if $b = 1$, it is *an odd coloring*. The points $F_k = (\xi_k, \eta_k)$, $k = 1, \cdots, K$ determine $K$ distinct squares, which may be interpreted as $K$ windows in a mask obscuring the plane. The terms of (2) are the values seen through the $K$ windows when the mask is translated bodily, the origin moving to $(\xi, \eta)$. Because the set of windows $\phi = \{F_1, F_2, \cdots, F_k\}$ moves to all locations of the plane to produce all the equations of the recurrence system (2), $\phi$ is called the *floater* of the coloring.

To produce a coloring, colors may first be prescribed in some set $I$ and then extended to a coloring function $a(P)$ satisfying (2) throughout the plane. There are many solutions of (2) if $I$ is too small and none if $I$ is too big. $I$ is an *initial set* if every coloring of $I$ extends to a unique coloring of the plane.

For example if the floater is a $2 \times 2$ square, $\phi = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, then the row $\eta = 0$ and column $\xi = 0$ together comprise one initial set $I$. The unique solution of (2) that extends a prescribed coloring of $I$ is

$$a(\xi, \eta) = a(\xi, 0) + a(0, \eta) + a(0, 0) + b\xi\eta.$$

This example illustrates a special kind of initial set, called a *standard column initial set*, that can be constructed for any floater $\phi$ as follows.

Suppose that the floater $\phi$ is $W$ columns wide. Suppose that the leftmost column of the floater is contained in $H_L$ consecutive rows. Likewise, let $H_R$ be the number of consecutive rows required to contain the rightmost column. For the floater in Fig. 1, $W = 4$, $H_L = 4$, $H_R = 3$. Select $W - 1$ consecutive columns of the plane and make them part of $I$. Also assign to $I$ the parts of any $H_L - 1$ consecutive rows lying to the left of the selected columns and any $H_R - 1$ consecutive rows to the right of the selected columns. In Fig. 1 only these columns and half-rows are shown. With the floater in the starting

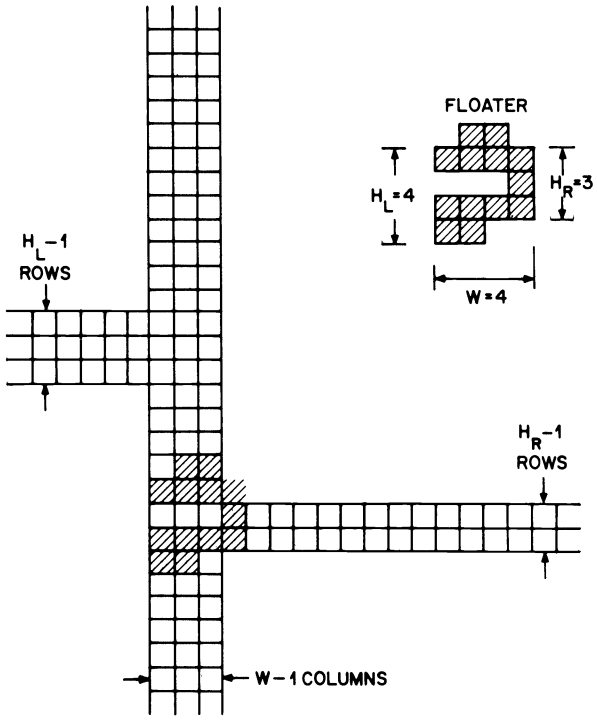



FIG. 1. *A standard column initial set.*

position drawn, only one square of the floater lies outside $I$; then (2) determines $a(P)$ for that square. By moving the floater up and down, an entire column may be colored one square at a time. The floater may then be moved sideways to color other columns.

Clearly one could also construct an initial set in a similar way starting with some complete rows and adding half-columns above and below these rows. That set will be called a *standard row initial set.*

If an initial set $I$ is given, then the colors $a(P_1), \cdots, a(P_n)$ at any $n$ given squares can be determined from the colors $c_1, \cdots, c_J$ of some finite number $J$ of squares of $I$. Because (2) is linear there will be a relationship

$$a(P_i) = \sum_{j=1}^{J} m_{ij}c_j + d_i b, \qquad i = 1, 2, \cdots, n. \tag{3}$$

with $(m_{ij})$ a binary matrix (the term $d_i b$ represents the solution with $I$ colored completely white). Thus one may regard colors $a(P_i)$ either as linear functions of colors $c_j$ on $I$ or, with $I$ colored at random, as random variables. If squares of $I$ are colored independently, with 0 and 1 equally likely, then 0 and 1 will also be equally likely at each other point $P_i$ of the plane. However, colors of squares outside $I$ need not be statistically independent.

LEMMA 1. *Let colors in an initial set $I$ be determined at random, independently, with* 0 *and* 1 *equally likely. Either the colors at n given squares are statistically independent or they satisfy a linear relationship derivable from* (2).

*Proof.* Let $P_1, \cdots, P_n$ specify the $n$ squares so that (1) is the condition that their colors are statistically independent. Let $r$ denote the rank of the matrix $(m_{ij})$.

If $r < n$, then the colors $a(P_i)$ satisfy a linear relationship and (1) fails.

If $r = n$, then the system (3) can be solved for $n$ of the initial colors $c_1, \cdots, c_J$ as linear functions of the remaining $J - n$ initial colors and the $n$ colors $a(P_i)$. Each of the $2^n$ color combinations $a_1, \cdots, a_n$ then arises in $2^{J-n}$ ways, each with probability $2^{-J}$, that correspond to color combinations of $J - n$ squares of $I$. Then (1) holds.

Section 4 will give a test for independence that is more convenient than evaluating the rank of (3).

**3. Translation.** A point $Q = (\xi_0, \eta_0)$ of the plane determines a translation that transforms each point $P = (\xi, \eta)$ to

$$P' = P + Q = (\xi + \xi_0, \eta + \eta_0). \tag{4}$$

Translation leaves the system of (2) invariant; replacing $P$ by $P + Q$ in (2) merely produces another one of the equations belonging to the same system (2). That suggests that a random coloring function $a(P)$ satisfying (2) may be *stationary*, i.e.,

$$p(P_1 + Q, \cdots, P_n + Q, a_1, \cdots, a_n) = p(P_1, \cdots, P_n, a_1, \cdots, a_n) \tag{5}$$

may hold for all $n$, all $2^n$ color combinations $a_1, \cdots, a_n$, all points $P_1, \cdots, P_n$, and all translations $Q$. However, $a(P)$ is generated from values in a fixed initial set $I$, not translation invariant, and so stationarity is only obtained under special conditions.

THEOREM 1. *A random coloring function $a(P)$ satisfying* (2) *is stationary if its values on an initial set $I$ are independent and equally likely to be* 0 *or* 1.

The proof will follow from two preliminary lemmas.

LEMMA 2. *Let $I + Q$ denote the set of points $P + Q$ obtained by translating all points $P$ of an initial set $I$. Then $I + Q$ is an initial set.*

The proof follows simply from the fact that (2) is translation invariant.

LEMMA 3. *If colors are statistically independent and* 0, 1 *equally likely on one initial set* $I_1$, *they are statistically independent and* 0, 1 *equally likely on any other initial set* $I_2$.

*Proof.* The colors of any number $k$ of squares belonging to $I_2$ cannot satisfy a linear dependence. Suppose the contrary, i.e., (2) implies such a dependence. It is possible to color $I_2$ in a way that violates the dependence. But that coloring certainly fails to extend to a solution of (2). Then $I_2$ was not an initial set, a contradiction. Since colors on $I_2$ are not linearly dependent, Lemma 1 shows that they are statistically independent.

To prove Theorem 1 consider any $n$ squares $P_1, \cdots, P_n$ and a translation vector $Q$. The result to be proved is (5). Solving (2) leads to a system (3) relating $a(P_i)$ to colors $c_j = a(T_j)$, where $T_1, \cdots, T_J$ belong to an initial set $I$. Because the system (2) is translation invariant, (3) remains true if all squares $P_i$, $T_j$ are translated. That leads to another system

$$a(P_i + Q) = \sum_{j=1}^{J} m_{ij} c'_j + d_i b, \qquad i = 1, 2, \cdots, n \tag{6}$$

with $c'_j = a(T_j + Q)$ and with the same parameters $m_{ij}$, $d_i$ as in (3). The points $T_j + Q$ belong to $I + Q$, which Lemma 2 shows to be an initial set. Then Lemma 3 shows that $c'_1, \cdots, c'_J$ are independent colors. Now the two probability distributions in (5) must indeed be equal; for $a(P_i)$ and $a(P_i + Q)$ are determined from independent colors ($c_j$ or $c'_j$) by the same equations ((3) or (6)).

Theorem 1 and Lemma 3 sometimes provide quick proofs of the $n$th order randomness property (1). For example, suppose the floater is a $2 \times 2$ square. Standard column initial sets contain one column, one half-row to the right, and another half-row to the left. Given any three points, some standard initial set includes all three points. Then Lemma 3 shows that (1) holds for these points if colors on $I$ are independent and equally likely.

All the colors in any $W - 1$ consecutive columns are independent if the floater is $W$ columns wide, for the columns in question can be made part of a standard column initial set. Likewise, any $H - 1$ consecutive rows can be made the rows of a standard row initial set if the floater is $H$ rows high. Then any $H - 1$ consecutive rows have independent colors. These $H - 1$ rows need not be independent of the $W - 1$ columns in general unless $H_L = H_R = H$.

**4. Generating functions.** An algebraic test for $n$th order randomness can be formulated in terms of generating functions. Associate to each set $\sigma$ of squares $(\xi, \eta)$, a generating function

$$G_\sigma = \sum_{\sigma} x^{\xi} y^{\eta} \tag{7}$$

in which the sum extends over all squares of $\sigma$. These expressions will be added and multiplied formally by the usual rules for power series, the coefficients being treated as integers modulo 2. If $\sigma$ contains only a finite number of squares, then $G_\sigma$ contains finitely many terms and will be called a *polynomial.* Because coordinates $\xi$ or $\eta$ may be negative, a "polynomial" can contain negative powers of $x$ or $y$.

The polynomial of main interest will be $G_\phi$, the generating function of the floater $\phi$. Also $G_\psi$ will denote the polynomial for a set $\psi = \{P_1, P_2, \cdots, P_n\}$ of squares with colors $a(P_1), \cdots, a(P_n)$ to be tested for statistical independence. Under the conditions of Lemma 1, $a(P_1), \cdots, a(P_n)$ must be tested for linear independence. Because the system (2) is invariant under translation, any translation $\phi + Q$ of the floater is another floater producing the same equations (2). Likewise a linear dependence between colors

of squares of $\psi$ implies a linear dependence between colors of squares of $\psi + Q$. Then it will suffice to consider sets $\phi$ and $\psi$ lying in the positive quadrant, in which case $G_\phi$ and $G_\psi$ contain no negative powers.

The product $G_1G_2$ of two generating functions is well determined by the usual rules for multiplying series as long as one of $G_1$ or $G_2$ is a polynomial. For then each coefficient of $G_1G_2$ is a sum of only finitely many terms. As with ordinary polynomials, a polynomial $D$ will be said to *divide* another polynomial $G$ if there is a third polynomial $F$ for which $G = FD$.

THEOREM 2. *A linear dependence between colors $a(P)$ in a set $\psi$ follows from* (2) *if and only if $G_\phi$ divides $G_\psi$.*

*Proof.* Define a new sum

$$A = \sum_{\xi,\eta} a(\xi, \eta)x^{-\xi}y^{-\eta}. \tag{8}$$

$A$ is a generating function, of the form (7), for the set of black squares obtained after rotating the coloring 180°. The series $A$ ordinarily does not converge. Nevertheless, as is usual with generating functions, $A$ may be manipulated formally to derive identities.

Multiply (2) by $x^{-\xi}y^{-\eta}$ and sum over $-\infty < \xi < \infty$, $-\infty < \eta < \infty$. The result is

$$AG_\phi = bU, \tag{9}$$

where $U = \sum x^\xi y^\eta$, containing a term for each square $(\xi, \eta)$.

If $G_\phi$ divides $G_\psi$, then $G_\psi = G_\phi R$ for some polynomial $R$. From (9) one obtains

$$AG_\psi = bUR. \tag{10}$$

From the way $U$ was defined, the right-hand side of (10) is either 0 or $bU$, depending on whether $R$ contains an even or odd number of terms. The left-hand side, $AG_\psi$, is a generating function with coefficients that are linear combinations of certain colors $a(P)$. A linear dependence

$$\sum_\psi a(P_i) = 0 \text{ or } b \tag{11}$$

follows by equating coefficients of $x^0y^0$ in (10).

Conversely suppose a linear dependence, involving the sum (11), follows from (2). Any derivation of this dependence must express the sum (11) as a sum of functions on the left side of certain equations in the system (2). Let the equations in question be the ones obtained by substituting $Q_1, \cdots, Q_M$ for $(\xi, \eta)$ in (2), i.e.,

$$\sum_{m=1}^{M} \sum_{k=1}^{K} a(Q_m + F_k) = \sum_\psi a(P_i). \tag{12}$$

For (12) to hold as an identity, each $P_i$ must appear an odd number of times among the terms $Q_m + F_k$, and any other square must appear an even number of times. But that condition is simply

$$G_\rho G_\phi = G_\psi, \tag{13}$$

where $\rho = \{Q_1, \cdots, Q_M\}$, i.e., $G_\phi$ divides $G_\psi$.

COROLLARY. *A linear dependence* (11) *cannot hold for an odd number $n$ of squares $P_1, \cdots, P_n$ if the floater $\phi$ contains an even number $K$ of squares.*

*Proof.* Any dependence would imply a polynomial identity (13). Set $x = y = 1$. If $n$ is odd and $K$ is even, then $G_\psi = 1$ and $G_\phi = 0$, contradicting (13).

**5. Examples.** Some examples follow to illustrate Theorem 2. It will be helpful to translate the floater $\phi$ and the set $\psi$ down and to the left as far as possible in the positive quadrant. Then $G_\phi$ and $G_\psi$ are always polynomials not divisible by $x$ or $y$, and containing no negative powers of $x$ or $y$.

*Example* a. The polynomial $G_\phi = 1+x+y$ describes an $L$-shaped floater. To verify the randomness property of order 2 one may show that $G_\phi$ divides no $G_\psi$ of the form $x^a y^b + x^c y^d$. By factoring out powers of $x$ and $y$ from $G_\psi$, one can reduce $G_\psi$ to one of four forms $1+x^a$, $1+y^b$, $1+x^a y^b$, or $x^a + y^b$. The first two can be ruled out because any product of $G_\phi$ and another polynomial $G_\rho$ contains terms in which both $x$ and $y$ appear. $G_\phi$ cannot divide $1+x^a y^b$ or $x^a + y^b$ because, substituting 0 for $y$, $1+x$ does not divide 1 or $x^a$. Randomness of order 2 may also be proved by observing that any two points $P_1$, $P_2$ lie in some standard initial set.

*Example* b. Suppose the floater $\phi$ is a rectangle, $\alpha$ cells wide and $\beta$ cells high with $\alpha \geqq 2$ and $\beta \geqq 2$. Then

$$G_\phi = (1+x+\cdots+x^{\alpha-1})(1+y+\cdots+y^{\beta-1})$$

and

$$(1+x)(1+y)G_\phi = (1+x^\alpha)(1+y^\beta). \tag{14}$$

If $\psi$ is the set $\{(0,0), (\alpha,0), (0,\beta), (\alpha,\beta)\}$ then (14) shows $G_\phi$ divides $G_\psi$. The colors at the four points of $\psi$ are linearly dependent. Conversely any three points of the plane belong to some standard initial set and so a rectangular floater can generate a coloring with 3rd order randomness.

Section 2 gave a simple solution of (2) in the special case of a $2\times 2$ square ($\alpha = \beta = 2$). That solution can be derived by considering

$$\begin{aligned} G_\psi &= (1+x^{\xi_0})(1+y^{\eta_0}) \\ &= (1+x+\cdots+x^{\xi_0-1})(1+y+\cdots+y^{\eta_0-1})G_\phi. \end{aligned}$$

The linear dependence that follows from (10) is just the solution given in § 2. The even texture with this floater has a very regular appearance in spite of having 3rd order randomness. The coloring resembles a checkerboard, but made from black and white rectangles of random dimensions (see Julesz, Gilbert, and Victor [7]). The odd coloring function ($b = 1$) differs from an even one by an added term $\xi\eta$ which breaks up the rectangles.

This example generalizes to any $G_\phi$ that is a product

$$G_\phi = f_1(x) f_2(y).$$

Any polynomial $f_1(x)$ of degree $d_1$ divides a polynomial of the form $1+x^{T_1}$ for some $T_1 < 2^{d_1}$. Then $G_\phi$ divides a polynomial $(1+x^{T_1})(1+y^{T_2})$. Although four colors have now been shown linearly dependent, their cells may be far apart because the periods $T_1$, $T_2$ can grow exponentially with the degrees of $f_1$ and $f_2$.

*Example* c. If $G_\phi = 1+x+x^2+xy$ the floater has four squares arranged as a letter $T$ upside down. The same argument used in Example a shows that no two cells have linearly dependent colors. Then the corollary shows that any three colors are linearly independent.

*Example* d. With $G_\phi = 1+x+y+x^2+xy$, $(1+x)G_\phi = 1+y+x^3+x^2y$. Then there is a linear dependence between four colors. Any three cells have independent colors because three cells always belong to some standard row initial set.

*Example* e. A floater of five squares arranged as a plus sign has $G_\phi = x+y+xy+xy^2+x^2y$. Consider any identity $G_\rho G_\phi = G_\psi$. The terms of highest degree in $G_\psi$ arise as products of terms of highest degree from $G_\rho$ and $G_\phi$. The same is true of the terms of lowest degree. $G_\psi$ contains highest degree terms that are multiple of $xy^2+x^2y = xy(x+y)$ and also lower degree terms that are a multiple of $x+y$. Then $G_\psi$ contains at least four terms. Any three colors are independent.

*Example* f.

$$G_\phi = 1+x+y+x^2+y^2+x^3+x^2y+xy^2+y^3$$

and

$$(1+x+y)G_\phi = 1+x^4+y^4.$$

Thus this complicated floater of 9 cells has 3 linearly dependent colors.

With a bit more difficulty, one can show that a six cell floater, with polynomial $(x+y)(1+x)(1+y)$, generates colorings with randomness of order 5. For every $n$ there may exist floaters that produce randomness of order $n$, but that has not been proved. Floaters with $n+1$ cells would be especially interesting. If such floaters exist they cannot approximate simple regions like squares or ellipses, for $(1+x)G_\phi$, $(1+y)G_\phi$ and $(x+y)G_\phi$ would then be polynomials with relatively few terms, identifying cells only near the boundary of the region.

**6. Even and odd.** If there is a dependence between $n$ colors, it can still happen that the even ($b=0$) and odd ($b=1$) colorings have the same $n$-gram probabilities, even though (1) fails to hold. These colorings are still useful for texture discrimination experiments.

If $b=1$ and the colors $a(P_1), \cdots, a(P_n)$ satisfy a linear dependence (11), the right-hand side is 0 or 1 depending on whether the polynomial $R$ in (10) (called $G_\rho$ in (13)) has an even or odd number of terms. If the number of terms is even, then both even and odd colorings satisfy the same dependence. The even and odd textures fail to have the same $n$-gram probabilities only if (13) holds for some $G_\psi$ with $\leqq n$ terms and for some $G_\rho$ with an odd number of terms.

To decide whether the even and odd colorings have equal $n$-gram probabilities it suffices to check that there are no identities (13) with $G_\rho$ having an odd number of terms. If such an identity exists, consider the effect of substituting $x=y=1$. $G_\rho = 1$ because it has an odd number of terms; $G_\phi$ and $G_\psi$ become the number $K$ of cells in the floater and the number $n$. Thus the only dependences to consider are those with $n$ of the same parity as $K$.

In Example b, any 3 colors were independent. If the rectangle contains an odd number of cells, the even and odd colorings have the same 4-gram probabilities.

In Examples d and e, $K=5$ and any three cells have independent colors. Then the even and odd colorings have the same 4-gram probabilities.

In Example f, there were three linearly dependent colors, found by using a polynomial $G_\rho = 1+x+y$ with an odd number of terms. The even and odd colorings have different trigram probabilities.

## REFERENCES

[1] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw Hill, New York, 1968.

[2] W. DIFFIE AND M. E. HELLMAN, *Privacy and authentication: An introduction to cryptography*, Proc. IEEE, 67 (1979), pp. 397–427.

[3] S. W. GOLOMB, *Shift Register Sequences*, Holden-Day, San Francisco, 1967.

[4] B. JULESZ, *Visual pattern discrimination*, IRE Trans. Information Theory, IT8 (1962), pp. 84–92.
[5] ———, *Experiments in the visual perception of textures*, Scientific American, 232 (1975), pp. 34–43.
[6] B. JULESZ, E. N. GILBERT, L. A. SHEPP AND H. L. FRISCH, *Inability of humans to discriminate between visual textures that agree in second-order statistics*, Perception, 2 (1973), pp. 391–405.
[7] B. JULESZ, E. N. GILBERT AND J. D. VICTOR, *Visual discrimination with identical third-order statistics*, Biol. Cybernet. 31 (1978), pp. 137–140.
[8] M. ROSENBLATT AND D. SLEPIAN, *N-th order Markov chains with any set of N variables independent*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 537–549.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980

0196-5212/80/0102-0003 $01.00/0

# SOME ERDÖS–KO–RADO THEOREMS FOR CHEVALLEY GROUPS*

DENNIS STANTON†

**Abstract.** For each infinite family of Chevalley groups over a finite field an Erdös–Ko–Rado theorem is given. The technique uses orthogonal polynomials to find upper bounds for the independence number of specific graphs. In all but one case the bound is realizable.

Erdös, Ko and Rado [5], [6, Thm. 5.3] proved the following theorem for subsets of a set:

THEOREM. *Let $F$ be a collection of $k$-subsets of an $n$-set, $k \leqq n/2$. If $A, B \in F$ implies $|A \cap B| \neq 0$, then $|F| \leqq \binom{n-1}{k-1}$. Equality occurs if $F$ consists of all $k$-subsets which contain a fixed point.*

Lovasz [9] has given an analytic proof of this theorem by computing the capacity of an associated graph. In this paper, we use this technique to derive analogous theorems for Chevalley groups.

First we give a modified version of Lovasz's proof [9]. Let $G_s$, $1 \leqq s \leqq k$, be the graph whose nodes are the $k$-subsets of an $n$-set. Two nodes $A$ and $B$ are joined by an edge if $|A \cap B| = k - s$. An upper bound for the independence number of any regular graph [9, Thm. 9] is $-v\lambda_{\min}/(\lambda_{\max} - \lambda_{\min})$, where $v$ = number of nodes, and $\lambda_{\max(\min)}$ is the maximum (minimum) eigenvalue of the adjacency matrix of the graph. It is clear that $G_s$ is regular. It is also well-known [2, p. 48] that the eigenvalues of the adjacency matrix for $G_s$ are Hahn polynomials:

$$\lambda_j(s) = \binom{k}{s}\binom{n-k}{s} Q_j(s; k-n-1, -k-1, k)$$

$$= \binom{k}{s}\binom{n-k}{s} {}_3F_2\left(\begin{matrix} -j, j-n-1, -s \\ -k, k-n \end{matrix} \middle| 1\right), \qquad j = 0, 1, \cdots, k.$$

(See [3, p. 356] for a short exposition of these polynomials.) For the Erdös–Ko–Rado theorem, we put $s = k$ and evaluate the resulting ${}_2F_1$ [1, p. 3] to obtain

$$\lambda_j(k) = (-1)^j \binom{n-k-j}{k-j}, \qquad j = 0, 1, \cdots, k.$$

Since $k \leqq n/2$, we have $\lambda_{\max} = \binom{n-k}{k}$ and $\lambda_{\min} = -\binom{n-k-1}{k-1}$. With $v = \binom{n}{k}$ this yields an upper bound of $\binom{n-1}{k-1}$. In fact Lovasz's theorem shows that the capacity of this graph is $\binom{n-1}{k-1}$.

Next we state and prove the Erdös–Ko–Rado theorem for a Chevalley group of type $A_{n-1}$ over $GF(q)$. For $k \neq n/2$ this theorem is due to Hsieh [7] (see also [6, Thm. 5. 7]).

* Received by the editors September 18, 1979.
† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This research was supported in part by the National Science Foundation under Grant MCS-78-18222.

THEOREM. *Let $F$ be a collection of $k$-dimensional subspaces of an $n$-dimensional vector space over $GF(q)$, $k \leqq n/2$. If $A, B \in F$ implies* $\dim(A \cap B) \neq 0$, *then* $|F| \leqq \begin{bmatrix} n-1 \\ k-1 \end{bmatrix}$.

To prove this theorem, we define the analogous regular graph $G_s$. The eigenvalues are $q$-Hahn polynomials [4, Thm. 3.7]

$$\lambda_j(s) = \begin{bmatrix} k \\ s \end{bmatrix}\begin{bmatrix} n-k \\ s \end{bmatrix} q^{s^2} Q_j(q^{-s}; q^{k-n-1}, q^{-k-1}, n; q)$$

$$= \begin{bmatrix} k \\ s \end{bmatrix}\begin{bmatrix} n-k \\ s \end{bmatrix} q^{s^2} {}_3\phi_2\left(\begin{matrix} q^{-j}, q^{j-n-1}, q^{-s} \\ q^{-k}, q^{k-n} \end{matrix} \middle| q; q\right), \qquad j = 0, 1, \cdots, k.$$

As before, we put $s = k$ and evaluate the ${}_2\phi_1$ [1, p. 68] to obtain

$$\lambda_j(k) = (-1)^j \begin{bmatrix} n-k-j \\ k-j \end{bmatrix} q^{\binom{j}{2} + k^2 - kj}.$$

It is easy to check that $|\lambda_{j+1}(k)/\lambda_j(k)| \leqq 1$ if and only if $q^{n-k-j} + q^{-k+j} \geqq 2$. However, $q \geqq 2$, $j+1 \leqq k$, and $k \leqq n/2$ imply that this inequality always holds. Thus $\lambda_{\max} = \begin{bmatrix} n-k \\ k \end{bmatrix} q^{k^2}$, $\lambda_{\min} = -\begin{bmatrix} n-k-1 \\ k-1 \end{bmatrix} q^{k^2-k}$, and $v = \begin{bmatrix} n \\ k \end{bmatrix}$ yield a bound of $\begin{bmatrix} n-1 \\ k-1 \end{bmatrix}$.

There is a variant of the Erdös–Ko–Rado theorem related to the Weyl groups of the other simple Lie algebras. We let $H(N, 2)$, $N \geqq 2$, be the binary Hamming scheme whose elements are the $N$-tuples of 0's and 1's. We consider collections $F$ of $N$-tuples such that if $A, B \in F$, then $A$ and $B$ agree in at least one entry. The theorem states that $|F| \leqq 2^{N-1}$. Although this is trivial the $q$-analog will not be so simple. It is clear that the bound $2^{N-1}$ can be attained by letting $F$ consist of all $N$-tuples with one entry fixed. However, for $N$ odd we could also let $F$ be the set of all $N$-tuples with an even number of 1's. There are $q$-analogs of both cases. The bounds are not the same.

In passing, we mention that the Hamming scheme $H(N, t)$ has been considered by Livingston [8], who showed that for $t \geqq 3$ the extremal $F$ are exactly those with one, entry constant. The bound $t^{N-1}$ could be obtained by the previous technique. The eigenvalues are Krawtchouk polynomials [3, Prop. 1.13]

$$\lambda_j(s) = \binom{N}{s}(t-1)^s K_j(s, (t-1)/t; N)$$

$$= \binom{N}{s}(t-1)^s {}_2F_1\left(\begin{matrix} -s, -j \\ -N \end{matrix} \middle| \frac{t}{t-1}\right), \qquad j = 0, 1, \cdots, N.$$

Putting $s = N$, we have $\lambda_j(N) = (-1)^j(t-1)^{N-j}$, which clearly gives the desired bound. Unfortunately the $q$-analog of $H(N, t)$ for $t \geqq 3$ is not known.

We concentrate on the $q$-analogs of $H(N, 2)$. There are six [10] which are associated with the Chevalley groups of types $B_N$, $C_N$, $D_N$, ${}^2D_{N+1}$, ${}^2A_{2N-1}$, and ${}^2A_{2N}$ over a finite field $K$. Let $p = |K|$, so that $p = q$ for $B_N$, $C_N$, $D_N$, and ${}^2D_{N+1}$; and $p = q^2$ for ${}^2A_{2N-1}$ and ${}^2A_{2N}$. Let $V$ be the appropriate vector space over $K$ for each type, and let $B$ be the corresponding bilinear form. We assume for each case that the dimension of a maximal isotropic subspace of $V$ is $N$. Then the $q$-analog of $H(N, 2)$ is the collection $X_N$ of all maximal isotropic subspaces of $V$ [10]. Furthermore $|X_N| = \prod_{i=1}^{N}(1 + cp^i)$, where $c = 1$ for types $B_N$ and $C_N$, $c = q^{-1}$ for types $D_N$ and ${}^2A_{2N-1}$, and $c = q$ for types ${}^2D_{N+1}$ and ${}^2A_{2N}$.

For an Erdös–Ko–Rado theorem consider collections $F$ of maximal isotropic subspaces such that if $A, B \in F$, $\dim(A \cap B) \neq 0$. Let $G_s$ be the graph with maximal isotropic subspaces as nodes, and edges $(A, B)$ if $\dim(A \cap B) = N - s$. This graph is regular, and the eigenvalues are $q$-Krawtchouk polynomials [10, Thm. 5.4]

$$\lambda_j(s) = \begin{bmatrix} N \\ s \end{bmatrix} p^{\binom{s+1}{2}} c^s K_j(p^{-s}; c^{-1}p^{-N-2}, N; p)$$

$$= \begin{bmatrix} N \\ s \end{bmatrix} p^{\binom{s+1}{2}} c^s {}_3\phi_2\left( \begin{matrix} p^{-s}, p^{-j}, -c^{-1}p^{j-N-1} \\ 0, p^{-N} \end{matrix} \middle| p; p \right), \qquad j = 0, 1, \cdots, N.$$

Evaluating at $s = N$ [1, p. 68], we obtain

$$\lambda_j(N) = (-1)^j p^{\binom{N+1}{2} + j(j-N-1)} c^{N-j} \qquad j = 0, 1, \cdots, N.$$

It is easy to find $\lambda_{\max}$ and $\lambda_{\min}$ for each case and thus find a bound on the independence number.

THEOREM. *Let $F$ be a collection of $N$-dimensional maximal isotropic subspaces such that if $A, B \in F$, then* $\dim(A \cap B) \neq 0$. *Then*

(1) $$|F| \leqq \prod_{i=1}^{N-1} (1 + q^i) \qquad \textit{for types } B_N \textit{ and } C_N,$$

(2) $$|F| \leqq \prod_{i=1}^{N-1} (1 + q^{i+1}) \qquad \textit{for type } {}^2D_{N+1},$$

(3) $$|F| \leqq \prod_{i=1}^{N-1} (1 + q^{2i+1}) \qquad \textit{for type } {}^2A_{2N},$$

(4)
$$|F| \leqq \prod_{i=1}^{N-1} (1 + q^{i-1}) \qquad \textit{if } N \textit{ is even,}$$
$$|F| \leqq \prod_{i=1}^{N-1} (1 + q^i) \qquad \textit{if } N \textit{ is odd, for type } D_N,$$

(5)
$$|F| \leqq \prod_{i=1}^{N-1} (1 + q^{2i-1}) \qquad \textit{if } N \textit{ is even, and}$$
$$|F| \leqq \prod_{\substack{i=1 \\ 2i-1 \neq N}}^{N} (1 + q^{2i-1}) \qquad \textit{if } N \textit{ is odd, for type } {}^2A_{2N-1}.$$

Except for the two cases when $N$ is odd, the bounds are attained by letting $F$ be all maximal isotropic subspaces containing a given vector. The odd case in (4) is realized by taking the family $F$ of maximal isotropic subspaces such that $A, B \in F$ implies $N$-$\dim(A \cap B)$ is even [9, § 5]. We do not have an extremal family $F$ for (5) if $N$ is odd.

For the examples in this paper we have found the capacity of the associated graphs. This answers a question at the end of [9], namely, to find other examples for which the upper bound gives information about the independence number. Other values of $s$ could be substituted to obtain upper bounds. For example, for subsets $s = k - 1$ gives $\lambda_j(k-1) = \binom{n-k}{k-1}\binom{k}{j}(k^2 + j(j-n-1))(-1)^j/k\binom{n-k}{j}$. Also we could replace $|A \cap B| = k - s$ by $|A \cap B| \leqq k - s$ in the definition of the graph $G_s$. The eigenvalues are then sums of Hahn polynomials, because the eigenspace for $\lambda_j(s)$ is an irreducible representation for $S_N$ [3], which is independent of $s$.

**Acknowledgment.** The author would like to thank L. Lovasz for pointing out his analytic proof of the Erdös–Ko–Rado theorem.

## REFERENCES

[1] W. BAILEY, *Generalized Hypergeometric Series*, Cambridge Tracts in Mathematics No. 32, Cambridge University Press, Cambridge, 1935.

[2] PH. DELSARTE, *An algebraic approach to the association schemes of coding theory*, Philips Res. Rep. Suppl., 10 (1973).

[3] C. DUNKL, *A Krawtchouk polynomial addition theorem and wreath products of symmetric groups*, Indiana Univ. Math. J., 25 (1976), pp. 335–358.

[4] ———, *An addition theorem for some q-Hahn polynomials*, Monatsh. Math., 85 (1978), pp. 5–37.

[5] P. ERDÖS, C. KO AND R. RADO, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Ser. (2), 12 (1961), pp. 313–318.

[6] C. GREENE AND D. KLEITMAN, *Proof techniques in the theory of sets*, Studies in Combinatorics, Mathematical Association of America Studies in Mathematics Vol. 17, Mathematical Association of America, Washington, DC, 1978, pp. 22–79.

[7] W. HSIEH, *Intersection theorems for systems of finite vector spaces*, Discrete Math., 12 (1975), pp. 1–16.

[8] M. LIVINGSTON, *An ordered version of the Erdös-Ko-Rado theorem*, J. Combinatorial Theory Ser. A, 26 (1979), pp. 162–165.

[9] L. LOVASZ, *On the Shannon capacity of a graph*, IEEE Trans. Information Theory, IT-25 (1979), pp. 1–7.

[10] D. STANTON, *Some q-Krawtchouk polynomials on Chevalley groups*, preprint.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0004 $01.00/0

# A NEW LOWER BOUND FOR THE NUMBER OF SWITCHES IN REARRANGEABLE NETWORKS*

NICHOLAS PIPPENGER†

**Abstract.** For the commonest model of rearrangeable networks with $n$ inputs and $n$ outputs, it is shown that such a network must contain at least $6n \log_6 n + O(n)$ switches. Similar lower bounds for other models are also presented.

**1. Introduction.** The lower bound referred to in the title will be established by modeling a rearrangeable network as a directed graph in which vertices represent wires and edges represent switches. A number of alternative models will be considered later.

A *n-network* $N = (G, A, B)$ comprises a directed graph $G = (V, E)$, with vertices $V$ and edges $E$, a set $A$ of $n$ distinguished vertices called *inputs*, and a set $B$, disjoint from $A$, of $n$ distinguished vertices called *outputs*.

A *request* for $N$ is an ordered pair $(a, b)$ comprising an input $a$ and an output $b$. An *assignment* for $N$ is a set of requests for $N$, no two having an input or output in common. *A k-assignment* for $N$ is an assignment containing exactly $k$ requests.

A *route* in $N$ is a directed path in $G$, starting at an input and ending at an output. A *state* of $N$ is a set of routes in $N$, no two having a vertex in common. The set of states of $N$ will be denoted $\Omega$. A *k-state* of $N$ is a state of $N$ containing exactly $k$ routes. The set of $k$-states of $N$ will be denoted $\Omega_k$.

An assignment is said to be *realized* by a state if, for every request $(a, b)$ in the assignment, there is a route from $a$ to $b$ in the state. An $n$-network $N$ is an *n-connector* if each of the $n!$ $n$-assignments for $N$ is realized by some state of $N$.

An $n$-connector must satisfy the lower bound

$$|E| \geqq 3n \log_3 n + O(n) \tag{1}$$

$(3/\ln 3 = 2.730\cdots)$; this follows from the inequality

$$|E| \geqq 3 \log_3 |\Omega_n|$$

(attributed to R. L. Dobrushin by Bassalygo and Tsybakov [1]), from the obvious inequality

$$|\Omega_n| \geqq n!$$

(distinct assignments must be realized by distinct states), and from the estimate

$$\log n! = n \log n + O(n)$$

(due to Stirling [7, p. 137]).

The purpose of this note is to derive the improved lower bound

$$|E| \geqq 6n \log_6 n + O(n) \tag{2}$$

$(6/\ln 6 = 3.348\cdots)$; this will follow from the improved inequality

$$|E| \geqq 6 \log_6 |\Omega_n|. \tag{3}$$

* Received by the editors October 3, 1979, and in revised form October 24, 1979.

† Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

These lower bounds may be compared with the upper bound for $n$-connectors,

$$|E| \leqq 6n \log_3 n + O(n)$$

($6/\ln 3 = 5.461 \cdots$; see Pippenger and Valiant [5, Remark 2.2.6]).

The qualitative significance of these improvements may be seen by comparing these bounds with the corresponding bounds for $n$-shifters ($n$-networks that need not have states realizing all $n!$ assignments, but only the $n$ assignments corresponding to cyclic permutations). The inequality (1) actually follows immediately from the even sharper lower bound for $n$-shifters,

$$|E| \geqq 3n \log_3 n$$

(see Pippenger and Valiant [5, Corollary 2.2.2]). This may be compared with the upper bound for $n$-shifters,

$$|E| \leqq 3n \log_3 n + O(n)$$

(see Pippenger and Valiant [5, Remark 2.2.5]). The results of this note thus show that $n$-connectors require more edges than $n$-shifters, a plausible conclusion which was, however, not deducible from previous results.

**2. The new lower bound.** One may assume, without loss of generality, that no edge is directed into an input or directed out of an output, for no such edge can occur in an $n$-state.

If there is any vertex $v$ in $V-(A \cup B)$ out of which no edge is directed, one may omit $v$ from $V$ and omit each edge of the form $(u, v)$ from $E$. If there is any vertex $v$ in $V-(A \cup B)$ out of which exactly one edge $(v, w)$ is directed, one may omit $v$ from $V$ and replace each edge of the form $(u, v)$ by the edge $(u, w)$ in $E$. In either case one obtains a network with at most as many edges and just as many $n$-states. Thus one may assume, without loss of generality, that at least two edges are directed out of each vertex in $V-(A \cup B)$.

Let $f: B \to A$ be an arbitrary bijection. Let $G^* = (V^*, E^*)$ be the directed graph with vertices $V^*$ and edges $E^*$ obtained from $N$ as follows. Let $V^*$ be obtained from $V$ by omitting the vertices in $B$. Let $E^*$ be obtained from $E$ by replacing each edge of the form $(v, b)$ in $V \times B$ by the edge $(v, f(b))$ in $V \times A$, and by adding the edge $(v, v)$ for each vertex $v$ in $V-(A \cup B)$.

The edges of the form $(v, v)$ added to $E^*$ are directed out of vertices $v$ in $V-(A \cup B)$, and $E$ contains at least two edges directed out of each such vertex. Thus

$$|E^*| \leqq \tfrac{3}{2}|E|. \tag{4}$$

A set of closed directed paths in $G^*$ containing exactly one edge directed into each vertex and one edge directed out of each vertex will be called a *circulation* in $G^*$. The set of circulations in $G^*$ will be denoted $\Omega^*$. Each $n$-state of $N$ corresponds to a circulation in $G^*$ (by replacing edges of the form $(v, b)$ by edges $(v, f(b))$ and adding edges of the form $(v, v)$ as necessary), and distinct $n$-states correspond to distinct circulations. Thus

$$|\Omega^*| \geqq |\Omega_n|.$$

By virtue of these inequalities, it will suffice to show

$$|E^*| \geqq 9 \log_6 |\Omega^*| \tag{5}$$

for an arbitrary directed graph $G^*$.

Let $M$ be the $(0, 1)$-matrix with rows and columns indexed by $V^*$ and with $(v, w)$th entry $M_{v,w}$ equal to 1 or 0 according as $(v, w)$ does or does not appear in $E^*$. Let

$$L_v = \sum_{w \in V^*} M_{v,w}$$

denote the sum of the entries in the $v$th row of $M$. Then

$$|E^*| = \sum_{v \in V^*} L_v.$$

On the other hand,

$$|\Omega^*| = \operatorname{per} M,$$

where $\operatorname{per} M$ denotes the permanent of $M$, since both sides count the number of permutations $g$ of $V^*$ for which $(v, g(v))$ appears in $E^*$ for each $v$ in $V^*$. Thus it will suffice to show

$$\sum_{v \in V^*} L_v \geqq 9 \log_6 \operatorname{per} M \tag{6}$$

for an arbitrary $(0, 1)$-matrix $M$.

The inequality

$$\sum_{v \in V^*} (\log (L_v!))/L_v \geqq \log \operatorname{per} M$$

for an arbitrary $(0, 1)$-matrix $M$ was conjectured by Minc [3] and proved by Bregman [2] (see Schrijver [6] for a particularly simple and elegant proof). Since the expression $(\log (L!))/L^2$ assumes its maximum over integers $L$ at $L = 3$,

$$\begin{aligned}\sum_{v \in V^*} L_v &\geqq (3^2/\log (3!)) \sum_{v \in V^*} (\log (L_v!))/L_v \\ &\geqq 9 \sum_{v \in V^*} (\log_6 (L_v!))/L_v \\ &\geqq 9 \log_6 \operatorname{per} M.\end{aligned}$$

This proves (6), and thus establishes (5), (3) and (2) in turn.

**3. Other new lower bounds.** The argument of this note is easily extended to a number of other models of rearrangeable networks. The most interesting of these is obtained by replacing directed graphs and directed paths by undirected graphs and undirected paths. A directed graph $G = (V, E)$ can be obtained from an undirected graph $G' = (V', E')$ by setting $V = V'$ and replacing each undirected edge $\{v, w\}$ in $E'$ by a pair of directed edges $(v, w)$ and $(w, v)$, so that

$$|E'| \geqq \tfrac{1}{2}|E|.$$

In this way, a directed $n$-connector $N = (G, A, B)$ can be obtained from an undirected $n$-connector $N' = (G', A', B')$ by setting $A = A'$ and $B = B'$. One may assume, without loss of generality, that at least three undirected edges are incident with each vertex in $V' - (A' \cup B')$, so that

$$|E^*| \leqq \tfrac{4}{3}|E|. \tag{7}$$

Continuing with the argument of § 2 leads to the lower bound

$$|E'| \geqq \tfrac{27}{8} n \log_6 n + O(n)$$

$(27/8 \ln 6 = 1.883 \cdots)$ for undirected $n$-connectors. This may be compared with the previous bound

$$|E'| \geqq \tfrac{5}{2} n \log_4 n + O(n)$$

$(5/2 \ln 4 = 1.803 \cdots)$ which applies even to undirected $n$-shifters (see Pippenger and Valiant [5, Thm. 2.2.3]). No better upper bounds are known for undirected $n$-connectors and $n$-shifters than for their directed counterparts.

Other, even easier, extensions are to consider "single-ended" or "undifferentiated" $n$-connectors in which the $n$ inputs and $n$ outputs are replaced by a single undifferentiated set of $n$ distinguished vertices called "terminals" (this reduces the leading terms of lower bounds by a factor of 2) and to bound $\log |\Omega|$ rather than merely $\log |\Omega_n|$ (this affects only the $O(n)$ terms). These extensions yield improvements of the results in Pippenger [4].

James Shearer, the referee for this paper, has pointed out some improvements to the foregoing results. In the directed case, a vertex $w$ in $V - (A \cup B)$ into which only two edges $(u, w)$ and $(v, w)$ are directed and out of which only two edges $(w, x)$ and $(w, y)$ are directed can be omitted, the edges being replaced by $(u, x)$, $(v, x)$, $(u, y)$ and $(v, y)$. Repeating this transformation as long as possible yields a graph with just as many edges and $n$-states but in which a total of at least five edges are directed into or out of each vertex in $V - (A \cup B)$. This allows (4) to be sharpened to

$$|E^*| \leqq \tfrac{7}{5} |E|,$$

and results in a lower bound of

$$|E| \geqq \tfrac{45}{7} n \log_6 n + O(n)$$

$(45/7 \ln 6 = 3.587 \cdots)$. Similarly, in the undirected case, a vertex $w$ in $V - (A \cup B)$ incident with only three edges $\{w, x\}$, $\{w, y\}$ and $\{w, z\}$ can be omitted, the edges being replaced by $\{x, y\}$, $\{y, z\}$ and $\{z, x\}$. This yields a graph in which every vertex in $V - (A \cup B)$ is incident with at least four edges, allows (7) to be sharpened to

$$|E^*| \leqq \tfrac{5}{4} |E|,$$

and results in a lower bound of

$$|E'| \geqq \tfrac{18}{5} n \log_6 n + O(n)$$

$(18/5 \ln 6 = 2.009 \cdots)$.

## REFERENCES


[1] L. A. BASSALYGO AND B. S. TSYBAKOV, *Blocking probability for a rearrangeable switching system*, Problems of Information Transmission, 6 (1973), pp. 336–348.

[2] L. M. BREGMAN, *Certain properties of nonnegative matrices and their permanents*, Soviet. Math. Dokl., 14 (1973), pp. 945–949.

[3] H. MINC, *Upper bounds for permanents of* (0, 1)*-matrices*, Bull. Amer. Math. Soc., 69 (1963), pp. 789–791.

[4] N. PIPPENGER, *The complexity of seldom-blocking networks*, IEEE Internat. Comm. Conf., 12 (1976), pp. 7-8—7-12.

[5] N. PIPPENGER AND L. G. VALIANT, *Shifting graphs and their applications*, J. Assoc. Comput. Mach., 23 (1976), pp. 423–432.

[6] A. SCHRIJVER, *A short proof of Minc's conjecture*, J. Combinatorial Theory (A), 25 (1978), pp. 80–83.

[7] J. STIRLING, *Methodus Differentialis*, London, 1730.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0005 $01.00/0

# WEYL GROUPS, THE HARD LEFSCHETZ THEOREM, AND THE SPERNER PROPERTY*

RICHARD P. STANLEY†

**Abstract.** Techniques from algebraic geometry, in particular the hard Lefschetz theorem, are used to show that certain finite partially ordered sets $Q^X$ derived from a class of algebraic varieties $X$ have the $k$-Sperner property for all $k$. This in effect means that there is a simple description of the cardinality of the largest subset of $Q^X$ containing no $(k+1)$-element chain. We analyze, in some detail, the case when $X = G/P$, where $G$ is a complex semisimple algebraic group and $P$ is a parabolic subgroup. In this case, $Q^X$ is defined in terms of the "Bruhat order" of the Weyl group of $G$. In particular, taking $P$ to be a certain maximal parabolic subgroup of $G = SO(2n+1)$, we deduce the following conjecture of Erdös and Moser: Let $S$ be a set of $2\ell+1$ distinct real numbers, and let $T_1, \cdots, T_k$ be subsets of $S$ whose element sums are all equal. Then $k$ does not exceed the middle coefficient of the polynomial $2(1+q)^2(1+q^2)^2\cdots(1+q^\ell)^2$, and this bound is best possible.

**1. The Sperner property.** Let $P$ be a finite partially ordered set (or *poset*, for short), and assume that every maximal chain of $P$ has length $n$. We say that $P$ is *graded of rank* $n$. Thus $P$ has a unique *rank function* $\rho: P \to \{0, 1, \cdots, n\}$ satisfying $\rho(x) = 0$ if $x$ is a minimal element of $P$, and $\rho(y) = \rho(x) + 1$ if $y$ covers $x$ in $P$ (i.e., if $y > x$ and no $z \in P$ satisfies $y > z > x$). If $\rho(x) = i$, then we say that $x$ has *rank* $i$. Define $P_i = \{x \in P: \rho(x) = i\}$ and set $p_i = p_i(P) = \text{card } P_i$. The polynomial $F(P, q) = p_0 + p_1 q + \cdots + p_n q^n$ is called the *rank-generating function* of $P$. We say that $P$ is *rank-symmetric* if $p_i = p_{n-i}$ for all $i$, and that $P$ is *rank-unimodal* if $p_0 \leqq p_1 \leqq \cdots \leqq p_i \geqq p_{i+1} \geqq \cdots \geqq p_n$ for some $i$.

*An antichain* (also called a *Sperner family* or *clutter*) is a subset $A$ of $P$, such that no two distinct elements of $A$ are comparable. The poset $P$ is said to have the *Sperner property* (or *property* $S_1$) if the largest size of an antichain is equal to $\max\{p_i: 0 \leqq i \leqq n\}$. More generally, if $k$ is a positive integer then $P$ is said to have the *$k$-Sperner property* (or property $S_k$) if the largest subset of $P$ containing no $(k+1)$-element chain has cardinality $\max\{p_{i_1} + \cdots + p_{i_k}: 0 \leqq i_1 < \cdots < i_k \leqq n\}$. If $P$ has property $S_k$ for all $k \leqq n$, then following [21] we say that $P$ has *property* S. For further information concerning the Sperner property and related concepts, see for instance [15], [16], [17].

Using some results from algebraic geometry, we will give several new classes of graded posets which have property S. These posets will all be rank-symmetric and rank-unimodal. First we must consider a property of posets related to property S. Suppose $P$ is graded of rank $n$ and is rank-symmetric. Again following [21], we say that $P$ has *property* T if for all $0 \leqq i \leqq [n/2]$, there exist $p_i$ pairwise disjoint saturated chains $x_i < x_{i+1} < \cdots < x_{n-i}$ where $x_j \in P_j$. It is clear that $P$ is then rank-unimodal.

LEMMA 1.1. *Let $P$ be a finite graded rank-symmetric poset of rank $n$. The following three conditions are equivalent:*

(i) *$P$ is rank-unimodal and has property* S.

(ii) *$P$ has property* T.

(iii) *Let $V_i$ be the complex vector space with basis $P_i$. Then for $0 \leqq i < n$, there exist linear transformations $\varphi_i: V_i \to V_{i+1}$ satisfying the following two properties:*

(a) *If $0 \leqq i \leqq [n/2]$, then the composite transformation $\varphi_{n-i-1}\varphi_{n-i-2}\cdots\varphi_{i+1}\varphi_i: V_i \to V_{n-i}$ is invertible.*

(b) *Let $x \in P_i$ and $\varphi_i(x) = \sum_{y \in P_{i+1}} c_y y$. Then $c_y = 0$ unless $x < y$.*

*Proof.* (i)⇔(ii). This is a special case of [21, Thms. 2 and 3].

* Received by the editors June 1, 1979.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The research was supported in part by the National Science Foundation under Grant MCS 77-01947.

(iii)$\Rightarrow$(ii). (I am grateful to Joseph Kung for supplying the following argument, which is considerably simpler than my original proof.) Assume (iii). Identify $\varphi_i$ with its matrix with respect to the bases $P_i$ and $P_{i+1}$. If $\varphi$ is a matrix whose rows are indexed by a set $S$ and whose columns are indexed by $T$, and if $S' \subset S$ and $T' \subset T$, then let $\varphi[S', T']$ denote the submatrix of $\varphi$ with rows indexed by $S'$ and columns by $T'$. By the Binet-Cauchy theorem (e.g., [1, § 36]) we have

$$\det(\varphi_{n-i-1} \cdots \varphi_i) = \sum (\det \varphi_i[Q_i, Q_{i+1}]) \cdot (\det \varphi_{i+1}[Q_{i+1}, Q_{i+2}]) \cdots (\det \varphi_{n-i-1}[Q_{n-i-1}[Q_{n-i}]),$$

where the sum is over all sequences of subsets $Q_i = P_i$, $Q_{i+1} \subset P_{i+1}$, $Q_{i+2} \subset P_{i+2}, \cdots, Q_{n-i-1} \subset P_{n-i-1}$, $Q_{n-i} = P_{n-i}$ such that $|Q_{i+1}| = |Q_{i+2}| \cdots = |Q_{n-i-1}| = p_i$. By (a), some term in the above sum is nonzero. Hence, the expansion of each factor $\det \varphi_k[Q_k, Q_{k+1}]$ in this term contains a nonzero term. By (b), this nonzero term defines a map $\sigma: Q_k \to Q_{k+1}$ such that $x < \sigma(x)$ for all $x \in Q_k$. Piecing together these two-element chains over all $k$ yields (ii).

(ii)$\Rightarrow$(iii). The steps of the above argument can be reversed, provided we pick the $\varphi_i$'s as generically as possible, i.e., all the entries of the matrices $\varphi_0, \varphi_1, \cdots, \varphi_{n-1}$ should be chosen to be algebraically independent over $\mathbb{Q}$, except for entries forced to equal 0 by condition (b). This completes the proof. □

**2. Varieties with cellular decompositions.** We now are in a position to invoke algebraic geometry. Let $X$ be a complex projective variety of complex dimension $n$. Suppose that there are finitely many pairwise-disjoint subsets $C_i$ of $X$, each isomorphic as an algebraic variety to complex affine space of some dimension $n_i$, such that (i) the union of the $C_i$'s is $X$, and (ii) $\bar{C}_i - C_i$ is a union of some of the $C_j$'s. (Here $\bar{C}_i$ denotes the closure of $C_i$ either in the Hausdorff or Zariski topology—under the present circumstances the two closures coincide.) Following [4, p. 500], we then say that the $C_i$'s form a *cellular decomposition* of $X$. The simplest and most familiar example is complex projective space $\mathbb{P}^n$ itself. Recall that $\mathbb{P}^n$ may be regarded as the set of nonzero $(n+1)$-tuples $x = (x_0, x_1, \cdots, x_n) \in \mathbb{C}^{n+1}$, modulo the equivalence relation $x \sim \lambda x$ $(\lambda \in \mathbb{C}^*)$. The set of elements of $\mathbb{P}^n$ of the form $(0, \cdots, 0, 1, x_{n-i+1}, \cdots, x_n)$ forms a subvariety isomorphic to $\mathbb{C}^i$. Hence we have the cellular decomposition $\mathbb{P}^n = \mathbb{C}^n \cup \mathbb{C}^{n-1} \cup \cdots \cup \mathbb{C}^0$.

If $X$ is any complex projective variety and $Y$ is a closed subvariety, then e.g., by [4] or [18, Chap. 5, § 4], $Y$ represents an element (cocycle) $[Y]$ of the cohomology group $H^*(X, \mathbb{C})$. If $X$ is irreducible of (complex) dimension $n$, and $Y$ is irreducible of dimension $m$, then in fact $[Y] \in H^{2(n-m)}(X, \mathbb{C})$. If $X$ is irreducible of dimension $n$ and has a cellular decomposition $\{C_i\}$, it follows that the closures $\bar{C}_i$ represent cohomology classes $[\bar{C}_i] \in H^{2(n-m)}(X, \mathbb{C})$ where $C_i \cong \mathbb{C}^m$. (For this fact, we don't need condition (ii) in our definition of cellular decomposition.) The following fundamental result concerning varieties with a cellular decomposition appears in [4, p. 501], [22, § 6] in the case when $X$ is nonsingular. The extension to singular varieties follows from [14]. (Again, condition (ii) is not actually necessary.)

THEOREM 2.1. *Let $X$ be a complex projective variety of complex dimension $n$, and suppose that $X$ has a cellular deomposition $\{C_i\}$. Then the cohomology classes $[\bar{C}_i]$ form a basis (over $\mathbb{C}$) for $H^*(X, \mathbb{C})$. In particular, $H^{2m+1}(X, \mathbb{C}) = 0$ for all $m \in \mathbb{Z}$, while if $X$ is irreducible then $H^{2(n-m)}(X, \mathbb{C})$ has a basis consisting of those classes $[\bar{C}_i]$ for which $C_i \cong \mathbb{C}^m$.* □

Now given a cellular decomposition $\{C_i\}$ of $X$, define a partial ordering $Q^X = Q^X(C_1, C_2, \cdots)$ on the $C_i$'s by setting $C_i \geqq C_j$ in $Q^X$ if $C_i \subset \bar{C}_j$. If $X$ is irreducible of

dimension $n$, then it can be shown, using standard techniques from algebraic geometry, that $Q^X$ is graded of rank $n$, with the rank function given by $\rho(C) = n\text{-dim } C$. If, moreover, $X$ is nonsingular, then Poincaré duality implies that $Q^X$ is rank-symmetric. Theorem 2.1 then implies that we may identify the vector space $V_i$ of Lemma 1.1 (iii) with $H^{2i}(X, \mathbb{C})$ by identifying $C \in Q_i^X$ with $[\bar{C}] \in H^{2i}(X, \mathbb{C})$.

We now wish to define linear transformations $\varphi_i: V_i \to V_{i+1}$ (or equivalently, $\varphi_i: H^{2i}(X, \mathbb{C}) \to H^{2(i+1)}(X, \mathbb{C})$) satisfying conditions (a) and (b) of Lemma 1.1 (iii). This will enable us to conclude that $Q^X$ has property $S$. Let $Y$ be a hyperplane section of $X$, i.e., the intersection of $X$ (regarded as being imbedded in some projective space $\mathbb{P}^N$) with a hyperplane of $\mathbb{P}^N$. If $X$ is irreducible, then $Y$ is a closed subvariety of $X$ of dimension $n-1$ which represents a cohomology class $[Y] \in H^2(X, \mathbb{C})$. The cup product operation on cohomology then yields a linear transformation $\varphi_i: H^{2i}(X, \mathbb{C}) \to H^{2(i+1)}(X, \mathbb{C})$ defined as multiplication by $[Y]$. In other words, $\varphi_i(K) = [Y] \cdot K$. We now verify that when $X$ is nonsingular and irreducible (so $Q^X$ is graded and rank-symmetric), then these linear transformations $\varphi_i$ satisfy conditions (a) and (b) of Lemma 1.1 (iii). First we dispose of condition (b). I am grateful to Steve Kleiman for providing a proof of this result.

LEMMA 2.2. *Let $X$ be a complex projective variety with a cellular decomposition $\{C_i\}$, and let $Y$ be a hyperplane section (or in fact any closed subvariety) of $X$. If $[Y] \cdot [\bar{C}_i] = \sum \alpha_j [\bar{C}_j]$ in $H^*(X, \mathbb{C})$, then $\alpha_j = 0$ unless $C_j \subset \bar{C}_i$.*

*Proof.* Let $A(W)$ denote the Chow group of the variety $W$, i.e., the group of cycles modulo rational equivalence. If $W$ is nonsingular and has a cellular decomposition $\{D_i\}$, then it is mentioned in [22, § 6] that the cycles $\bar{D}_i$ form a basis for $A(W)$, and that the corresponding map $A(W) \to H^*(W, \mathbb{Z})$ is an isomorphism of groups. It follows from [14] that this result continues to hold when $W$ is singular. Now returning to our hypotheses, the $C_j$'s contained in $\bar{C}_i$ form a cellular decomposition of $\bar{C}_i$. Hence a hyperplane section of $\bar{C}_i$ is rationally equivalent to a linear combination of the $\bar{C}_j$ that are contained in $\bar{C}_i$. A priori, the rational equivalence is on $\bar{C}_i$, but it may be considered as a rational equivalence on $X$. Hence $\alpha_j = 0$ unless $C_j \subset \bar{C}_i$ because the $[\bar{C}_i]$ are linearly independent in $H^*(X, \mathbb{C})$. □

Lemma 2.2 shows that condition (b) of Lemma 1.1 (iii) holds for $Q^X$ (assuming $X$ is nonsingular and irreducible, so we know $Q^X$ is graded and rank-symmetric). Condition (a) is implied by the following basic result, known as the "hard Lefschetz theorem" (although the first rigorous proof was given by Hodge). See [34] for a brief history and survey of this theorem, and for its extension to characteristic $p$. Other references include [24, p. 187], [29], [10, Corollary, p. 75], [30, p. 44], [19, Chap. 0, § 7].

LEMMA 2.3 (the hard Lefschetz theorem). *Let $X$ be a nonsingular irreducible complex projective variety of complex dimension $n$. Let $Y$ be a hyperplane section of $X$. If $0 \leqq i \leqq n$, then the linear transformation $H^i(X, \mathbb{C}) \to H^{2n-i}(X, \mathbb{C})$ given by multiplication by $[Y]^{n-i}$ is an isomorphism.*

Putting Lemmas 1.1, 2.2, and 2.3 together, we obtain the main result of this paper.

THEOREM 2.4. *Let $X$ be a nonsingular irreducible complex projective variety of complex dimension $n$ with a cellular decomposition $\{C_i\}$. Then $Q^X$ is graded of rank $n$, rank-symmetric, rank-unimodal, and has property* S.

For future use, we record the following simple result. The proof is evident.

PROPOSITION 2.5. *Let $X$ and $Y$ be complex projective varieties, with cellular decompositions $\{C_i\}$ and $\{D_j\}$ respectively. Then the product variety $X \times Y$ has a cellular decomposition with cells $C_i \times D_j$, and $Q^{X \times Y} \cong Q^X \times Q^Y$.*

It follows from Theorem 2.4 and Proposition 2.5 that if $P = Q^X$ and $P' = Q^Y$ for nonsingular irreducible complex projective varieties $X$ and $Y$, each having a cellular

decomposition, then $P \times P'$ has property S. More generally, Canfield [7] and independently Proctor, Saks, and Sturtevant [36] have shown that the product $P \times P'$ of any two graded, rank-symmetric, rank-unimodal posets $P$ and $P'$, each with property S, also has property S. (An even more general result has subsequently been proved by Saks [37].) For our purposes, however, it suffices to consider only Proposition 2.5.

**3. Weyl groups.** It remains to find interesting examples of varieties $X$ with cellular decompositions and to describe the resulting posets $Q^X$. The best known examples of such varieties are the following. Let $G$ be a complex semisimple algebraic group, and let $P$ be a *parabolic subgroup* of $G$ (i.e., a closed subgroup which contains a maximal solvable subgroup $B$ of $G$. $B$ is known as a *Borel subgroup.*) Then the coset space $G/P$ has the structure of a non-singular irreducible complex projective variety, and the Bruhat decomposition of $G$ affords a cellular decomposition $\{C_i\}$ of $G/P$. The cells $C_i$ are known as *generalized Schubert cells.* See [5, § 3] for further details.

When $X = G/P$, a description of the poset $Q^X$ can be given in terms of the Weyl groups $W$ of $G$, and $W_J$ of $P$ [5, § 3], [11] as follows. Every Weyl group $W$ is a finite Coxeter group, i.e., $W$ is a finite group with a finite set $S = \{s_1, \ldots, s_m\}$ of generators such that for all $1 \leqq k \leqq m$, $1 \leqq i < j \leqq m$ and certain integers $n_{ij} \geqq 2$, $W$ is defined by the relations $s_k^2 = 1$ and $(s_i s_j)^{n_{ij}} = 1$. The pair $(W, S)$ is called a *Coxeter system.*

A *parabolic subgroup* of $W$ (with respect to $S$) is any subgroup $W_J$ generated by a subset $J$ of $S$. Thus $W_\phi = \{1\}$ and $W_S = W$. The *length* $\ell(w)$ of an element $w \in W$ is the smallest integer $q \geqq 0$ for which $w$ is a product of $q$ elements of $S$. Define a partial order, called the *Bruhat order,* on $W$ as follows. We say $w \leqq w'$ if there exist conjugates $t_1, \cdots, t_j$ of the elements of $S$ such that $w' = wt_1t_2 \cdots t_j$ and $\ell(wt_1t_2 \cdots t_{i+1}) > \ell(wt_1t_2 \cdots t_i)$ for all $0 \leqq i < j$. The following properties (among others) of the Bruhat order of a finite Coxeter group $W$ are known:

1. The Bruhat order makes $W$ into a graded poset (which we still call $W$).

2. The function $\ell$ is the rank function of $W$, and the rank-generating function of $W$ is given by

$$F(W, q) = \prod_{i=1}^{m} (1 + q + q^2 + \cdots + q^{e_i}) \tag{1}$$

for certain positive integers $e_i$ known as the *exponents* of $W$. One may regard (1) as the definition of the exponents. For other equivalent definitions, see, e.g., [6, Chap. 5, § 6.2] or [8, Chap. 10]. Note that (1) implies the well-known fact that $|W| = \prod (e_i + 1)$, and that $W$ has rank $e_1 + \cdots + e_m$.

3. If $J \subset S$, then each coset $wW_J$ of $W_J$ in $W$ contains a unique element $w_J$ of minimal length. For any $v \in W_J$ we have $\ell(w_J v) = \ell(w_J) + \ell(v)$.

4. Let $W^J$ be the set of minimal length coset representatives $w_J$. Then $W^J$ is a graded subposet of $W$ such that the rank function of $W^J$ is the restriction of the rank function of $W$.

5. $(W_J, J)$ is itself a finite Coxeter system, say with exponents $f_1, \cdots, f_t$. Then $W^J$ has the rank-generating function

$$F(W^J, q) = \frac{F(W, q)}{F(W_J, q)} = \frac{\prod_{i=1}^{m} (1 + q + q^2 + \cdots + q^{e_i})}{\prod_{j=1}^{t} (1 + q + q^2 + \cdots + q^{f_j})}. \tag{2}$$

For proofs of these results and further information on Coxeter groups, see e.g., [6], [8], [11]. For a connection between the posets $W^J$ and combinatorics, different from the one given here, see [23].

Now we return to the varieties $X = G/P$, where $G$ is a complex semisimple algebraic group and $P$ a parabolic subgroup of $G$. It is known [6, p. 29], [5, § 3] that the parabolic subgroups of $G$ containing a given Borel subgroup $B$ are in one-to-one

correspondence with the parabolic subgroups $W_J$ of the Weyl group $W$ of $G$ (with respect to a fixed set $S$ of Coxeter generators of $W$). Moreover, the poset $Q^X$ corresponding to the cellular decomposition of $X = G/P$ obtained from the Bruhat decomposition of $G$ is isomorphic to the partial order on $W^J$ defined above. Hence from Theorem 2.4 we conclude:

THEOREM 3.1. *Let $(W, S)$ be a Coxeter system for which $W$ is a Weyl group. Let $J \subset S$ and let $W^J$ be the poset defined above. Then $W^J$ is rank-symmetric, rank-unimodal, and has property* S.

A Coxeter system $(W, S)$ is *irreducible* if one cannot write $S$ as a nontrivial disjoint union $T \cup T'$ such that $W = W_T \times W_{T'}$. If $(W, S)$ is reducible, say $W = W_T \times W_{T'}$, then we also have $W = W_T \times W_{T'}$ as posets, and similarly for $W^J$. Thus by Proposition 2.5 nothing is lost by considering only irreducible Coxeter systems. Now all finite irreducible Coxeter systems are known (e.g., [6, p. 193]). There are the infinite families of type $A_n (n \geqq 1)$, $B_n (n \geqq 2)$, and $D_n (n \geqq 4)$, together with seven "exceptional" systems $E_6$, $E_7$, $E_8$, $F_4$, $G_2$, $H_3$, $H_4$ and the dihedral groups $I_2(p)$ of order $2p$ for $p = 5$ or $p \geqq 7$. ($I_2(3)$ coincides with $A_2$, $I_2(4)$ with $B_2$, and $I_2(6)$ with $G_2$.) For all of these systems $(W, S)$, $W$ is a Weyl group except for $H_3$, $H_4$, $I_2(p)$, $p = 5$ or $p \geqq 7$. It is easy to check that Theorem 3.1 remains valid for the dihedral groups $I_2(p)$, and for $H_3$. Presumably the remaining case $H_4$ can also be checked directly, so in fact one could determine those finite Coxeter systems (probably all of them) for which Theorem 3.1 remains valid.

**4. Type $A_n$.** We now want to describe the posets $W^J$ in greater detail, for the types $A_n$, $B_n$, $D_n$. First consider $A_{n-1}$. Then $W$ is the symmetric group $\mathfrak{S}_n$ of all permutations of $\{1, 2, \cdots, n\}$. The exponents are $1, 2, \cdots, n-1$, and as Coxeter generators we may take the "adjacent transpositions" $s_i = (i, i+1)$, $1 \leqq i \leqq n-1$. Regard a permutation $\pi \in \mathfrak{S}_n$ as a linear array $a_1 a_2 \cdots a_n$, where $\pi(i) = a_i$. Then a direct translation of the definition of the Bruhat order yields the following: $\pi \leqq \sigma$ in $W$ if $\sigma$ can be obtained from $\pi$ by a sequence of operations which interchange $i$ and $j$ in a permutation $a_1 a_2 \cdots a_n$ provided $i$ appears to the left of $j$ and $i < j$. We abbreviate this operation as

$$i < j \longrightarrow j > i. \tag{3}$$

Thus the notation "$i < j$" in (3) means that $i$ and $j$ appear in the given order (i.e., $i$ to the left of $j$) and $i < j$. For instance, $213 \leqq 312$ (obtained by $2 < 3 \longrightarrow 3 > 2$) and $24153 \leqq 35241$ (obtained, e.g., by $2 < 3 \to 3 < 2$, $1 < 2 \to 2 > 1$, $4 < 5 \to 5 > 4$). The rank $\ell(\pi)$ of $\pi = a_1 a_2 \cdots a_n \in W$ is equal to the number $i(\pi)$ of *inversions* of $\pi$, i.e., the number of pairs $(i, j)$ for which $i < j$ and $a_i > a_j$. Thus $12 \cdots n$ is the unique permutation of rank 0 and $n \cdots 21$ is the unique permutation of highest rank $\binom{n}{2}$. It is well-known (e.g., [9, § 6.4]) that

$$\sum_{\pi \in \mathfrak{S}_n} q^{i(\pi)} = (1+q)(1+q+q^2) \cdots (1+q+\cdots+q^{n-1}),$$

which of course agrees with (1). Figure 1 depicts the Bruhat order of $\mathfrak{S}_3$.

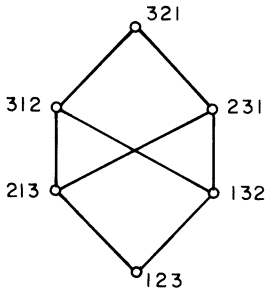



FIG. 1

Now let $J \subset S = \{s_1, \cdots, s_{n-1}\}$ where $s_i = (i, i+1)$. If we let $\mathfrak{S}(a, b)$ denote the group of all permutations of $\{a, a+1, \cdots, b\}$, then it is clear that $W_J = \mathfrak{S}(1, c_1) \times \mathfrak{S}(c_1+1, c_2) \times \cdots \times \mathfrak{S}(c_{j-1}+1, n)$ for some integers $1 \leqq c_1 < c_2 < \cdots < c_{j-1} < n$, where $j = n - |J|$. If $\pi = a_1 a_2 \cdots a_n \in W$, then the coset $\pi W_J$ consists of all $c_1!(c_2 - c_1)! \cdots (n - c_{j-1})!$ permutations obtained from $\pi$ by permuting among themselves the elements within the sets $N_1 = \{1, 2, \cdots, c_1\}$, $N_2 = \{c_1+1, \cdots, c_2\}$, $\cdots$, $N_j = \{c_{j-1}+1, \cdots, n\}$. The coset representative $\pi_J \in \pi W_J$ with the least number of inversions is that element of $\pi W_J$ for which the elements of the above sets $N_i$ appear in their natural order. Hence $W^J$ consists of those $n!/c_1!(c_2 - c_1)! \cdots (n - c_{j-1})!$ permutations for which the elements of each of the sets $N_i$ appear in their natural order; or, as it is sometimes called, the set of *shuffles* of $N_1, \cdots, N_j$. The rank-generating function of $W^J$ is given by

$$F(W^J, q) = \frac{\mathbf{(n)!}}{\mathbf{(c_1)!(c_2 - c_1)! \cdots (n - c_{j-1})!}} \tag{4}$$

where $\mathbf{(k)!} = (1-q)(1-q^2) \cdots (1-q^k)$. The right-hand side of (4) is known as a *q-multinomial coefficient* and is commonly denoted $\begin{bmatrix} n \\ c_1, c_2 - c_1, \cdots, n - c_{j-1} \end{bmatrix}$. Figure 2 illustrates the poset $W^J$ in the case $n = 4$, $J = \{(12)\}$.

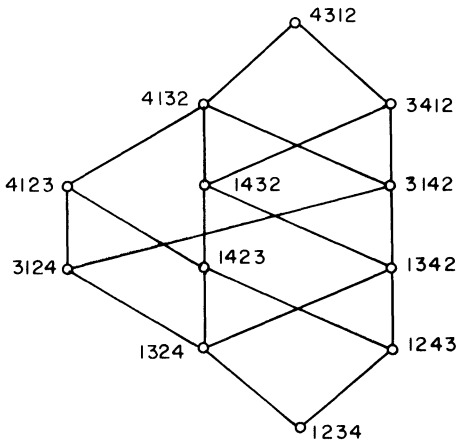



FIG. 2

If we take $W_J$ to be a *maximal* parabolic subgroup above, i.e., $|J| = n-2$, then the poset $W^J$ has an interesting alternative description. Suppose $J = S - \{(n-k, n-k+1)\}$, so $N_1 = \{1, 2, \cdots, n-k\}$ and $N_2 = \{n-k+1, \cdots, n\}$. If $\pi = a_1 a_2 \cdots a_n \in W^J$ and $1 \leqq i \leqq k$, then set

$$\ell_i(\pi) = \text{card}\{j : j \text{ appears to the right of } n-i+1 \text{ and } j < n-i+1\}. \tag{5}$$

Clearly $\ell(\pi) = \sum_{i=1}^{k} \ell_i(\pi)$. The mapping $\pi \mapsto (\ell_1(\pi), \cdots, \ell_k(\pi))$ is a bijection between $W^J$ and all integer sequences $0 \leqq \ell_1 \leqq \cdots \leqq \ell_k \leqq n-k$. Moreover, $\pi \leqq \pi'$ in $W^J$ if and only if $\ell_i \leqq \ell_i'$ for $1 \leqq i \leqq k$. Hence, $W^J$ is isomorphic to the poset of all partitions of integers into at most $k$ parts, with largest part at most $n-k$, i.e., a partition whose Ferrers diagram (e.g., [9, § 2.4]) fits into a $k \times (n-k)$ rectangle. These partitions are ordered by inclusion of their Ferrers diagrams. Since the union and intersection of Ferrers diagrams is again a Ferrers diagram, it follows that the poset $W^J$ is actually a distributive lattice, which we will denote by $L(k, n-k)$. Figure 3 depicts $L(2, 3)$.

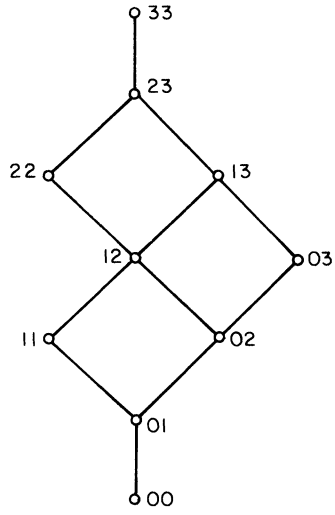



FIG. 3

In terms of the characterization [3, Thm. 3, p. 46] of a finite distributive lattice $L$ as the lattice $\mathbf{2}^P$ of semi-ideals (also called "order ideals" or "decreasing subsets") of a poset $P$, we have $L(k, n-k) = \mathbf{2}^{\mathbf{k}\times(\mathbf{n}-\mathbf{k})}$, where $\mathbf{i}$ denotes an $i$-element chain. The rank-generating function of this lattice is the $q$-binomial coefficient $\begin{bmatrix} n \\ k \end{bmatrix} = \mathbf{(n)!/(k)!(n-k)!}$. It is by no means a priori obvious that $W^J$ is rank-unimodal; this was first shown essentially by Sylvester in 1878 (see [40] for historical details) and no combinatorial proof is known. I am grateful to Tony Iarrobino for originally calling to my attention that the hard Lefschetz theorem implies the unimodality of the coefficients of $\begin{bmatrix} n \\ k \end{bmatrix}$. It was my attempt to understand this fact which eventually led to the present paper.

By applying Theorem 3.1 to the lattice $L(k, m)$, we can deduce a "multiset analogue" to a conjecture of Erdös and Moser [13, (12)]. (Regarding their actual conjecture, see Corollary 5.3 below.) I am grateful to Ranee Gupta for her comments on this result.

COROLLARY 4.1. *Fix positive integers $k$, $m$, and $j$. Let $A = \{a_0, a_1, \cdots, a_m\}$ be a set of $m+1$ distinct real numbers. Let $B_1, \cdots, B_r$ be subsets of $A$ with exactly $k$ elements* ***with repeated elements allowed.*** *(One may think of $B_s$ as being an $m+1$-tuple $(\alpha_0, \cdots, \alpha_m)$ of nonnegative integers such that $\sum \alpha_i = k$, where $\alpha_i$ is the number of repetitions of $a_i$.) Let $\sum B_s$ denote the sum of the elements of $B_s$, i.e., $\sum B_s = \sum \alpha_i a_i$. Suppose that there are at most $j$ distinct numbers among $\sum B_1, \cdots, \sum B_r$. Then $r$ is less than or equal to the sum of the $j$ middle coefficients of the polynomial $\begin{bmatrix} m+k \\ k \end{bmatrix}$. Moreover, this value of $r$ is achieved by taking $A = \{0, 1, \cdots, m\}$ and $B_1, \cdots, B_r$ to have element sums consisting of the $j$ middle elements of the set $\{0, 1, \cdots, km\}$. (If $km - j$ is even, then there are two equivalent choices of the "$j$ middle coefficients" and "$j$ middle elements.")*

*Proof.* Regarding $B_s = (\alpha_0, \cdots, \alpha_m)$ associate with $B_s$ the sequence $\lambda_s = (\ell_1, \cdots, \ell_k) \in L(k, m)$ defined by setting exactly $\alpha_i$ of the $\ell_k$'s equal to $i$. It is easy to see that the subset $\{\lambda_1, \cdots, \lambda_r\}$ of $L(k, m)$ contains no $(j+1)$-element chain provided there are only $j$ distinct numbers among $\sum B_1, \cdots, \sum B_r$. The proof now follows from Theorem 3.1 and the fact that the rank-generating function of $L(k, m)$ is $\begin{bmatrix} k+m \\ k \end{bmatrix}$. □

As a variation of the preceding corollary, we have

COROLLARY 4.2. *Fix positive integers $k$, $m$, and $j$. Let $A' = \{a_1, \cdots, a_m\}$ be a set of $m$ distinct nonzero real numbers. Let $B_1, \cdots, B_r$ be subsets of $A'$ with* ***at most*** *$k$ elements with repeated elements allowed. Suppose that there are at most $j$ distinct numbers among $\sum B_1, \cdots, \sum B_r$. Then $r$ is less than or equal to the sum of the $j$ middle coefficients of the polynomial $\begin{bmatrix} m+k \\ k \end{bmatrix}$. Moreover, this value of $r$ is achieved by taking $A' = \{1, \cdots, m\}$ and $B_1, \cdots, B_r$ to have element sums consisting of the $j$ middle elements of the set $\{0, 1, \cdots, km\}$.*

*Proof.* Apply Corollary 4.1 to the set $A = A' \cup \{0\}$. □

*Remark.* The cellular decomposition of $G/P$ in the case $W(G) = \mathfrak{S}_n$ and $W(P) = \mathfrak{S}_k \times \mathfrak{S}_{n-k}$ can be described quite concretely. The group $G$ is given by $SL(n, \mathbb{C})$, which acts linearly on $n$-dimensional complex projective space $\mathbb{P}^{n-1}$. Let $V$ be a $(k-1)$-dimensional subspace (or $(k-1)$-plane) of $\mathbb{P}^{n-1}$, and let $P$ be the subgroup of $G$ leaving $V$ invariant. (Then $P$ is a maximal parabolic subgroup of $G$.) The coset $\phi P$ transforms $V$ into the subspace $\phi V$, and this sets up a one-to-one correspondence between $X = G/P$ and the $(k-1)$-planes in $\mathbb{P}^{n-1}$. Hence $X$ is the *Grassmann manifold* $G(k-1, n-1)$ of all $(k-1)$-planes in $\mathbb{P}^{n-1}$. Regard the elements of $\mathbb{P}^{n-1}$ as (equivalence classes of) $n$-tuples $(x_1, \cdots, x_n) \in \mathbb{C}^n - \{0\}$. A $(k-1)$-plane $V$ in $\mathbb{P}^{n-1}$ has a unique ordered basis $w_1, \cdots, w_k$ for which the matrix $\begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}$ is in row-reduced echelon form. Choose integers $0 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_k \leqq n-k$, and suppose we specify that for each $i$, the first 1 in $w_i$ occurs in coordinate $a_i + i$. The set of all such $V$ forms a subset $C(a_1, \cdots, a_k)$ of $G(k-1, n-1)$ isomorphic to $\mathbb{C}^{k(n-k)-a_1-\cdots-a_k}$; indeed, there are $n-k-a_i$ coordinates in $w_i$ which can be specified arbitrarily, and the remaining coordinates are predetermined. By considering all sequences $0 \leqq a_1 \leqq \cdots \leqq a_k \leqq n-k$, we obtain a cellular decomposition of $G(k-1, n-1)$. Thus the cells $C(a_1, \cdots, a_k)$ are in one-to-one correspondence with the elements $(a_1, \cdots, a_k)$ of $L(k, n-k)$. For instance, when $k = 2$ and $n = 4$ the cells correspond to the following row-reduced echelon matrices:

$$\underset{C(0,0)}{\begin{bmatrix} 1 & 0 & * & * \\ 0 & 1 & * & * \end{bmatrix}}, \underset{C(0,1)}{\begin{bmatrix} 1 & * & 0 & * \\ 0 & 0 & 1 & * \end{bmatrix}}, \underset{C(0,2)}{\begin{bmatrix} 1 & * & * & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}},$$

$$\underset{C(1,1)}{\begin{bmatrix} 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \end{bmatrix}}, \underset{C(1,2)}{\begin{bmatrix} 0 & 1 & * & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}, \underset{C(2,2)}{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}.$$

A little thought shows that $\overline{C(a_1, \cdots, a_k)} \supset C(b_1, \cdots, b_k)$ if and only if $a_i \leqq b_i$ for $1 \leqq i \leqq k$. Thus we see directly that $Q^X \cong L(k, n-k)$. The closure of the cell $C(a_1, \cdots, a_k)$ is called a *Schubert variety*, and its cohomology class is called a *Schubert cycle*, which we shall denote by $\Omega(a_1, \cdots, a_k)$. (A more common notation is $\Omega(a'_1, \cdots, a'_k)$ where $a'_i = n-k+i-1-a_{k-i+1}$.) The Schubert cycle $\omega = \Omega(0, 0, \cdots, 0, 1) \in H^2(X, \mathbb{C})$ turns out to be the class of a hyperplane section. According to a special case of Pieri's formula in the Schubert calculus, the product of $\Omega(a_1, \cdots, a_k)$ with $\omega$ in $H^*(X, \mathbb{C})$ is equal to the sum of all $\Omega(b_1, \cdots, b_k)$ such that $b_i \geqq a_i$ and $\sum b_i = 1 + \sum a_i$. In other words, $\omega \cdot \Omega(a_1, \cdots, a_k) = \sum \Omega(b_1, \cdots, b_k)$, where the sum is over all sequences $(b_1, \cdots, b_k)$ covering $(a_1, \cdots, a_k)$ in $L(k, n-k)$. Thus we

have a direct verification of Lemma 2.2. For further information on these matters, see, for example, [26], [27], [41].

**5. Type $B_n$.** We next turn our attention to type $B_n$. In this case $W$ is the group of all $n \times n$ signed permutation matrices (i.e., matrices with entries $0, \pm 1$ with one nonzero entry in every row and column). $W$ has order $2^n n!$ and exponents $1, 3, 5, \cdots, 2n-1$. Identify the matrix $(m_{ij}) \in W$ with the ordered pair $(\pi, \varepsilon)$, where $\pi \in \mathfrak{S}_n$ is given by $m_{i,\pi(i)} = \pm 1$ and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \in \{\pm 1\}^n$ by $\varepsilon_i = m_{i,\pi(i)}$. We then have the multiplication rule $(\pi, \varepsilon)(\pi', \varepsilon') = (\pi\pi', \delta)$, where $\delta_i = \varepsilon_{\pi'(i)}\varepsilon'_i$. We sometimes will abbreviate a group element such as $(24513, (-1, 1, -1, -1, 1))$ by $\bar{2}\ 4\ \bar{5}\ \bar{1}\ 3$, and thus regard $W$ as consisting of all "barred permutations" of $\{1, 2, \cdots, n\}$. For the Coxeter generators of $W$ we take the set $S = \{s_1, \cdots, s_n\}$, where $s_i$ is the adjacent transposition $(i, i+1)$, $1 \leqq i \leqq n-1$, and $s_n = \bar{1}\ 2\ 3 \cdots n$. A little thought shows that $\pi \leqq \sigma$ in $W$ if $\sigma$ can be obtained from $\pi$ by a sequence of the following seven types of operations on barred permutations:

a) $i \longrightarrow \bar{i}$,
b) $i < j \longrightarrow j > i$,
c) $\bar{i} < j \longrightarrow \bar{j} > \bar{i}$,
d) $\bar{i} < j \longrightarrow \bar{j} > i$,
e) $\bar{i} > j \longrightarrow j < \bar{i}$,
f) $i > \bar{j} \longrightarrow j < \bar{i}$,
g) $\bar{i} > \bar{j} \longrightarrow \bar{j} < \bar{i}$.

For instance, Fig. 4 illustrates $W$ when $n = 2$.

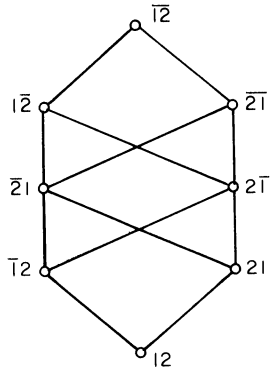



FIG. 4

If $(\pi, \varepsilon) \in W$, then one can check that

$$\ell(\pi) = i(\pi) + \sum_j (2d_j + 1), \tag{6}$$

where $i(\pi)$ is the number of inversions of $\pi$, $j$ ranges over all integers for which $\varepsilon_j = -1$, and $d_j$ is the number of $k$'s appearing in $\pi = a_1 a_2 \cdots a_n$ to the left of $a_j$ for which $k < a_j$. For instance, $\ell(\bar{3}\ 1\ 5\ \bar{4}\ 2) = 11$, since $i(\pi) = 5$, $d_1 = 0$, $d_4 = 2$. It is easy to give a direct combinatorial proof that

$$\sum_{\pi \in W} q^{\ell(\pi)} = \prod_{i=1}^{n} (1 + q + q^2 + \cdots + q^{2i-1}),$$

agreeing with (1).

Now let $J \subset S$. Let $\bar{\mathfrak{S}}(a, b)$ denote the group of all signed permutations of $\{a, a+1, \cdots, b\}$. Then $W_J$ has the form

$$W_J = \bar{\mathfrak{S}}(1, c_1) \times \mathfrak{S}(c_1+1, c_2) \times \mathfrak{S}(c_2+1, c_3) \times \cdots \times \mathfrak{S}(c_{j-1}+1, n), \tag{7}$$

where $0 \leqq c_1 < c_2 < \cdots < c_{j-1} < n$. The case $c_1 = 0$ corresponds to $s_n \notin J$. If $c_1 = 0$ then $j = n - |J|$; otherwise $j = n - |J| + 1$. Set $N_1 = \{1, 2, \cdots, c_1\}$, $N_2 = \{c_1+1, \cdots, c_2\}, \cdots, N_j = \{c_{j-1}+1, \cdots, n\}$. One can check that $W^J$ consists of all $(a_1 a_2 \cdots a_n, \varepsilon) \in W$ satisfying:

(i) $\varepsilon_i = 1$ if $a_i \in N_1$.

(ii) If $a_r, a_s \in N_i$ with $r < s$ and $\varepsilon_r = \varepsilon_s = 1$, then $a_r < a_s$.

(iii) If $a_r, a_s \in N_i$ with $r < s$ and $\varepsilon_r = \varepsilon_s = -1$, then $a_r > a_s$.

(iv) If $a_r, a_s \in N_i$ and $\varepsilon_r = 1$, $\varepsilon_s = -1$, then $a_r > a_s$.

For instance, if $W_J = \bar{\mathfrak{S}}(1, 2) \times \mathfrak{S}(3, 7) \times \mathfrak{S}(8, 9)$, then a typical element of $W^J$ is $5\ \bar{4}\ 1\ \bar{8}\ 6\ 2\ 7\ 9\ \bar{3}$. The letters 1, 2 are unbarred and appear in increasing order. Similarly 3, 4 are barred and decrease, 5, 6, 7 are unbarred and increase, 8 is barred and "decreases," and 9 is unbarred and "increases."

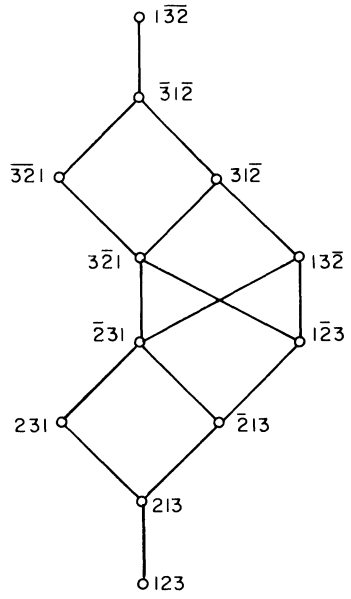



FIG. 5

Figure 5 illustrates $W^J$ when $n = 3$ and $J = \{s_1, s_3\}$. We see that, unlike the situation for $A_n$, $W^J$ need not be a distributive lattice (or even just a lattice) when $J$ is a maximal subset of $S$. There is one case, however, in which $W^J$ is a distributive lattice, viz., $J = \{s_1, s_2, \cdots, s_{n-1}\}$, so $W_J = \bar{\mathfrak{S}}(1, n)$. In this case we will denote $W^J$ by $M(n)$. To see that $M(n)$ is indeed a distributive lattice, observe that for every sequence $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \in \{\pm 1\}^n$, there is a unique $\pi \in \mathfrak{S}_n$ for which $(\pi, \varepsilon) \in M(n)$. Identify $\varepsilon$ with the subset of $\{1, 2, \cdots, n\}$ consisting of those integers $t$ for which $\varepsilon_t = -1$. Then the partial order on $M(n)$ is given by $\{a_1, \cdots, a_j\} \leqq \{b_1, \cdots, b_k\}$ if $a_1 < \cdots < a_j$, $b_1 < \cdots < b_k$, $j \leqq k$, and $a_{j-i} \leqq b_{k-i}$ for $0 \leqq i \leqq j-1$. It is then easily seen that $M(n)$ is a distributive lattice. The poset $P$ for which $M(n) = \mathbf{2}^P$ is given by $P = \mathbf{2}^{\mathbf{2} \times (\mathbf{n}-\mathbf{1})}$. Figure 6 illustrates $M(4)$.

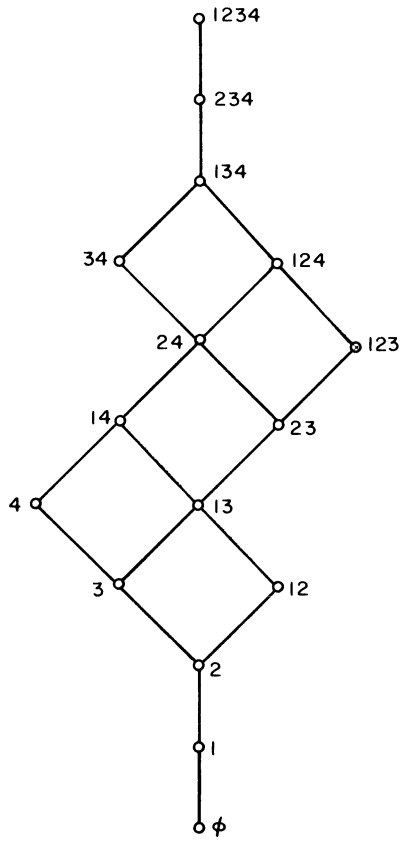



FIG. 6

Lindström [30] conjectured that $M(n)$ has property $S_1$, while in fact we now know that $M(n)$ has property $S$ and is rank-unimodal. (I am grateful to Larry Harper for calling my attention to Lindström's conjecture.) The rank-generating function of $M(n)$ is $(1+q)(1+q^2)\cdots(1+q^n)$. The unimodality of the coefficients of this polynomial was first explicitly proved by Hughes [25], based on a result of Dynkin (see [40] for further information). Presumably, however, this result could also be proved analytically using the methods of [12]. Lindström [30], [31] shows that the structure of $M(n)$ is related to a conjecture [13, (12)] of Erdös and Moser (see also [12], [38], [42]). In fact, Corollary 5.3 below provides a more general result. I am grateful to Ranee Gupta for pointing out an error in my original treatment of the Erdös–Moser conjecture.

COROLLARY 5.1. *Let $A$ be a set of distinct real numbers. Assume that $\nu$ elements of $A$ are negative, $\zeta$ are equal to 0 (so $\zeta = 0$ or 1), and $\pi$ are positive. Let $B_1, \cdots, B_r$ be subsets of $A$ whose element sums take on at most $k$ distinct values. Then $r$ does not exceed the sum of the $k$ middle coefficients of the polynomial*

$$G_{\nu\zeta\pi}(q) = 2^{\zeta}(1+q)(1+q^2)\cdots(1+q^{\nu})\cdot(1+q)(1+q^2)\cdots(1+q^{\pi})$$

*(there being two equivalent choices of the "$k$ middle coefficients" when $\binom{\nu+1}{2}+\binom{\pi+1}{2}-k$ is even). Moreover, this value of $r$ is achieved by taking $A = \{-1, -2, \cdots, -\nu\}\cup\{1, 2, \cdots, \pi\}\cup Z$, where $Z = \phi$ or $\{0\}$ depending on whether $\zeta = 0$ or 1.*

*Proof.* Since 0 can be adjoined to a set without affecting its element sum we may assume $\zeta = 0$. Let $M(\nu)^*$ denote the order-dual of $M(\nu)$.(The elements of $M(\nu)$ and $M(\nu)^*$ coincide, but $C \leqq C'$ in $M(\nu)^*$ if and only if $C \geqq C'$ in $M(\nu)$.) Regard elements of the product $M(\nu)^* \times M(\pi)$ as consisting of pairs $(C, D)$, where $C$ is a subset of $\{1, 2, \cdots, \nu\}$, and $D$ is a subset of $\{1, 2, \cdots, \pi\}$. Suppose that the elements of $A$ are $\alpha_\nu < \cdots < \alpha_1 < 0 < \beta_1 < \cdots < \beta_\pi$ and that $B_s = \{\alpha_{i_1}, \cdots, \alpha_{i_h}, \beta_{j_1}, \cdots, \beta_{j_m}\}$. Associate with $B_s$ the set $(C_s, D_s) = (\{i_1, \cdots, i_h\}, \{j_1, \cdots, j_m\}) \in M(\nu)^* \times M(\pi)$. It is easy to see that the subset $\{(C_1, D_1), \cdots, (C_r, D_r)\}$ of $M(\nu)^* \times M(\pi)$ contains no $(k+1)$-element chain provided there are most $k$ distinct element sums of $B_1, \cdots, B_r$. Now it is not difficult to see that $M(\nu)^* \cong M(\nu)$. (For instance, given the set $T = \{i_1, \cdots, i_h\} \in M(\nu)$ with $1 \leqq i_1 < \cdots < i_h \leqq \nu$, define $T^*$ to be the set of nonzero parts of the partition $\lambda$ which is conjugate (in the sense of [9, p. 100]) to the partition whose parts are $\nu - i_h, \nu - 1 - i_{h-1}, \cdots, \nu - h + 1 - i_1, \nu - h, \nu - h - 1, \cdots, 1$. Then the mapping $T \to T^*$ is an isomorphism $M(\nu) \to M(\nu)^*$. See also § 7 for a more general result.) The proof now follows from Theorem 3.1 and Proposition 2.5 (or from Theorem 3.1 alone applied to the appropriate *reducible* Weyl group) and the fact that the rank-generating function of $M(\nu)^* \times M(\pi)$ is $G_{\nu 0 \pi}(q)$. □

We now want to consider the situation where $\nu + \zeta + \pi$ is fixed, but $\nu$, $\zeta$, and $\pi$ can vary. First we need:

LEMMA 5.2. *Let $G(q)$ be a polynomial of degree $d$ with symmetric unimodal coefficients. Fix positive integers $j$ and $k$. Then the sum of the middle $k$ coefficients of $G(q)(1+q^{j+1})$ does not exceed the sum of the middle $k$ coefficients of $G(q)(1+q^j)$.*

*Proof.* Let $G(q) = \alpha(0) + \alpha(1)q + \cdots + \alpha(d)q^d$. For simplicity of notation we assume $d = 2d'$, $j = 2j'$, $k = 2k'$. The other cases are done similarly. The middle $k$ coefficients of $G(q)(1+q^j)$ are

$$\alpha(d' + j' - k' + i) + \alpha(d' - j' - k' + i), \qquad 0 \leqq i \leqq k-1.$$

The middle $k$ coefficients of $G(q)(1+q^{j+1})$ are

$$\alpha(d' + j' - k' + i + 1) + \alpha(d' - j' - k' + i), \qquad 0 \leqq i \leqq k-1.$$

(Here we set $\alpha(t) = 0$ if $t < 0$.) If $\Omega$ applied to a polynomial denotes the sum of its middle $k$ coefficients, then

$$\Omega G(q)(1+q^j) - \Omega G(q)(1+q^{j+1}) = \alpha(d' + j' - k') - \alpha(d' + j' + k').$$

Since $\alpha(i) = \alpha(d - i)$ and $\alpha(0) \leqq \alpha(1) \leqq \cdots \leqq \alpha(d')$, it follows that $\alpha(d' + j' - k') \geqq \alpha(d' + j' + k')$, completing the proof. □

COROLLARY 5.3. *Let $A$ be a set of $n$ distinct real numbers, and let $B_1, \cdots, B_r$ be subsets of $A$ whose element sums take on at most $k$ distinct values. Let $\nu = [(n-1)/2]$ and $\pi = [n/2]$. Then $r$ does not exceed the sum of the $k$ middle coefficients of the polynomial*

$$2(1+q)(1+q^2)\cdots(1+q^\nu)\cdot(1+q)(1+q^2)\cdots(1+q^\pi).$$

*Moreover, this value of $r$ is achieved by choosing $A = \{-\nu, -\nu+1, \cdots, \pi\}$.*

*Proof.* For fixed $n = \nu + \zeta + \pi$, it follows from Lemma 5.2 that the sum of the middle $k$ coefficients of $G_{\nu\zeta\pi}(q)$ is maximized by choosing $\zeta = 1$, $\nu = [(n-1)/2]$, $\pi = [n/2]$. The proof follows from Corollary 5.1. □

The actual conjecture [13, (12)] of Erdös and Moser is equivalent to the case $k = 1$, and $n$ odd, of Corollary 5.3. A purely combinatorial derivation of the Erdös–Moser conjecture from the fact that $M(n)$ has property S appears in [35].

**6. Type $D_n$.** If $(W, S)$ is a Coxeter system of type $D_n$, then $W$ is the subgroup of the group $W'$ of type $B_n$ consisting of all $(\pi, \varepsilon)$ such that $\prod_{i=1}^{n} \varepsilon_i = +1$. $W$ has order $2^{n-1}n!$ and exponents $1, 3, 5, \cdots, 2n-5, 2n-3, n-1$. We may take $S = \{s_1, \cdots, s_n\}$ where $s_i = (i, i+1)$ if $1 \leqq i \leqq n-1$ (as in type $B_n$) and $s_n = \bar{2}\ \bar{1}\ 3\ 4 \cdots n$. We then have the following seven transformation rules for obtaining $w'$ from $w$ when $w \leqq w'$ in $W$:

a) $i < j \longrightarrow \bar{j} > \bar{i}$,
b) $i < j \longrightarrow j > i$,
c) $i < j \longrightarrow j > \bar{i}$,
d) $\bar{i} < j \longrightarrow \bar{j} > i$,
e) $\bar{i} > j \longrightarrow j < \bar{i}$,
f) $i > \bar{j} \longrightarrow j < \bar{i}$,
g) $\bar{i} > \bar{j} \longrightarrow \bar{j} < \bar{i}$.

Note that rules b–g coincide with those for $B_n$, and that rule a for $D_n$ is obtained by applying rule b and rule a twice for $B_n$. It follows that if $\pi \leqq \sigma$ in $W$ then $\pi \leqq \sigma$ in $W'$. The converse, however, is false. For instance, $21 < \bar{2}\bar{1}$ in $W'$ but $21$ and $\bar{2}\bar{1}$ are incomparable in $W$. Figure 7 depicts $W$ when $n = 2$.

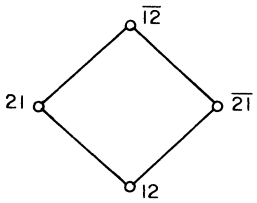



FIG. 7

If $(\pi, \varepsilon) \in W$, then

$$\ell(\pi) = i(\pi) + 2\sum_j d_j,$$

where $i(\pi)$ and $d_j$ have the same meaning as in (6). For instance, $\ell(\bar{3}\ 1\ 5\ \bar{4}\ 2) = 9$ for $D_5$, while $\ell(\bar{3}\ 1\ 5\ \bar{4}\ 2) = 11$ for $B_5$.

Now let $J \subset S$. In so far as describing the poset $W^J$ is concerned, we may assume that if $s_n = \bar{2}\bar{1}34 \cdots n \in J$ then also $s_1 = 213 \cdots n \in J$, since interchanging $s_1$ and $s_n$ induces an automorphism of the Coxeter system $(W, S)$. Thus if we let $\hat{\mathfrak{S}}(a, b)$ denote the group of all signed permutations of $\{a, a+1, \cdots, b\}$ with an even number of $-1$'s, then $W_J$ has the form

$$W_J = \hat{\mathfrak{S}}(1, c_1) \times \mathfrak{S}(c_1+1, c_2) \times \cdots \times \mathfrak{S}(c_{j-1}+1, n),$$

where $0 \leqq c_1 < c_2 < \cdots < c_{j-1} < n$ and $c_1 \neq 1$. The case $c_1 = 0$ corresponds to $s_n \notin J$. Defining $N_1 = \{1, 2, \cdots, c_1\}$, $N_2 = \{c_1+1, \cdots, c_2\}$, $\cdots$, $N_i = \{c_{i-1}+1, \cdots, n\}$, one can check that $W^J$ consists of all $(a_1 a_2 \cdots a_n, \varepsilon) \in W$ satisfying:

(i) $\varepsilon_1 = 1$ if $a_i \in N_1$ and $a_i > 1$.
(ii)–(iv) Same as for type $B_n$.
(v) 1 precedes every other element of $N_1$ (even if 1 is barred).

For instance, Fig. 8 depicts $W^J$ when $n = 3$ and $J = \{12\}$, i.e., $W_J = \hat{\mathfrak{S}}(1, 2) \times \mathfrak{S}(3, 3)$, so $N_1 = \phi$, $N_2 = \{1, 2\}$, $N_3 = \{3\}$. Note that this poset is isomorphic to that of Fig. 2; this is no accident since Coxeter systems of types $A_3$ and $D_3$ are isomorphic. (Recall that to obtain nonisomorphic systems, one may take $A_n$ for $n \geqq 1$, $B_n$ for $n \geqq 2$, and $D_n$ for $n \geqq 4$.)

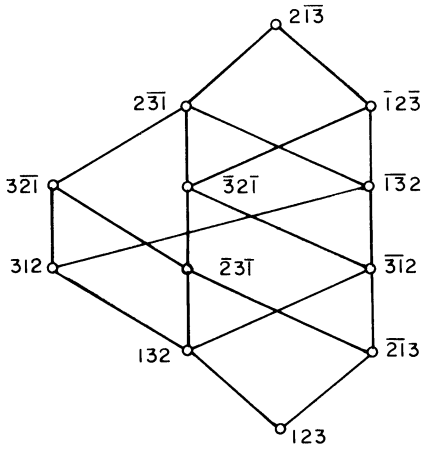



FIG. 8

As in the case of $B_n$, $W^J$ need not be a distributive lattice when $J$ is maximal. For instance, take $n=4$ and $J=\{s_1, s_3, s_4\}=S-\{(23)\}$, so $W_J=\mathfrak{S}(1,2)\times\mathfrak{S}(3,4)$. Then the rank-generating function of $W^J$ is given by

$$F(W^J, q)=1+q+3q^2+3q^3+4q^4+4q^5+3q^6+3q^7+q^8+q^9,$$

and it is easy to check that there does not exist a distributive lattice with this rank-generating function. As in the situation for $B_n$, there is one special case for which $W^J$ *is* a distributive lattice. Take $J=\{s_1, s_2, \cdots, s_{n-1}\}$, so $W_J=\mathfrak{S}(1, n)$. If we regard $M(n)$ (as defined in the previous section) as consisting of all subsets of $\{1, 2, \cdots, n\}$, then $W^J$ turns out to be the subposet of $M(n)$ consisting of all sets of even cardinality. But it is easily seen that this subposet is isomorphic to $M(n-1)$, so nothing new is obtained.

**7. Final comments.** In view of the examples $L(m, n)$ and $M(n)$, it is natural to ask under what circumstances is $W^J$ a distributive lattice. I am grateful to Robert Proctor for supplying the following answer to this question. The Coxeter generators $S$ of an irreducible Weyl group $W$ correspond to the *fundamental representations* $\lambda_i (1\leqq i\leqq n)$ of a certain complex simple Lie algebra $\mathfrak{g}$. By direct computation facilitated by representation theory, Proctor has shown that (except for the representations $\lambda_1$ and $\lambda_2$ of $G_2$) $W^J$ is distributive if and only if the irreducible representation of $\mathfrak{g}$ with highest weight $\sum_{i\notin J}\lambda_i$ is *miniscule*, as defined in [6, p. 226]. These representations have special significance in other contexts; see [39] and more generally [28]. It turns out that for all the distributive $W^J$'s except $L(m, n)$ and $M(n)$, it is easy to check Property $S$ directly.

Proctor has also shown that if $W$ is a Weyl group with largest element $v$ (in the Bruhat order) and if $W^J$ (for any $J\subset S$) has largest element $y$, then the bijection from $W^J$ to $W^J$ given by $w\to vwy^{-1}v^{-1}$ is an anti-automorphism of $W^J$. Thus $W^J$ is self-dual whenever $W$ is a Weyl group. We do not know whether the more general posets $Q^X$ of Theorem 2.4 need always be self-dual.

We conclude with an open problem. Let $P$ be a finite graded rank-symmetric poset of rank $n$, with rank function $\rho$. $P$ is called a *symmetric chain order* (e.g., [17, §3], [20], [21]) if it can be partitioned into pairwise disjoint saturated chains $x_i<x_{i+1}<\cdots<x_{n-i}$ such that $\rho(x_j)=j$. It is easy to see that a symmetric chain order satisfies Property $T$ and hence is rank-unimodal. Easy examples show that a rank-symmetric poset satisfying Property $T$ need not be a symmetric chain order.

Our open problem is the following: Are all the posets $Q^X$ of Theorem 2.4 (or at least the special cases $W^J$ of Theorem 3.1) symmetric chain orders? Since any poset $Q^X$ given by Theorem 2.4 has property $T$, there are pairwise disjoint chains connecting all of $Q_i^X$ to $Q_{i+1}^X$ when $i < n/2$, and all of $Q_i^X$ to $Q_{i-1}^X$ when $i > n/2$. Piecing together these chains yields a partition of $Q^X$ into saturated chains all of which pass through the middle rank (when $n$ is even) or middle two ranks (when $n$ is odd). However, it is by no means clear whether these chains may be chosen to be symmetric about the middle.

Emden Gansner has pointed out to me that for type $A_n$, there is a rank-preserving, order-preserving bijection $\mathbf{1}\times\mathbf{2}\times\cdots\times\mathbf{n}\xrightarrow{\varphi} W=\mathfrak{S}_n$, where $\mathbf{1}\times\mathbf{2}\times\cdots\times\mathbf{n}=\{(b_1,\cdots,b_n)\colon 0\leqq b_i<i\}$. Namely, $\varphi(b_1,\cdots,b_n)$ is that permutation $\pi=a_1a_2\cdots a_n\in\mathfrak{S}_n$ such that $b_i$ is the number of elements $j$ appearing in $\pi$ to the right of $i$ and satisfying $j<i$. Since any product of chains is a symmetric chain order (e.g., [17, pp. 30–31]), it follows that $\mathfrak{S}_n$ (with the Bruhat order) is also a symmetric chain order. A similar argument for types $B_n$ and $D_n$ produces rank-preserving order-preserving bijections $\mathbf{2}\times\mathbf{4}\times\cdots\times\mathbf{2n}\to\tilde{\mathfrak{S}}_n$ and $\mathbf{2}\times\mathbf{4}\times\cdots\times\mathbf{2(n-1)}\times\mathbf{n}\to\hat{\mathfrak{S}}_n$. Hence $\tilde{\mathfrak{S}}_n$ and $\hat{\mathfrak{S}}_n$ are also symmetric chain orders. However, we do not know for instance whether $L(m, n)$ and $M(n)$ are always symmetric chain orders. Lindström [32] has shown that $L(3, n)$ is a symmetric chain order, and D. West [44] has shown that $L(4, n)$ is a symmetric chain order. Littlewood [33, pp. 193–203] claims to prove that $L(m, n)$ is indeed a symmetric chain order for all $m$ and $n$. However, his proof is invalid. Specifically, it relies on the "method of chains" of Aitken [45], and this method is not correct as stated by Aitken. For the reader's benefit we will discuss the nature of Aitken's error in more detail. Let $P=\{x_1,\cdots,x_n\}$ be a finite poset, and let $\Phi=(a_{ij})$ be the $n\times n$ matrix defined by $a_{ij}=0$ unless $x_i<x_j$ in $P$; otherwise the $a_{ij}$'s are independent indeterminates over $\mathbb{Q}$. Remove a chain $C_1$ of maximum cardinality $c_1$ from $P$, then remove a chain $C_2$ of maximum cardinality $c_2$ from $P-C_1$, etc. Aitken essentially claims first that the numbers $c_1, c_2,\cdots$, are independent of the choice of chains $C_1, C_2,\cdots$, and second that the numbers $c_1, c_2,\cdots$ are the sizes of the Jordan blocks of $\Phi$. The first claim is clearly false. However, Littlewood's proof would still be valid if there were *some* way of choosing $C_1, C_2,\cdots$ so that the second claim is true. Even this weaker result is false. Let $P$ be the poset of Fig. 9. We have no choice but to take $c_1=4$, $c_2=1$, $c_3=1$. However, the Jordan block sizes of $\Phi$ are 4 and 2. A corrected version of Aitken's result appears in [37]. If this corrected result is used in conjunction with Littlewood's method, it yields the result that $L(m, n)$ has property T. Thus we have an alternative proof, avoiding the hard Lefschetz theorem (though actually Littlewood's method essentially proves the hard Lefschetz theorem for the Grassmann variety), that $L(m, n)$ has property T.

A further property of posets which implies the Sperner property is the LYM *property* [17, § 4]. However, Griggs has observed that $L(4, 3)$ fails to satisfy the LYM property.

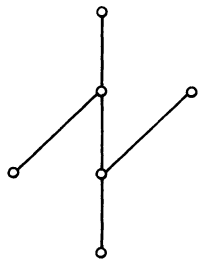


FIG. 9

*Note added in proof.* A proof that $L(3, m)$ and $L(4, m)$ have symmetric chain decompositions was first given by W. Riess, *Zwei Optimierungsprobleme auf Ordnungen*, Arbeitsberichte des Institute für Mathematische Maschinen und Datenverarbeitung (Informatik) 11, Number 5, Erlangen, April 1978.

## REFERENCES


[1] A. C. AITKEN, *Determinants and Matrices*, 3rd ed., Oliver and Boyd, London, 1944.

[2] G. E. ANDREWS, *A theorem on reciprocal polynomials with applications to permutations and compositions*, Amer. Math. Monthly, 82 (1975), pp. 830–833.

[3] G. BIRKHOFF, *Lattice Theory*, third ed., American Mathematical Society, Providence, RI, 1967.

[4] A. BOREL AND A. HAEFLIGER, *La classe d'homologie fondamentale d'un espace analytique*, Bull. Soc. Math. France, 89 (1961), pp. 461–513.

[5] A. BOREL AND J. TITS, *Compléments à l'article: "Groupes réductifs,"* Inst. Hautes Études Sci. Publ. Math., No. 41 (1972), pp. 253–276.

[6] N. BOURBAKI, *Groupes et algèbres de Lie*, Éléments de Mathématique, Fasc. XXXIV, Hermann, Paris, 1968.

[7] E. R. CANFIELD, *A Sperner property preserved by product*, Linear and Multilinear Algebra, to appear.

[8] R. W. CARTER, *Simple Groups of Lie Type*, Wiley, New York, 1972.

[9] L. COMTET, *Advanced Combinatorics*, Reidel, Boston, 1974.

[10] M. CORNALBA AND P. A. GRIFFITHS, *Some transcendental aspects of algebraic geometry*, Algebraic Geometry, Arcata 1974, R. Hartshorne, ed., American Mathematical Society, Providence, RI, 1975, pp. 3-110.

[11] V. V. DEODHAR, *Some characterizations of Bruhat ordering on a Coxeter group and determination of the relative Möbius function*, Invent. Math., 39 (1977), pp. 187–198.

[12] R. C. ENTRINGER, *Representation of m as* $\sum_{k=-n}^{n} \varepsilon_k k$, Canad. Math. Bull., 11 (1968), pp. 289–293.

[13] P. ERDÖS, *Extremal problems in number theory*, Theory of Numbers, A. L. Whiteman, ed., American Mathematical Society, Providence, RI, 1965, pp. 181–189.

[14] W. FULTON, *Rational equivalence on singular varieties*, Inst. Hautes Etudes Sci. Publ. Math., no. 45 (1975), pp. 147–167.

[15] C. GREENE, *Some partitions associated with a partially ordered set*, J. Combinatorial Theory, 20 (1976), pp. 69–79.

[16] C. GREENE AND D. J. KLEITMAN, *The structure of Sperner k-families*, J. Combinatorial Theory, 20 (1976), pp. 41–68.

[17] C. GREENE AND D. J. KLEITMAN, *Proof techniques in the theory of finite sets*, Studies in Combinatorics, G.-C. Rota, ed., Mathematical Association of America, Washington, DC, 1978, pp. 22–79.

[18] P. GRIFFITHS AND J. ADAMS, *Topics in Algebraic and Analytic Geometry*, Princeton University Press, Princeton, NJ, 1974.

[19] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, Wiley, New York, 1978.

[20] J. R. GRIGGS, *Sufficient conditions for a symmetric chain order*, SIAM J. Appl. Math., 32 (1977), pp. 807–809.

[21] J. R. GRIGGS, *On chains and Sperner k-families in ranked posets*, J. Combinatorial Theory, to appear.

[22] A. GROTHENDIECK, *Sur quelques propriétés fondamentales en théorie des intersections*, Séminaire C. Chevalley, Ecole Normale Supérieure, Paris, 2 (1958), Chap. 4.

[23] L. H. HARPER, *Stabilization and the edgesum problem*, Ars Combinatoria, 4 (1977), pp. 225–270.

[24] W. V. D. HODGE, *The Theory and Applications of Harmonic Integrals*, 2nd ed., Cambridge University Press, London, 1952.

[25] J. W. B. HUGHES, *Lie algebraic proofs of some theorems on partitions*, Number Theory and Algebra, H. Zassenhaus, ed., Academic Press, New York, 1977, pp. 135–155.

[26] S. L. KLEIMAN, *Problem 15. Rigorous foundation of Schubert's enumerative calculus*, *Mathematical Developments Arising from Hilbert Problems*, F. E. Browder, ed., American Mathematical Society, Providence, RI, 1976, pp. 445–482.

[27] S. L. KLEIMAN AND D. LAKSOV, *Schubert calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061–1082.

[28] V. LAKSHMIBAI, C. MUSILI AND C. S. SESHADRI, *Geometry of G/P*, Bull. Amer. Math. Soc., 1 (1979), pp. 432–435.

[29] S. LEFSCHETZ, *L'Analysis situs et la Géométrie Algébrique*, Gauthier-Villars, Paris, 1924; reprinted in Selected Papers, Chelsea, New York, 1971.

[30] B. LINDSTRÖM, *Conjecture on a theorem similar to Sperner's*, Combinatorial Structures and their Applications, R. Guy, H. Hanani, N. Sauer, and J. Schonheim, eds., Gordon and Breach, New York, 1970, p. 241.

[31] B. LINDSTRÖM, *Om en elementär sats i kombinatoriken och några tillämpningar*, Nordisk Mat. Tidskr., 17 (1969), pp. 61–70.

[32] ———, *A partition of $L(3, n)$ into saturated symmetric chains*, European J. Combinatorics, to appear.

[33] D. E. LITTLEWOOD, *The Theory of Group Characters and Matrix Representations of Groups*, 2nd ed., Oxford University Press, London, 1950.

[34] W. MESSING, *Short sketch of Deligne's proof of the hard Lefschetz theorem*, Algebraic Geometry, Arcata 1974, R. Hartshorne, ed., American Mathematical Society, Providence, RI, 1975, pp. 563–580.

[35] G. W. PECK, *Erdös' conjecture on sums of distinct numbers*, Studies in Applied Math., to appear.

[36] R. A. PROCTOR, M. E. SAKS AND D. G. STURTEVANT, *Product partial orders with the Sperner property*, Discrete Math., to appear.

[37] M. E. SAKS, *Dilworth numbers, incidence maps and product partial orders*, this Journal, this issue pp. 211–215.

[38] A. SÁRKÖZI AND E. SZEMERÉDI, *Über ein Problem von Erdös und Moser*, Acta Arith., 11 (1966), pp. 205–208.

[39] C. S. SESHADRI. *Geometry of $G/P$-I (Standard monomial theory for a miniscule $P$)*, C. P. Ramanujan: A Tribute, Springer-Verlag, 1978, published for the Tata Institute of Fundamental Research, Bombay.

[40] R. STANLEY, *Unimodal sequences arising from Lie algebras*, Proc. Young Day, to appear.

[41] ———, *Some combinatorial aspects of the Schubert calculus*, Combinatoire et Représentation du Groupe Symétrique, D. Foata, ed., Lecture Notes in Math. No. 579, Springer-Verlag, New York, 1977, pp. 217–251.

[42] J. H. VAN LINT, *Representation of 0 as $\sum_{k=-N}^{N} \varepsilon_k k$*, Proc. Amer. Math. Soc., 19 (1967), pp. 182–184.

[43] A. WEIL, *Introduction à l'étude des variétés Kählerienes*, Hermann, Paris, 1958.

[44] D. B. WEST, *A symmetric chain decomposition of $L(4, n)$*, preprint.

[45] A. C. AITKEN, *The normal form of compound and induced matrices*, Proc. London Math. Soc., (2) 38 (1934), pp. 354–376.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0006 $01.00/0

# THE GROWTH OF POWERS OF A NONNEGATIVE MATRIX*

SHMUEL FRIEDLAND† AND HANS SCHNEIDER†

**Abstract.** Let $A$ be a nonnegative $n \times n$ matrix. In this paper we study the growth of the powers $A^m$, $m = 1, 2, 3, \cdots$ when $\rho(A) = 1$. These powers occur naturally in the iteration process

$$x^{(m+1)} = Ax^{(m)}, \qquad x^{(0)} \geqq 0,$$

which is important in applications and numerical techniques. Roughly speaking, we analyze the asymptotic behavior of each entry of $A^m$. We apply our main result to determine necessary and sufficient conditions for the convergence to the spectral radius of $A$ of certain ratios naturally associated with the iteration above.

## 1. Introduction.

Let $A$ be a nonnegative $n \times n$ matrix. In the iteration process

$$x^{(m+1)} = Ax^{(m)}, \qquad x^{(0)} \geqq 0, \tag{1.1}$$

which is important in applications and numerical techniques, the powers $A^m$, $m = 1, 2, \cdots$ occur naturally. In this paper, we study the growth of these powers. In the literature there are several studies of the growth of $A^m$ when the elementary divisors belonging to the spectral radius $\rho(A)$ of $A$ are linear. For example, see Gantmacher [7, Chap. 13, § 5–7] Varga [19, pp. 32–34] when $A$ is irreducible, and Meyer–Plemmons [10] when $\lim_{m\to\infty} A^m$ exists. We deal here with the general nonnegative case, when the elementary divisors belonging to $\rho(A)$ may have degrees greater than 1. At the cost of ignoring nilpotent $A$, where the problem is trivial, we assume that $\rho(A) > 0$.

For a complex $n \times n$ matrix $A$, with $\rho(A) = 1$, there is a least integer $k$ for which $m^{-k}A^m$ is bounded, $m = 1, 2, \cdots$. However, even in the simple case of an imprimitive, irreducible nonnegative $A$, $\lim_{m\to\infty}\|m^{-k}A^m\|$ and, a fortiori $\lim_{m\to\infty} m^{-k}A^m$, do not in general exist. To obtain precise results for general nonnegative $A$ with $\rho(A) = 1$, it is thus necessary to introduce some smoothing. For example, in [14] Rothblum considered Cesaro means of powers of $A$. In this paper we study the growth of

$$B^{(m)} = A^m(I + \cdots + A^{q-1}), \qquad m = 1, 2, \cdots, \tag{1.2}$$

where $q$ is a certain positive integer.

After some preliminaries in § 2, we use elementary analytic methods in § 3 to prove a theorem on the growth of $B^{(m)}$. As corollary, we obtain a known theorem on the index of the eigenvalue 1 of $A$, cf. Schaefer [17, Chap. 1, Thm. 2.7]. We also give a local form of the theorem; that is, we show that for $1 \leqq i, j \leqq n$ there exist integers $k = k(i, j)$ and $q = q(i, j) > 0$ such that the element $b_{ij}^{(m)}$ of the matrix given by (1.2) satisfies

$$\lim_{m\to\infty} m^{-k} b_{ij}^{(m)} > 0. \tag{1.3}$$

The analytic results of § 3 motivate the investigations in the rest of the paper.

The main thrust of the paper is the use of the graph structure of the matrix $A$ to decrease the integer $q(i, j)$ and to determine the integer $k(i, j)$ in (1.3). The requisite graph theoretic concepts are developed in § 4, and in § 5 we state our main result, Theorem (5.10). As a corollary, we obtain a striking theorem on the index of 1 due to

* Received by the editors June 14, 1979, and in revised form December 4, 1979.

† Mathematics Department and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. This research was supported in part by the United States Army under Contract DAAG29-75-C-0024 and by the National Science Foundation under Grant MCS78-01087.

Rothblum [13]. Our results are related to those of U. G. Rothblum [14], [15], and in some instances, would also follow from his. But where Rothblum considers $A^{qm}$, $m = 1, 2, \cdots$, we consider $B^{(m)}$ and this allows us to choose a smaller integer $q$. Our definitions of $q(i, j)$ involves the greatest common divisor (g.c.d.) of certain periods where one might expect the least common multiple (l.c.m.). Consider the example

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

Then, by direct computation, for $1 \leqq i, j \leqq 2$, $\lim_{m\to\infty} b_{ij}^{(m)} = 1$, where $B^{(m)} = A^m(I + A)$. Thus $k(i, j) = 0$, and we may choose $q(i, j) = 2$ if $1 \leqq i, j \leqq 2$. Similarly $k(i, j) = 0$, $q(i, j) = 3$ if $3 \leqq i, j \leqq 5$. Yet $\lim_{m\to\infty} m^{-1} a_{ij}^{(m)} = \frac{1}{6}$ if $1 \leqq i \leqq 2$, $3 \leqq j \leqq 6$, and so we have $k(i, j) = 1$, $q(i, j) = 1$. We might add that it may be possible that our choice of $q(i, j)$ can be improved in the general case where we use an l.c.m. of certain g.c.d.'s.

In § 6, we apply our results to the iteration process (1.1) for any nonnegative matrix $A$ satisfying $\rho(A) > 0$. For $x \geqq 0$, $x \neq 0$ denote

$$r(x) = \sup\{\mu : \mu x \leqq Ax\}, \tag{1.4i}$$

$$R(x) = \inf\{\mu : \mu x \geqq Ax\}. \tag{1.4ii}$$

In Theorem 6.8, we find necessary and sufficient conditions for $r(A^m x)$ and $R(A^m x)$ to converge to the spectral radius of $A$. We show that whether or not this happens depends only on what is in general a small part of the vector $x$. In § 7, we show that a theorem due to D. H. Carlson [3] on the existence of nonnegative solutions $y$ for $(I - A)y = x$, $x \geqq 0$, $\rho(A) = 1$ is a consequence of our main results and we extend the theorem.

## 2. Preliminaries.

**Notations.** Let $\varphi(1), \varphi(2), \cdots$, be a sequence of nonnegative numbers and $k \geqq 0$ be an integer.

(2.1)
(i) $\varphi(m) = O(m^k)$ will denote that $\varphi(m)/m^k$, $m = 1, 2, \cdots$, is bounded.
(ii) $\varphi(m) = o(m^k)$ will denote that $\lim_{m\to\infty} \varphi(m)/m^k = 0$.
(iii) $\varphi(m) \approx m^k$ will denote that $\lim_{m\to\infty} \varphi(m)/m^k$ exists and is positive.
(iv) The above notations will also be used for $k = -1, -\infty$. In case that $k = -1$ $\varphi(m) = O(m^k)$, $\varphi(m) = o(m^k)$, $\varphi(m) \approx m^k$ will each indicate that there exists $\rho$, $0 < \rho < 1$, such that $\varphi(m)\rho^{-m} = O(1)$. In case that $k = -\infty$ the above notations will mean that $\varphi(m) = 0$ for all sufficiently large $m$. (Thus $\varphi(m) \approx m^{-\infty}$ implies $\varphi(m) \approx m^{-1}$.)
(v) The notation $A(m) \approx m^k$ will be used for a sequence of nonnegative matrices $A(1), A(2), \cdots$ to indicate the relation holds for each element.

**Combinatorial result.** Let $r \geqq 0$ and $t > 0$ be integers. Then

$$\Gamma_t^r = \sum_{p_1+\cdots+p_t=r} 1^{p_1} 1^{p_2} \cdots 1^{p_t}, \tag{2.2}$$

where the summation is taken over all nonnegative integers $p_1, \cdots, p_t$ whose sum is $r$. That is, $\Gamma_t^r$ is the number of collections of $r$ objects chosen from $t$ distinct objects, with

repetitions allowed. It is well-known that

$$\Gamma_t^r = \binom{r+t-1}{r}. \tag{2.3}$$

A simple way to prove this equality is by considering the coefficient of $x^r$ of both sides of the identity

$$\sum_{r=0}^{\infty} \binom{r+t-1}{r} x^r = \left(\sum_{r=0}^{\infty} x^r\right)^{-t}$$

which is derived from $(1-x)^{-t} = (1-x)^{-1} \cdots (1-x)^{-1}$. For a purely combinatorial proof see for example Brualdi [2, p. 37]. For $t = 0$ the above formula implies $\Gamma_0^r = 1$ for all $r \geqq 0$.

We shall also need some results on the convergence of series.

LEMMA 2.4. *Given integers $k \geqq 1$, $q > 0$, and let $b_p \geqq 0$, $p = 0, 1, 2, \cdots$ be a sequence such that*

$$\lim_{p\to\infty} p^{-(k-1)}(b_p + \cdots + b_{p+q-1}) = v, \tag{2.5}$$

*where $q > 0$. Then*

$$\lim_{m\to\infty} m^{-k} \sum_{p=1}^{m} b_p = \frac{v}{kq}. \tag{2.6}$$

*Proof.* Elementary. Alternatively, check that $c_{m,p} = m^{-k}kp^{k-1}$ satisfies the assumptions of Hardy [8, Thm. 2, p. 43]. □

LEMMA 2.7. *Suppose* (2.5) *holds. If* $\lim_{m\to\infty} a_m = u$ *then*

$$\lim_{m\to\infty} m^{-k} \sum_{p=1}^{m} a_p b_{m-p} = \frac{uv}{kq}. \tag{2.8}$$

*Proof.* According to Hardy [8, Thm. 16, p. 64]

$$\lim_{m\to\infty} \frac{\sum_{p=1}^{m} a_p b_{m-p}}{\sum_{p=1}^{m} b_p} = u \tag{2.9}$$

since

$$0 \leqq \frac{b_m}{\sum_{p=1}^{m} b_p} \leqq \frac{b_m + \cdots + b_{m+q-1}}{\sum_{p=1}^{m} b_p} \leqq \frac{2vm^{(k-1)}}{v(2kq)^{-1}m^k},$$

and the last expression tends to 0. If we apply (2.6) to (2.9) we obtain (2.8). □

**3. Analytic approach.** By $\mathbb{R}$, resp. $\mathbb{C}$, we denote the real, resp. complex field, and by $\mathbb{R}_+$ the nonnegative numbers. The set of real, resp. complex, nonnegative $r \times n$ matrices will be denoted by $\mathbb{R}^{rn}$, resp. $\mathbb{C}^{rn}$, $\mathbb{R}_+^{rn}$. We also write $A \geqq 0$ for $A \in \mathbb{R}_+^{rn}$ ($A$ is nonnegative) and $A > 0$ when $A$ is positive ($a_{ij} > 0$, $i = 1, \cdots, r$, $j = 1, \cdots, n$).

*Let* $A \in \mathbb{C}^{nn}$. By spec $A$ we denote the set of eigenvalues of $A$. Suppose that spec $A = \{\lambda_1, \cdots, \lambda_r\}$, where the $\lambda_\alpha$ are pairwise distinct. It is known (cf. Gantmacher [7, Chap. 5, § 3]) that there exist nonnegative integers $p_1, \cdots, p_r$ and unique matrices $Z^{(\alpha\beta)} \in \mathbb{C}^{nn}$, $\beta = 0, \cdots, p_\alpha$, $\alpha = 1, \cdots, r$ which are linearly independent such that for each polynomial $f(\tau)$,

$$f(A) = \sum_{\alpha=1}^{r} \sum_{\beta=0}^{p_\alpha} f^{(\beta)}(\lambda_\alpha) Z^{(\alpha\beta)}. \tag{3.1}$$

The $Z^{(\alpha\beta)}$ are polynomials in $A$, $p_\alpha+1$ is the size of a largest Jordan-block belonging to $\lambda_\alpha$. The columns of $Z^{(\alpha p_\alpha)}$ are eigenvectors of $A$ corresponding to the eigenvalue $\lambda_\alpha$, the rank of $Z^{\alpha p_\alpha}$ is equal to the number of Jordan blocks of size $p_\alpha+1$ corresponding to $\lambda_\alpha$. (The simplest way to obtain (3.1) is by assuming that $A$ is in Jordan form.) As usual we define

$$\text{index}\,(\lambda_\alpha)=p_\alpha+1.$$

That is, $p_\alpha+1$ is the multiplicity of $\lambda_\alpha$ in the minimal polynomial of $A$. We shall also use a localized index. For $1\leqq i,j\leqq n$ we put

$$\text{index}_{ij}(\lambda_\alpha)=1+\max\{\beta\colon z_{ij}^{(\alpha\beta)}\neq 0,\beta=0,\cdots,p_\alpha\},$$

where $\text{index}_{ij}(\lambda_\alpha)=0$ if $z_{ij}^{(\alpha\beta)}=0,\beta=0,\cdots,p_\alpha$. If $A\in\mathbb{C}^{nn}$ and $m$ is any integer we shall denote the elements of $A^m$ by $a_{ij}^{(m)}$, $1\leqq i,j\leqq m$.

Let $A\in\mathbb{R}_+^{nn}$. We assume throughout the normalization $\rho(A)=1$. It is well-known (see Frobenius [6], Gantmacher [7, Chap. 13], Berman–Plemmons [1, Chap. 2]) that if $\lambda$ is an eigenvalue of $A$ and $|\lambda|=1$, then $\lambda$ is a root of 1. Hence, there is a positive integer $q$ such that $\lambda^q=1$, for all $\lambda\in\text{spec}\,A$, $|\lambda|=1$. The smallest such integer $q$ will be called the *period* of $A$. If $q=1$, $A$ will be called *aperiodic*. For an irreducible and aperiodic matrix $A\geqq 0$, the Frobenius theorem and the formula (3.1) imply

$$\lim_{m\to\infty}A^m=Z^{(10)}>0,$$

where $\lambda_1=1$, see for example Berman–Plemmons [1, Chap. 2, Thm. 4.1]. Theorem 3.4 extends the above equality in a local way. Part (i) of the theorem is an extension of the known inequality apparently due to Schaefer [16, Thm. 2.4, p. 264],

$$\text{index}\,(\lambda)\leqq\text{index}\,(1)\quad\text{if}\ |\lambda|=1, \tag{3.2}$$

for nonnegative matrices; see also Schaefer [17, Chap. 1, Thm. 2.7], Berman–Plemmons [1, Chap. 1, Thm. 3.2]. This result and part (i) of Theorem 3.4 could easily be deduced from the classical Pringsheim theorem on analytic functions; e.g., see Titchmarsh [18, p. 214]. The use of the Pringsheim theorem in analyzing the spectral properties of nonnegative matrices can be traced back to Ostrowski [11] (see also Karlin [9] and Schaefer [16, Appendix] for the infinite dimensional case). See Friedland [5] for a detailed analysis of the Pringsheim theorem for rational functions which has certain analogs to the Frobenius theorem. For sake of completeness we bring a short and elementary independent proof of Theorem 3.4. To do so we need an easy lemma which probably is known.

LEMMA 3.3. *Let $\lambda_\alpha$, $z_\alpha$, $\alpha=1,\cdots,r$ be complex numbers, where the $\lambda_\alpha$ are pairwise distinct. If $\lim_{m\to\infty}(\sum_{\alpha=1}^r\lambda_\alpha^m z_\alpha)$ exists, then $z_\alpha=0$ if $|\lambda_\alpha|\geqq 1$, $\lambda_\alpha\neq 1$.*

*Proof.* Since $\lim_{m\to\infty}\lambda_\alpha^m$ exists for $|\lambda_\alpha|<1$ or $\lambda_\alpha=1$, without loss of generality we may assume that $|\lambda_\alpha|\geqq 1$, $\lambda_\alpha\neq 1$, $\alpha=1,\cdots,r$. Put $z=(z_1,\cdots,z_r)^t\in\mathbb{C}^r$ and $u^{(m)}=(u_m,\cdots,u_{m+r-1})^t$, where $u_m=\sum_{\alpha=1}^r\lambda_\alpha^m z_\alpha$. Let $\Lambda=\text{diag}\{\lambda_1,\cdots,\lambda_r\}\in\mathbb{C}^{rr}$ and let $V=(v_{\alpha\beta})_1^r\in\mathbb{C}^{rr}$ be the Vandermond matrix given by $v_{\alpha\beta}=\lambda_\beta^{\alpha-1}$, $\alpha,\beta=1,\cdots,r$. Then

$$u^{(m)}=V\Lambda^m z.$$

The assumption of the lemma implies that $\lim_{m\to\infty}u^{(m)}$ exists. Since $V$ is nonsingular, $\lim_{m\to\infty}\Lambda^m z=\lim_{m\to\infty}V^{-1}u^{(m)}$ and so $z=0$. □

THEOREM 3.4. *Let $A\in\mathbb{R}_+^{nn}$ where $\rho(A)=1$. Let $1\leqq i,j\leqq n$.*

(i) *If $\lambda\in\text{spec}\,A$, $|\lambda|=1$, then $\text{index}_{ij}\,(\lambda)\leqq\text{index}_{ij}\,(1)$.*

(ii) *Let $q$ be a positive integer such that $\lambda^q=1$ if $\lambda\in\text{spec}\,A$, $|\lambda|=1$ and $\text{index}_{ij}(\lambda)=$*

$\text{index}_{ij}(1)$. *Put* $k+1=\text{index}_{ij}(1)$ *and let*

$$B^{(m)}=A^m(I+\cdots+A^{q-1}).$$

*Then* $b_{ij}^{(m)}\approx m^k$. *In particular,* $a_{ij}^{(m)}\neq o(m^k)$ *if* $k\geqq 0$.

*Proof.* (i) Let $\{\lambda_1,\cdots,\lambda_r\}$ be the eigenvalues with $|\lambda_\alpha|=1$, $\alpha=1,\cdots,r$, where the $\lambda_\alpha$ are pairwise distinct. Let

$$d+1=\max\{\text{index}_{ij}\{\lambda_\alpha\}:\alpha=1,\cdots,r\}.$$

If $d=-1$ then there is nothing to prove. So assume that $d\geqq 0$. Suppose that $z_\alpha\equiv z_{ij}^{(\alpha d)}\neq 0$ for $\alpha=1,\cdots,s$ where $1\leqq s\leqq r$ and $z_{ij}^{(\alpha d)}=0$ for $\alpha=s+1,\cdots,r$. It follows immediately from (3.1) that

$$a_{ij}^{(m)}=m^d\Big(\sum_{\alpha=1}^{s}\lambda_\alpha^{m-d}z_\alpha\Big)+o(m^d).$$

Hence, by Lemma (3.3), $a_{ij}^{(m)}\neq o(m^d)$.

Let $q$ be a positive integer such that $\lambda_\alpha^q=1$, $\alpha=1,\cdots,s$. Define

$$\varphi_m(\tau)=\tau^m(1+\tau+\cdots+\tau^{q-1}).$$

If we take the $d$th derivative of $\varphi_m(\tau)$, we obtain

$$\varphi_m^{(d)}(\tau)=m^d\varphi_{m-d}(\tau)+o(m^d)$$

for any fixed $\tau$, $|\tau|\leqq 1$, and also $\varphi_{m-d}(\lambda_\alpha)=0$ for $|\lambda_\alpha|=1$, $\lambda_\alpha\neq 1$, $1\leqq\alpha\leqq s$. Put $B^{(m)}=\varphi_m(A)$. By (3.1) and the equality above we have

$$b_{ij}^{(m)}=m^d\Big(\sum_{\alpha=1}^{r}\varphi_{m-d}(\lambda_\alpha)z_\alpha\Big)+o(m^d). \tag{3.5}$$

Now suppose that $\text{index}_{ij}(1)<d+1$. Then (3.5) implies that $b_{ij}^{(m)}=o(m^d)$. But $b_{ij}^{(m)}=a_{ij}^{(m)}+\cdots+a_{ij}^{(m+q-1)}\geqq a_{ij}^{(m)}\geqq 0$, and this is a contradiction. Thus $d=k$ and this proves (i).

(ii) Suppose that $\lambda_1=1$. If $k=-1$, by an argument like that above, $a_{ij}^{(m)}=b_{ij}^{(m)}\approx m^k$. Let $k\geqq 0$. By (3.5) and the preceding argument we obtain

$$b_{ij}^{(m)}=m^kqz_1+o(m^k),$$

where $z_1=z_{ij}^k>0$. This proves (ii). □

We now state a global version of Theorem 3.4 (ii) which follows immediately from Theorem 3.4.

THEOREM 3.6. *Let* $A\in\mathbb{R}_+^{nn}$ *where* $\rho(A)=1$. *Let* $q$ *be a positive integer such that* $\lambda^q=1$ *if* $\lambda\in\text{spec}\,A$, $|\lambda|=1$ *and* $\text{index}\,(\lambda)=\text{index}\,(1)=k+1$. *Let*

$$B^{(m)}=A^m(I+\cdots+A^{q-1}).$$

*Then*

$$\lim_{m\to\infty}m^{-k}B^{(m)}=F, \tag{3.7}$$

*where* $F\geqq 0$ *and* $F$ *is not identically zero.*

It should be noted that the assumption that $A$ is nonnegative was used crucially in the proof of Theorems 3.4 and 3.6. For example, let $A=-I$; then there are no $k$, $q$ for which the limit of (3.7) exists and is nonzero. Also, the assumption that $\rho(A)=1$ is used

in an essential way. Let

$$A=\begin{bmatrix}0 & 1\\ 4 & 0\end{bmatrix}.$$

Then $\lim_{m\to\infty}\rho(A)^{-2m}A^{2m}(I+A)$ and $\lim_{m\to\infty}\rho(A)^{-(2m+1)}A^{2m+1}(I+A)$ exist, but are distinct. It follows that no $k, q$ exist for which $\lim_{m\to\infty}\rho(A)^{-m}m^{-k}B^{(m)}$ exists and is nonzero.

Our subsequent work discusses the nature of $k$, $q$ and $F$.

**4. Graph theoretical concepts.** Let $A\in\mathbb{R}_+^{nn}$ and let $\rho(A)>0$. We may assume, without loss of generality, that after simultaneous permutations of rows and columns, $A$ is in the Frobenius [6] normal form which can be found in many references, e.g., Gantmacher [7, Vol. II, p. 75]. Thus

$$A=\begin{bmatrix}A_{11} & A_{12} & \cdots & A_{1\nu}\\ & A_{22} & \ddots & \\ & & \ddots & \\ 0 & & & A_{\nu\nu}\end{bmatrix}, \tag{4.1}$$

where the diagonal blocks $A_{\alpha\alpha}$, $\alpha=1,\cdots,\nu$ are irreducible and all subdiagonal blocks are 0. (The $1\times 1$ matrix 0 is considered to be irreducible.)

Let $A$ be in Frobenius normal form (4.1). Then the (reduced) *graph* $G(A)$ of $A$ is a subset of $\langle\nu\rangle\times\langle\nu\rangle$, where $\langle\nu\rangle=\{1,\cdots,\nu\}$ and $G(A)=\{(\alpha,\beta)\in\langle\nu\rangle\times\langle\nu\rangle: A_{\alpha\beta}\neq 0\}$. (Observe that many authors would call $G(A)$ the arcset of the graph $(\langle\nu\rangle, G(A))$, but we have no need to mention the vertex set $\langle\nu\rangle$ explicitly.)

If $(\alpha,\beta)\in G(A)$, we call $(\alpha,\beta)$ an *arc* of $G(A)$. If $(\alpha,\beta)$ is an arc of $G(A)$, then $\alpha\leqq\beta$; also $(\alpha,\alpha)\in G(A)$, $1\leqq\alpha\leqq\nu$, unless $A_{\alpha\alpha}$ is the $1\times 1$ matrix 0. Thus we define a (simple) *path from $\alpha$ to $\beta$ in $G(A)$* to be a sequence $\pi=(\alpha_0,\cdots,\alpha_s)$, where *either* $s\geqq 1$, $1\leqq\alpha=\alpha_0<\cdots<\alpha_s=\beta\leqq\nu$ and $(\alpha_{i-1},\alpha_i)\in G(A)$, $i=1,\cdots,s$, *or* $s=0$ and $\alpha=\alpha_0=\beta$ and $(\alpha,\alpha)\in G(A)$. The *support of* $\pi$ is the set $\operatorname{supp}\pi=\{\alpha_0,\cdots,\alpha_s\}\subseteq\{1,\cdots,\nu\}$. We always assume that the $\alpha_i$, $i=0,\cdots,s$, have been listed in *strictly ascending order.*

If $1\leqq\alpha\leqq\nu$, then we call $\alpha$ a *singular vertex* (of $G(A)$) if $\rho(A_{\alpha\alpha})=\rho(A)$. (This terminology is consistent with that of Richman–Schneider [12].) Let $1\leqq\alpha\leqq\beta\leqq\nu$. For any path $\pi$ from $\alpha$ to $\beta$ in $G(A)$, let $k(\pi)+1$ be the number of singular $\gamma$ in the support of $\pi$. (Thus note each distinct $\gamma$ is counted only once in $k(\pi)+1$.) Let $\alpha_{j_0}<\alpha_{j_1}<\cdots<\alpha_{j_k}$, where $k=k(\pi)$, be all singular vertices in $\operatorname{supp}\pi$. We set

$$k(\alpha,\beta)=\max\{k(\pi):\ \pi \text{ is a path from } \alpha \text{ to } \beta \text{ in } G(A)\}. \tag{4.2}$$

If there is no path from $\alpha$ to $\beta$ in $G(A)$ we put $k(\alpha,\beta)=-\infty$. We shall call $k(\alpha,\beta)$ the *singular distance* from $\alpha$ to $\beta$. If $(i,i)$ is a position in $A_{\alpha\alpha}$ and $(j,j)$ a position in $A_{\beta\beta}$ then we shall also call $k[i,j]=k(\alpha,\beta)$ the singular distance from $i$ to $j$ (note our use of square brackets).

A path $\pi$ from $\alpha$ to $\beta$ will be called a *maximal path* if the number of singular vertices in the support of $\pi$ is $k(\alpha,\beta)+1$. Let $1\leqq\alpha,\beta\leqq\nu$. Let $\mathcal{P}(\alpha,\beta)$ be the set of maximal paths from $\alpha$ to $\beta$. For each $\pi\in\mathcal{P}(\alpha,\beta)$ let $q(\pi)$ be the g.c.d. of periods of $A_{\gamma\gamma}$ with $\gamma\in\operatorname{supp}\pi$ and singular (viz. $\rho(A_{\gamma\gamma})=\rho(A)$).

Then we define

$$q(\alpha,\beta)=\text{l.c.m.}\{q(\pi):\ \pi\in\mathcal{P}(\alpha,\beta)\}. \tag{4.3}$$

We shall call $q(\alpha,\beta)$ the *local period* of $(\alpha,\beta)$. If $k(\alpha,\beta)<0$ then $q(\alpha,\beta)=1$. Also if

$(i, i)$ is a position in $A_{\alpha\alpha}$ and $(j, j)$ is a position in $A_{\beta\beta}$ then we shall put $q(\alpha, \beta) = q[i, j]$, the local period of $(i, j)$.

**5. The main results.** Let $A \in \mathbb{R}_+^{nn}$, where $\rho(A) = 1$, be in Frobenius normal form (4.1). It follows from the Perron–Frobenius theory for nonnegative matrices, e.g., Gantmacher [7, Chap. 13] that there is a diagonal matrix $X$ with positive diagonal elements so that, upon replacing $A$ by $X^{-1}AX$,

$$A_{\alpha\alpha} = \rho(A_{\alpha\alpha})A'_{\alpha\alpha}, \tag{5.1}$$

where $A'_{\alpha\alpha}$ is a stochastic matrix,

$$\|A_{\alpha\beta}\|_\infty \leqq \sigma, \qquad 1 \leqq \alpha < \beta \leqq \nu, \tag{5.2}$$

where $1 > \sigma$ and $\sigma > \max\{\rho(A_{\alpha\alpha}): \rho(A_{\alpha\alpha}) < 1, \alpha = 1, \cdots, \nu\}$ if such $\alpha$ exist. Here $\| \ \|_\infty$ is the $l_\infty$-operator norm,

$$\|Z\|_\infty = \max\left\{ \sum_{j=1}^{n} |z_{ij}|: i = 1, \cdots, r \right\} \quad \text{for } Z \in \mathbb{R}^{rn}.$$

The diagonal matrix $X$ can be constructed as follows. Let $u^{(\alpha)}$ be a positive vector satisfying $A_{\alpha\alpha}u^{(\alpha)} = \rho(A_{\alpha\alpha})u^{(\alpha)}$. Denote by $X_\alpha$ a diagonal matrix, whose diagonal entries are the elements of $u^{(\alpha)}$. Then $X$ is of the form diag $\{X_1, \varepsilon X_2, \cdots, \varepsilon^{\nu-1}X_\nu\}$ for some small enough positive $\varepsilon$. In our subsequent proofs we may assume that $A$ has been normalized as above.

Let $\pi$ be a path in $G(A)$. Denote by $s+1$ the cardinality of supp $\pi$. That is

$$\text{supp } \pi = \{\beta_0, \cdots, \beta_s\}, \qquad 1 \leqq \beta_0 < \beta_2 < \cdots < \beta_s \leqq \nu. \tag{5.3i}$$

We define the *path matrix* $A(\pi)$ by

$$\begin{aligned} &A_{ii}(\pi) = A_{\beta_i\beta_i}, && i = 0, \cdots, s, \\ &A_{i,i+1}(\pi) = A_{\beta_i\beta_{i+1}}, && i = 0, \cdots, s-1, \\ &A_{ij}(\pi) = 0, && i, j = 0, \cdots, s \text{ otherwise,} \end{aligned} \tag{5.3ii}$$

$$A(\pi) = (A_{ij}(\pi))_0^s. \tag{5.3iii}$$

Thus $A(\pi)$ is in Frobenius normal form and has $s+1$ irreducible diagonal blocks $A_{ii}(\pi) = A_{\beta_i\beta_i}$, $i = 0, \cdots, s$. To avoid ambiguity, we write $A(\pi)_{ij}^{(m)}$ for the $(i, j)$ block component of $A(\pi)^m$, $i, j = 0, \cdots, s$.

We now prove a sequence of lemmas for the path matrix $A(\pi)$ of a given path.

LEMMA 5.4. *Let $A \in \mathbb{R}_+^{nn}$ where $\rho(A) = 1$. Let $1 \leqq \alpha, \beta \leqq \nu$ and $\pi$ be a path in $G(A)$ from $\alpha$ to $\beta$. Put $k = k(\pi)$, where $k(\pi)+1$ is the number of singular vertices in* supp $\pi$. *If $A(\pi)$ is the path matrix given by* (5.3), *then $\|A(\pi)_{0s}^{(m)}\|_\infty = O(m^k)$.*

*Proof.* We note that

$$A(\pi)_{0s}^{(m)} = \sum_{p_0+\cdots+p_s=m-s} A_{00}^{p_0}(\pi)A_{01}(\pi)A_{11}^{p_1}(\pi)\cdots A_{(s-1)s}(\pi)A_{ss}^{p_s}(\pi). \tag{5.5}$$

So

$$\|A(\pi)_{0s}^{(m)}\|_\infty \leqq \sigma^s \sum_{p_0+\cdots+p_s=m-s} \|A_{00}(\pi)\|_\infty^{p_0} \cdots \|A_{ss}(\pi)\|_\infty^{p_s}.$$

Suppose first that $\pi$ does not contain singular vertices, i.e., $k = -1$. Then

$$\|A(\pi)_{0s}^{(m)}\|_\infty \leqq \sigma^m \sum_{p_0+\cdots+p_s=m-s} 1^{p_0} \cdots 1^{p_s} = \sigma^m \Gamma_{s+1}^{m-s},$$

where $\Gamma_t^r$ is given by (2.3). As $\Gamma_s^{m-s} \leqq m^s$ we immediately deduce

$$\lim_{m\to\infty} \tau^{-m} A(\pi)_{0s}^{(m)} = 0 \qquad \text{for any } \tau,\ \sigma < \tau < 1.$$

Suppose now that $k \geqq 0$. Then

$$\begin{aligned}\|A(\pi)_{0s}^{(m)}\|_\infty &\leqq \sigma^s \sum_{q_0+\cdots+q_s=m-s} 1^{q_0}\cdots 1^{q_k}\sigma^{q_{k+1}}\cdots\sigma^{q_s} \\ &= \sigma^s \sum_{u=0}^{m-s}\Big(\sum_{q_0+\cdots+q_k=u} 1^{q_0}\cdots 1^{q_k}\Big)\Big(\sum_{q_{k+1}+\cdots+q_s=m-s-u}\sigma^{q_{k+1}}\cdots\sigma^{q_s}\Big) \\ &= \sigma^s \sum_{u=0}^{m-s} \Gamma_{k+1}^u \Gamma_{s-k}^{m-s-u}\sigma^{m-s-u}.\end{aligned}$$

Hence

$$\|A(\pi)_{0s}^{(m)}\|_\infty \leqq \Gamma_{k+1}^{m-s}\Big(\sum_{\nu=0}^{\infty}\Gamma_{s-k}^{\nu}\sigma^{\nu+s}\Big).$$

The last series converges by the ratio test and $\Gamma_{k+1}^{m-s} \leqq m^k$. This establishes the lemma. □

LEMMA 5.6. *Let the assumptions of Lemma* 5.4 *hold. Assume furthermore that* $k \geqq 0$, *i.e., the support of* $\pi$ *contains singular vertices. Then, for sufficiently large* $m$,

$$\sum_{j=0}^{2(s+1)(n-1)} A(\pi)_{0s}^{(m+j)} \geqq Gm^k, \tag{5.7}$$

*where* $G$ *is a positive matrix.*

*Proof.* Let

$$B_{ii}(\pi) = I + A_{ii}(\pi) + \cdots + A_{ii}(\pi)^{(n-1)}, \qquad i = 1, \cdots, s.$$

Since $A_{ii}(\pi)$ is irreducible, and its dimension does not exceed $n$, we have $B_{ii}(\pi) > 0$, Wielandt [20], Berman–Plemmons [1, Chap. 2, Thm. 1.3]. Clearly (5.5) implies, for $t = 2(s+1)(n-1)$,

$$\begin{aligned}\sum_{j=0}^{t} A(\pi)_{0s}^{(m+j)} \geqq n^{-(s+1)} \sum_{p_0+\cdots+p_s=m-s} & B_{00}(\pi)A_{00}(\pi)^{p_0}B_{00}(\pi)A_{01}(\pi)B_{11}(\pi)A_{11}^{p_1}B_{11}(\pi)\cdots \\ & \cdot A_{s-1,s}(\pi)B_{ss}(\pi)A_{ss}^{p_s}(\pi)B_{ss}(\pi).\end{aligned}$$

For $i, j = 0, \cdots, s$, let $E_{ij}$ be the matrix all of whose entries equal 1 and whose dimension is that of $A_{ij}(\pi)$. Clearly $B_{00}(\pi) \geqq c_0' E_{00}$, $B_{ss}(\pi) \geqq c_s' E_{ss}$ where $c_0', c_s' > 0$. Since $A_{i,i+1}(\pi) \neq 0$, we have

$$B_{ii}(\pi)A_{i,i+1}(\pi)B_{i+1,i+1}(\pi) \geqq c_i E_{i,i+1},$$

where $c_i > 0$, $i = 1, \cdots, s-1$, and hence, for some $c > 0$,

$$\sum_{j=1}^{t} A(\pi)_{0s}^{(m+j)} \geqq c \sum_{p_0+\cdots+p_s=m-s} E_{00}A_{00}(\pi)^{p_0}E_0, \cdots, E_{s-1,s}A_{ss}(\pi)^{p_s}E_{ss}. \tag{5.8}$$

In the inequality (5.8) we may restrict the sum on the right-hand side by letting $p_j = 0$ if $\rho(A_{jj}(\pi)) < 1$. So let $\gamma_0 < \cdots < \gamma_k$ be the subscripts of $A_{ii}$ which are singular vertices

and put $\bar{A}_{ii} = A_{\gamma_i\gamma_i}(\pi)$. Since $E_{ij}E_{jk} \geqq E_{ik}$, it follows that

$$\sum_{j=0}^{t} A(\pi)_{0s}^{(m+j)} \geqq c' \sum_{p_0+\cdots+p_k=m-s} \bar{E}_{-1,0}\bar{A}_{00}(\pi)^{p_0}\bar{E}_{01}\cdots\bar{A}_{kk}^{p_k}\bar{E}_{k,k+1},$$

where $c'>0$ and the $\bar{E}_{i,i+1}$, $i=-1,\cdots,k$ are matrices all of whose entries are 1. But $\bar{A}_{ii}(\pi)$ is a stochastic matrix, $i=0,\cdots,k$, whence $\bar{A}_{ii}(\pi)^{p_0}\bar{E}_{i,i+1} = \bar{E}_{i,i+1}$, $i=0,\cdots,k$. It follows that

$$\sum_{j=0}^{t} A(\pi)_{0s}^{(m+j)} \geqq 2\Gamma_{k+1}^{m-s}G,$$

where $G>0$. The lemma now follows from (2.3) since $\Gamma_k^{m-s} \geqq \frac{1}{2}m^k$ for sufficiently large $m$. □

LEMMA 5.9. *Let the assumptions of Lemma* 5.4 *hold, and suppose that* $k = k(\pi) \geqq 0$. *Let* $q = q(\pi)$ *be the g.c.d. of periods of* $A_{\gamma\gamma}$ *for singular* $\gamma \in \operatorname{supp}\pi$. *Let*

$$B(\pi)^{(m)} = A(\pi)^m(I + A(\pi) + \cdots + A(\pi)^{q-1}).$$

(i) *If* $(i,j)$ *is any position in* $A(\pi)_{0s}$ *then, in* $A(\pi)$, *index*$_{ij}$ $(1) = k+1$.

(ii) $b(\pi)_{ij}^{(m)} \approx m^k$.

*Proof.* (i) Let $k^*+1 = \text{index}_{ij}(1)$ in $A(\pi)$. By Theorem 3.4 there is a positive integer $q^*$ such that for

$$B^*(\pi)^{(m)} = A(\pi)^m(I + A(\pi) + \cdots + A(\pi)^{q^*-1}),$$

we have $b^*(\pi)_{ij}^{(m)} \approx m^{k^*}$. But $k^* > k$ contradicts Lemma 5.4. Since the sum in (5.7) can be majorized by a sum of terms of the form $B^*(\pi)_{0s}^{(m+j)}$, $j=0,\cdots,2(s+1)(n-1)$, it follows that $k^* < k$ contradicts Lemma 5.6. Hence $k^* = k$.

(ii) Now suppose that $\lambda \in \operatorname{spec} A(\pi)$, $|\lambda| = 1$ and $\text{index}_{ij}(\lambda) = \text{index}_{ij}(1) = k+1$ in $A(\pi)$. Then

$$\text{index}_{ij}(\lambda) \leqq \text{index}(\lambda) \leqq \text{mult}(\lambda),$$

where mult $(\lambda)$ is the algebraic multiplicity of $\lambda$ in $A(\pi)$. But, by the Perron–Frobenius theorem for irreducible matrices,

$$\text{mult}(\lambda) \leqq \text{mult}(1) = k+1.$$

Hence mult $(\lambda) = k+1$ and, by Perron–Frobenius, $\lambda$ is an eigenvalue of every $A_{\gamma\gamma}$ for which $\gamma$ is singular. It follows that $\lambda^q = 1$, where $q = q(\pi)$. Hence the conditions of Theorem 3.4 (ii) are satisfied and the lemma follows.

We state our main result.

THEOREM 5.10. *Let* $A$ *be nonzero* $n \times n$ *matrix normalized by the condition* $\rho(A) = 1$. *Assume* $1 \leqq i, j \leqq n$. *Let* $k = k[i,j]$ *be the singular distance from* $i$ *to* $j$ *and* $q = q[i,j]$ *be the local period of* $(i,j)$. *Put* $B^{(m)} = A^m(I + A + \cdots + A^{q-1})$. *Then* $b_{ij}^{(m)} \approx m^k$.

*Proof.* As usual, we assume that $A$ is in the Frobenius form (4.1). Suppose that $(i,j)$ is a position in $A_{\alpha\beta}$. Denote by $\Pi(\alpha,\beta)$ the set of all paths connecting $\alpha$ to $\beta$. Then we obviously have

$$A_{\alpha\beta}^{(m)} = \sum_{\pi\in\Pi(\alpha,\beta)} A(\pi)_{0s(\pi)}^{(m)}.$$

So

$$B_{\alpha\beta}^{(m)} = \sum_{\pi\in\Pi(\alpha,\beta)} B(\pi)_{0s(\pi)}^{(m)}.$$

Assume first that $k = k(\pi) = -\infty$; then, clearly, $B_{\alpha\beta}^{(m)} = A_{\alpha\beta}^{(m)} = 0$. If $k = -1 \geqq k(\pi)$ then Lemma 5.4 implies that each $A(\pi)_{0s(\pi)}^{(m)} \approx m^{-1}$. So $A_{\alpha\beta}^{(m)} \approx m^{-1}$ and again $A_{\alpha\beta}^{(m)} = B_{\alpha\beta}^{(m)}$.

Assume now that $k \geqq 0$. If $k > k(\pi)$, Lemma 5.4 implies that $B(\pi)_{0s(\pi)}^{(m)} = O(m^k)$. However, if $k = k(\pi)$, then according to Lemma 5.9 $\lim_{m\to\infty} m^{-k} B(\pi)_{0s(\pi)}^{(m)} = F_{0s}(\pi) > 0$ as $q(\pi)$ divides $q(\alpha, \beta) = q[i, j]$. By the definition of $k(\alpha, \beta)$ there exists $\pi \in \Pi(\alpha, \beta)$ such that $k(\pi) = k(\alpha, \beta)$. So $\lim_{k\to\infty} m^{-k} B_{\alpha\beta}^{(m)} = F_{\alpha\beta} > 0$. □

COROLLARY 5.11. *Under the conditions of Theorem* 5.10,

$$\sum_{p=1}^{m} a_{ij}^{(p)} \approx m^{(k+1)}.$$

*Proof.* For $k \geqq 0$, the result is immediate by Lemma 2.4. If $k = -1$, then by Theorem 5.10 the nonnegative series above converges. The assumption $k = -1$ implies that at least one term is positive. Finally if $k = -\infty$, $a_{ij}^{(p)} = 0$, $p = 1, 2, \cdots$, and the result follows. □

Comparing Theorems 3.4 and 5.10 we first deduce a local version of Rothblum's equality and then the equality itself.

THEOREM 5.12. *Let* $A \in \mathbb{R}_+^{nn}$ *where* $\rho(A) = 1$. *Assume that* $1 \leqq i, j \leqq n$; *then*

$$\operatorname{index}_{ij}(1) = k[i, j] + 1.$$

COROLLARY 5.13 (Rothblum [13]). *Let* $A \in \mathbb{R}_+^{nn}$ *where* $\rho(A) = 1$. *Then* $\operatorname{index}(1) = \max_{1\leqq i,j\leqq n} \operatorname{index}_{ij}(1) = \max_{1\leqq i,j\leqq n} k[i, j] + 1$.

**6. Convergent iterative methods for the spectral radius of a nonnegative matrix.** Let $A \in \mathbb{R}_+^{nn}$ and assume that $\rho(A) > 0$. Let $r(x)$ and $R(x)$ be defined as in (1.4). Clearly $0 \leqq r(x) \leqq R(x) \leqq +\infty$. It is obvious that

$$r(x) \leqq r(Ax) \leqq R(Ax) \leqq R(x).$$

So the sequence $r(A^m x)$, $m = 0, 1, \cdots$ is an increasing sequence bounded above by $R(x)$, and the sequence $R(A^m x)$, $m = 0, 1, \cdots$ is a decreasing sequence bounded below by $r(x)$.

In [4], Collatz observed that, for $A \in \mathbb{R}_+^{nn}$ and $x > 0$,

$$r(x) \leqq \rho(A) \leqq R(x), \tag{6.1}$$

and when $A$ is irreducible, this inequality is valid for all $x \geqq 0$, $x \neq 0$; see Wielandt [20], Varga [19, p. 32]. Thus the question arises when, for $A \geqq 0$ and $x \geqq 0$, $x \neq 0$,

$$\lim_{m\to\infty} r(A^m x) = \rho(A) = \lim_{m\to\infty} R(A^m x). \tag{6.2}$$

Wielandt's [20] characterization of $\rho(A)$ for irreducible $A$ easily implies that (6.2) holds for primitive $A$ and all $x \in \mathbb{R}_+^n$, $x \geqq 0$, $x \neq 0$ (cf. Varga [19, p. 34]). This result follows from the fact that

$$\lim_{m\to\infty} \rho(A)^{-m} A^m = Z > 0$$

when $A$ is primitive, where $Z = uv^t$, $v > 0$, $Au = \rho(A)u$, $v > 0$, $v^t A = \rho(A)v^t$, $v^t u = 1$. If $A$ is irreducible but imprimitive then (6.2) does not hold unless $x$ is orthogonal on all eigenvectors of $A^t$ corresponding to $\lambda$ such that $|\lambda| = \rho(A)$ and $\lambda \neq \rho(A)$. We shall show

that this condition can be put in equivalent forms. If $A$ is irreducible and of period $q$, then by simultaneous permutations of rows and columns we now put $A$ into the form

$$\begin{bmatrix} 0 & A_{12} & 0 & \cdots & 0 \\ 0 & 0 & A_{23} & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & A_{q-1,q} \\ A_{q1} & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{6.3}$$

where the diagonal blocks 0 are square (see Frobenius [6], Gantmacher [7, Vol II, p. 62], Berman–Plemmons [1, Chap. 2, Thm. 2.20]).

LEMMA 6.4. *Let $A$ be an irreducible nonnegative matrix of period $q$ in form* (6.3), *and suppose that $\rho(A)=1$. Let $v^tA=v^t$, $Au=u$, where $v>0$, $u>0$, $v^tu=1$, $A^ty^j=\omega^jy^j$, $j=1,\cdots,q-1$, $\omega=e^{2\pi i/q}$. Let $0\neq x\in\mathbb{R}_+^n$ be partitioned conformally with $A$, $x^t=(x_{(1)}^t,\cdots,x_{(q)}^t)$. Then the following are equivalent*

(i) $\lim_{m\to\infty}A^mx=(v^tx)u$,

(ii) $\lim_{m\to\infty}A^mx$ *exists*,

(iii) $x^ty^j=0, \quad j=1,\cdots,q-1$,

(iv) $v_{(1)}^tx_{(1)}=\cdots=v_{(q)}^tx_{(q)}$,

(v) $\lim_{m\to\infty}R(A^mx)=\lim_{m\to\infty}r(A^mx)=1$,

*where $v^t=(v_{(1)}^t,\cdots,v_{(q)}^t)$ has been partitioned conformally with $A$.*

*Proof.* We first derive a formula for $A^mx$, $m=1,2,\cdots$. Let $\omega$ be a primitive $q$th root of unity. It is well-known that the eigenvalues of $A$ on the unit circle are $\lambda_\alpha=\omega^{\alpha-1}$, $\alpha=1,\cdots,q$ and that each $\lambda_\alpha$ is a simple zero of the characteristic polynomial. It follows, in the notation of § 3, that $p_\alpha=0$, $\alpha=1,\cdots,q$ and that

$$Z^{(\alpha 0)}=D^{\alpha-1}uv^tD^{1-\alpha}, \qquad \alpha=1,\cdots,q,$$

$$y^\alpha=D^{(1-\alpha)}v, \qquad \alpha=1,\cdots,q,$$

where

$$D=\begin{bmatrix} I_{11} & & & \\ & \omega I_{22} & & 0 \\ & & \ddots & \\ 0 & & & \omega^{q-1}I_{qq} \end{bmatrix},$$

and $I_{\alpha\alpha}$ is an identity matrix of the same order of $A_{\alpha\alpha}$, $\alpha=1,\cdots,q$.

Hence by (3.1),

$$A^m=\sum_{\alpha=0}^{q-1}\omega^{m\alpha}D^\alpha uv^tD^{-\alpha}+o(1);$$

and so

$$A^mx=\sum_{\alpha=0}^{q-1}\omega^{m\alpha}a_\alpha(D^\alpha u)+o(1), \tag{6.5}$$

where

$$a_\alpha=v^tD^{-\alpha}x=x^ty^\alpha, \qquad \alpha=0,\cdots,q-1. \tag{6.6}$$

Let

$$c_\beta=v_{(\beta+1)}^tx_{(\beta+1)}, \qquad \beta=0,\cdots,q-1.$$

Then it follows immediately from (6.6) that

$$a_\alpha = \sum_{\beta=0}^{q-1} \omega^{-\alpha\beta} c_\beta, \qquad \alpha = 0, \cdots, q-1. \tag{6.7}$$

We now prove the equivalence of our five conditions. We show (i)$\Rightarrow$(ii)$\Rightarrow$(iii)$\Rightarrow$(iv)$\Rightarrow$(i) and (i)$\Rightarrow$(v)$\Rightarrow$(iv).

(i)$\Rightarrow$(ii). Trivial.

(ii)$\Rightarrow$(iii). Since $\lim_{m\to\infty} A^m x$ exists, $\lim_{m\to\infty} v^t D^{-\alpha} A^m x$ also exists, $\alpha = 0, \cdots, q-1$. But $v^t u > 0$, and hence $a_\alpha = x^t y^\alpha = 0$, $\alpha = 1, \cdots, q-1$ by Lemma 3.3.

(iii)$\Rightarrow$(iv). Consider the identity (6.7). Since the Vandermonde matrix $q^{-1/2}(\omega^{-\alpha\beta})$, $\alpha, \beta = 0, \cdots, q-1$ is unitary the assumption $a_\alpha = x^t y^\alpha = 0$, $\alpha = 1, \cdots, q-1$ implies that $c_0 = c_1 = \cdots = c_{q-1}$, which proves (iv).

(iv)$\Rightarrow$(i). If (iv) holds, then $c_0 = c_1 = \cdots = c_{q-1}$ and (6.7) implies $a_1 = \cdots = a_{q-1} = 0$. This establishes (i) in view of (6.5) and (6.6).

(i)$\Rightarrow$(v). Trivial, since $v^t x > 0$ and $u > 0$.

(v)$\Rightarrow$(i). Let $m = ql + r$, $0 \leqq r \leqq q-1$. Then (6.5) implies

$$\lim_{l\to\infty} A^{ql+r} x = \tilde{x}^{(r)}, \qquad r = 0, \cdots, q-1$$

for some $\tilde{x}^{(r)} \geqq 0$, $\tilde{x}^{(r)} \neq 0$. Also

$$A^r \tilde{x}^{(0)} = \tilde{x}^{(r)}, \qquad r = 0, \cdots, q-1, \qquad A^q \tilde{x}^{(0)} = \tilde{x}^{(0)}.$$

As $A^q$ is a direct sum of $q$ irreducible and primitive matrices the assumption $x \geqq 0$, $x \neq 0$ implies that $\lim_{l\to\infty} (A^q)^l x = \tilde{x}^{(0)} \neq 0$. Obviously $x^{(0)} \geqq 0$.

Now (v) implies that

$$x^0 \leqq x^{(1)} = Ax^0 \leqq x^{(0)},$$

whence $x^{(1)} = x^{(0)}$ and thus $x^{(r)} = x^{(0)}$ for $r = 1, \cdots, q-1$. So $\lim_{m\to\infty} A^m x = x^{(0)}$ and (i) follows. □

In what follows, we give necessary and sufficient conditions on a reducible matrix $A$ to satisfy (6.2). To do so we need a few more graph theoretical concepts.

Let $G$ be a graph on $\langle\nu\rangle = \{1, \cdots, \nu\}$. Let $J$ be a nonvoid subset of $\langle\nu\rangle$. Then $\alpha \in J$ is called a final state with respect to $J$ if for any $\beta \neq \alpha$ and $(\alpha, \beta\rangle \in G$, $\beta \notin J$. Denoting by $\mathcal{F}(J)$ the set of all final states with respect to $J$. If $J = \langle\nu\rangle$, then $\alpha$ is called a final state, i.e., $(\alpha, \beta) \in G$ implies that $\beta = \alpha$. Define

$$d(\beta, J) = \max\{k(\beta, \alpha) : \alpha \in \mathcal{F}(J)\}.$$

If $J = \langle\nu\rangle$, then write $d(\beta)$ instead of $d(\beta, \langle\nu\rangle)$. Let $A \geqq 0$ be a reducible matrix. We assume that $A$ is in the Frobenius form (4.1).

As in § 4, denote by $G(A)$ the (reduced) graph of $A$. Let $x \geqq 0$, $x \neq 0$. Partition $x$ conformably with $A$ given by (4.1). That is $x^t = (x_{(1)}^t, \cdots, x_{(\nu)}^t)$. The support of $x$ is the set supp $x = \{\alpha_1, \cdots, \alpha_s\} \subseteq \{1, \cdots, \nu\}$ such that $x_{(i)} \neq 0$ if and only if $i \in$ supp $x$. We shall always assume that $\alpha_i$, $i = 1, \cdots, s$ have been listed in strictly ascending order.

THEOREM 6.8. *Let $A \in \mathbb{R}_+^{nn}$, $\rho(A) > 0$. Assume that $A$ is in the Frobenius form* (4.1). *Moreover, if $A_{ii}$ is imprimitive then $A_{ii}$ is the Frobenius form* (6.3). *Let $x \geqq 0$, $x \neq 0$. Then* (6.2) *holds if and only if any final state $\alpha$ with respect to the support of $x$ satisfies*

(i) *$\alpha$ is a singular vertex (i.e., $\rho(A_{\alpha\alpha}) = \rho(A)$),*

(ii) either *$A_{\alpha\alpha}$ is primitive,* or *$A_{\alpha\alpha}$ and $x_{(\alpha)}$ satisfy the condition* (iv) *of Lemma* 6.4.

*Proof.* Without loss of generality we may assume that $\rho(A) = 1$. Next we note that

$$(A^m x)_\alpha = \sum_{\beta \in \operatorname{supp} x} A^{(m)}_{\alpha\beta} x_{(\beta)}. \tag{6.9}$$

Suppose that $\alpha \in \mathscr{F}(\operatorname{supp} x)$. Then

$$(A^m x)_\alpha = A^m_{\alpha\alpha} x_{(\alpha)}.$$

By the definition of $R(x)$ and $r(x)$ we have

$$r(A^m x) A^m x \leqq A^{m+1} x \leqq R(A^m x) A^m x.$$

So

$$r(A^m x) A^m_{\alpha\alpha} x_{(\alpha)} \leqq A^{m+1}_{\alpha\alpha} x_{(\alpha)} \leqq R(A^m x) A^m_{\alpha\alpha} x_{(\alpha)}.$$

Hence, since $A_{\alpha\alpha}$ is irreducible, by (6.1),

$$r(A^m x) \leqq r(A^m_{\alpha\alpha} x_{(\alpha)}) \leqq \rho(A_{\alpha\alpha}) \leqq R(A^m_{\alpha\alpha} x_{(\alpha)}) \leqq R(A^m x).$$

Assume now that (6.2) holds. Then for any final state $\alpha$ with respect to supp $x$, we must have

$$\lim_{m\to\infty} r(A^m_{\alpha\alpha} x_{(\alpha)}) = \lim_{m\to\infty} R(A^m_{\alpha\alpha} x_{(\alpha)}) = \rho(A_{\alpha\alpha}) = 1.$$

So $\alpha$ is a singular vertex. If $A_{\alpha\alpha}$ is imprimitive, then the condition (v) of Lemma 6.4 holds. Hence, $A_{\alpha\alpha}$ and $x_{(\alpha)}$ satisfy (iv) of Lemma 6.4. This proves one direction of our theorem.

Assume now that if $\alpha \in \mathscr{F}(\operatorname{supp} x)$ then $\rho(A_{\alpha\alpha}) = 1$; and if $A_{\alpha\alpha}$ is not primitive then $A_{\alpha\alpha}$ and $x_{(\alpha)}$ satisfy the condition (iv) of Lemma 6.4.

Let $1 \leqq \beta \leqq \nu$. Let $d = d(\beta, J)$. By our assumption, $d \neq -1$. If $d = -\infty$, then $(A^m x)_\beta = 0$, $m = 1, 2, \cdots$. If $d \geqq 0$, then

$$m^{-d}(A^m x)_\beta = m^{-d} \sum_{\alpha \in K} A^{(m)}_{\beta\alpha} x_\alpha + o(1),$$

where $K = \{\alpha : k(\beta, \alpha) = d\}$. Clearly $K \subseteq \mathscr{F}(\operatorname{supp} x)$. Thus, to show

$$\lim_{m\to\infty} m^{-d}(A^m x)_\beta > 0, \tag{6.10}$$

it is enough to prove

$$m^{-d} A^{(m)}_{\beta\alpha} x_\alpha > 0 \tag{6.11}$$

for $\alpha \in \mathscr{F}(\operatorname{supp} x)$, $k(\beta, \alpha) = d$. To prove (6.11), let $D$ be the matrix obtained from $A$ by setting $D_{\alpha\alpha} = 0$ and $D_{\gamma\delta} = A_{\gamma\delta}$ in all other cases, $1 \leqq \gamma, \delta \leqq \nu$. We then have

$$m^{-d} A^{(m)}_{\beta\alpha} x_\alpha = m^{-d} \sum_{p=0}^{m} D^{(m-p)}_{\beta\alpha} A^p_{\alpha\alpha} x_\alpha.$$

Since in $D$, the singular distance from $\beta$ to $\alpha$ is $d-1$, we have, by Corollary 5.11,

$$\lim_{m\to\infty} m^{-d} \sum_{p=0}^{m} D^{(m-p)}_{\beta\alpha} = U_{\beta\alpha} > 0,$$

and by Lemma (6.4)

$$\lim_{p\to\infty} A^p_{\alpha\alpha} x_\alpha = v_\alpha > 0.$$

It easily follows from Lemma 2.7 that

$$\lim_{m\to\infty} m^{-d} A_{\beta\alpha}^{(m)} x_\alpha = \frac{1}{d} U_{\beta\alpha} v_\alpha > 0.$$

Thus, for each $\beta$, $1 \leqq \beta \leqq \nu$, either $(A^m x)_\beta = 0$, $m = 1, 2, \cdots$ or (6.10) is satisfied. From this (6.2) follows immediately. □

COROLLARY 6.12. *Let $A \in \mathbb{R}_+^{nn}$, $\rho(A) > 0$. Assume that $A$ is the Frobenius form* (4.1). *Let $J$ be an nonempty set of $\langle \nu \rangle$. Then for any $x \geqq 0$ whose support is the set $J$,* (6.2) *holds if and only if for all final states $\alpha$ with respect to $J$, $\rho(A_{\alpha\alpha}) = \rho(A)$ and $A_{\alpha\alpha}$ is primitive.*

COROLLARY 6.13. *Let $A \in \mathbb{R}_+^{nn}$, $\rho(A) > 0$. Assume that $A$ is in the Frobenius form* (4.1). *Then for any $x \geqq 0$, $x \neq 0$,* (6.2) *holds if and only if for each $\alpha$, $\alpha = 1, \cdots, \nu$, $\rho(A_{\alpha\alpha}) = \rho(A)$ and $A_{\alpha\alpha}$ is primitive.*

**7. Nonnegative solutions of $(I - A)y = x$.** As an application of our results we give a simple proof of a theorem concerning nonnegative solutions $y$ of $(I - A)y = x$ for given $x \geqq 0$. For $1 \leqq \alpha, \beta \leqq \nu$ we shall say that $\beta$ has *access* to $\alpha$ in $G(A)$ if there is a path from $\beta$ to $\alpha$ in $G(A)$, viz., $k(\beta, \alpha) \geqq -1$.

THEOREM 7.1. *Let $A \in \mathbb{R}_+^{nn}$ with $\rho(A) = 1$, and suppose that $A$ is in the Frobenius normal form* (4.1). *Let $x \in \mathbb{R}_+^n$. Then the following are equivalent*:

(i) *there is a $y \in \mathbb{R}_+^n$ such that $(I - A)y = x$;*

(ii) *no singular vertex $\beta$ has access in $G(A)$ to any $\alpha \in \operatorname{supp} x$;*

(iii) $\lim_{m\to\infty} (I + \cdots + A^m)x$ *exists*;

(iv) $\lim_{m\to\infty} A^m x = 0$.

*Further, if* (iii) *holds and $y = \lim_{m\to\infty} (I + A + \cdots + A^m)x$, then $(I - A)y = x$ and*

(7.2) $\qquad y_\beta = 0$ *if $\beta$ does not have access to any $\alpha \in \operatorname{supp} x$,*

(7.3) $\qquad y_\beta > 0$ *if $\beta$ has access to some $\alpha \in \operatorname{supp} x$.*

*Proof.* Let $S^{(m)} = I + A + \cdots + A^m$. If $1 \leqq \beta \leqq \nu$, then

$$(S^{(m)} x)_\beta = \sum_{\alpha \in \operatorname{supp} x} S_{\beta\alpha}^{(m)} x_\alpha, \tag{7.4}$$

and, by Corollary 5.11, for $k = k(\beta, \alpha) \geqq -1$,

$$\lim_{m\to\infty} m^{-(k+1)} S_{\beta\alpha}^{(m)} = U_{\beta\alpha} > 0; \tag{7.5i}$$

while for $k(\beta, \alpha) = -\infty$,

$$S_{\beta\alpha}^{(m)} = U_{\beta\alpha} = 0, \qquad m = 1, 2, 3, \cdots. \tag{7.5ii}$$

We shall prove (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i), (iii) $\Rightarrow$ (iv) $\Rightarrow$ (ii).

(i) $\Rightarrow$ (ii). Suppose that $(I - A)y = x$, where $y \geqq 0$. Then

$$S^{(m)} x = (I - A^{m+1})y \leqq y.$$

Let $\beta$ be a singular vertex. If $\beta$ has access to $\alpha$, then $k = k(\beta, \alpha) \geqq 0$ and, by (7.4) and (7.5),

$$y_\beta \geqq (S^{(m)} x)_\beta \geqq \tfrac{1}{2} m^{(k+1)} U_{\beta\alpha} x_\alpha$$

for large $m$. Hence $x_\alpha = 0$ and $\alpha \notin \operatorname{supp} x$.

(ii) $\Rightarrow$ (iii). Suppose (ii) holds and let $1 \leqq \beta \leqq \alpha$.

If $\alpha \in \operatorname{supp} x$, then $k = k(\beta, \alpha) = -1$, or $k = -\infty$. Hence, $\lim_{m\to\infty} S^{(m)}_{\beta\alpha} x_\alpha = U_{\beta\alpha} x_\alpha$ exists for $\alpha \in \operatorname{supp} x_\alpha$. So, by (7.5), $\lim_{m\to\infty} S^{(m)} x$ exists.

(iii)$\Rightarrow$(i). Let $y = \lim_{m\to\infty} S^{(m)} x$. Clearly $y \geqq 0$. Since $AS^{(m)} x = S^{(m+1)} x - x$, $y$ satisfies $(I - A)y = x$. This proves (i).

(iii)$\Rightarrow$(iv). Trivial.

(iv)$\Rightarrow$(ii). Suppose that (iv) holds but that (ii) is false. Then there exists a singular $\beta$ and an $\alpha \in \operatorname{supp} x$ such that $k(\beta, \alpha) \geqq 0$. Let $q = q(\beta, \alpha)$ be the local period and let $B^{(m)} = A^m(I + \cdots + A^{q-1})$. Then $\lim_{m\to\infty} B^{(m)} x = 0$. But by Theorem 5.10 for all sufficiently large $m$,

$$(B^{(m)} x)_\beta \geqq B^{(m)}_{\beta\alpha} x_\alpha \geqq c m^k x_\alpha,$$

where $c > 0$, and $x_\alpha \neq 0$. This is a contradiction, and the implication is proved.

To complete the proof of the theorem observe that, for $y = \lim_{m\to\infty} S^{(m)} x$,

$$y_\beta = \sum_{\alpha \in \operatorname{supp} x} U_{\beta\alpha} x_\alpha$$

in view of (ii) and (7.5). Since $U_{\beta\alpha} > 0$, if $\beta$ has access to $\alpha$ and $U_{\beta\alpha} = 0$ otherwise, we immediately obtain (7.2) and (7.3). □

The equivalence of conditions (i) and (ii) in Theorem 7.1 is due to D. H. Carlson [3]. We remark that Carlson also showed that if a nonnegative solution $y$ of $(I - A)y = x$ exists, then the solution satisfying (7.2) and (7.3) is unique. It should be observed that the assumption that $A$ is in Frobenius normal form is not needed for conditions (i), (iii) and (iv) of Theorem 7.1, which may easily be proved equivalent directly. Conditions (iii) and (iv) are equivalent for general $A \in \mathbb{R}^{nn}$ and $x \in \mathbb{R}^n$. We observe that for

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

there is a $y \in \mathbb{R}^n$ such that $(I - A)y = x$; yet the equivalent conditions (ii), (iii) and (iv) do not hold. Clearly, no $y$ satisfying $(I - A)y = x$ can be nonnegative.

## REFERENCES


[1] A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Academic Press, New York, 1979.

[2] R. A. Brualdi, *Introductory Combinatorics*, North Holland, Amsterdam, 1977.

[3] D. H. Carlson, *A note on M-matrix equations*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 1027–1033.

[4] L. Collatz, *Einschliessungssatz für die charakteristischen Zahlen von Matrizen*, Math. Z., 48 (1942), pp. 221–226.

[5] S. Friedland, *On an inverse problem for nonnegative matrices and eventually nonnegative matrices*, Israel J. Math., 29 (1978), pp. 43–60.

[6] G. F. Frobenius, *Über Matrizen aus nicht negativen Elementen*, S. B. Kön. Preuss. Akad. Wiss. Berlin, (1912), pp. 456–477; Gesammelte Abhandlungen, vol. 3, Springer, Berlin, 1968, pp. 546–567.

[7] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[8] G. H. Hardy, *Divergent Series*, Clarendon, Oxford, England, 1949.

[9] S. Karlin, *Positive operators*, J. Math. Mech., 8 (1959), pp. 907–937.

[10] C. D. Meyer and R. J. Plemmons, *Convergent powers of a matrix with applications to iterative methods for singular systems*, SIAM J. Numer. Anal., 14 (1977), pp. 699–705.

[11] A. Ostrowski, *Über die Determinanten mit überwiegender Hauptdiagonale*, Comment. Math. Helv., 10 (1937), pp. 69–96.

[12] D. Richman and H. Schneider, *On the singular graph and the Weyr characteristic of an M-matrix*, Aequationes Math., 17 (1978), pp. 208–234.

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

[14] U. G. ROTHBLUM, *Expansions of sums of matrix powers and resolvents,* Yale Univ. Rep., New Haven, CT, revised Jan. 1977; SIAM Rev., to appear.
[15] ———, *Sensitive growth analysis of multiplicative systems I: The dynamic approach,* Yale Univ. Rep., New Haven, CT, revised Jan. 1977.
[16] H. H. SCHAEFER, *Topological Vector Spaces,* Macmillan, New York, 1964.
[17] H. H. SCHAEFER, *Banach Lattices and Positive Operators,* Springer, New York, 1974.
[18] E. C. TITCHMARSH, *The Theory of Functions,* 2nd ed., Oxford University Press, London, 1939.
[19] R. S. VARGA, *Matrix Iterative Analysis,* Prentice Hall, Englewood Cliffs, NJ, 1962.
[20] H. WIELANDT, *Unzerlegbare, nicht negative Matrizen,* Math. Z., 52 (1950), pp. 642–648.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0007 $01.00/0

# DISCOUNTED STOCHASTIC RATIO GAMES*

V. AGGARWAL†, R. CHANDRASEKARAN‡ AND K. P. K. NAIR†

**Abstract.** In a recent work, the authors considered a finite state Markov ratio decision process in which the objective was to maximize the ratio of total discounted rewards. In this paper, discounted Markov ratio decision processes are generalized to discounted stochastic ratio games. These may also be viewed as generalizations of ratio games to a stochastic context where the payoff is the ratio of the two total discounted rewards. We show that in the discounted stochastic ratio game the players have stationary optimal strategies with a unique value. The solution may depend on the initial probability distribution. We also provide a convergent algorithm.

**1. Introduction.** In a recent work [1], we considered a finite state discounted Markov ratio decision process in which the objective was to maximize the ratio of total discounted rewards over an infinite horizon. We have shown that in this process, an optimal policy exists and the value is unique. The optimal policy is stationary and pure; but the solution may depend on the initial probability distribution. We also gave an algorithm for computing the solution.

In this paper, finite state discounted Markov ratio decision processes (abbreviated as DMRDP) are generalized to discounted stochastic ratio games (abbreviated as DSRG). The mechanics of motion of this game is similar to that of the stochastic game of Shapley [7]; but the payoff function, being a ratio, is different. Thus, in each state each of the two players has a finite number of alternative actions. The players' actions and the state jointly determine two rewards and probabilities of transition to the next state. Thus, there are two sequences of rewards. These sequences are discounted so that each discounted sequence sums to a finite quantity. In DSRG, the payoff function is the ratio of these two finite quantities.

We show that the players have stationary optimal strategies and the game has a unique value. In general, these are dependent on the probability distribution over the initial state. We also give a convergent algorithm for computing the solution.

The discounted stochastic ratio games may also be viewed as generalizations of two other interesting models. These, respectively, are the generalizations of ratio game [6] to a stochastic context, where the payoff is defined as the ratio of the two total discounted rewards, and Markov renewal process [5] to Markov renewal games. Of course, in the latter case one has to replace discounting by the assumption that the game is stopping. Here the two sequences, respectively, are one of rewards and the other of durations of stay in the states before the next transition. Thus the payoff is the ratio of the total expected reward to the total expected time to termination, that is, the average reward before termination.

An interesting application of the model may be possible in economic equilibrium theory. Von Neumann [8] developed a model of general economic equilibrium theory considering a static situation in which there are certain goods, and technologies for producing the goods. An optimal choice of goods and technologies are obtained by solving a ratio game. The results of the present work may be helpful in developing a dynamic economic equilibrium model considering a finite state space in which the economy makes probabilistic transitions.

* Received by the editors January 31, 1978 and in final revised form November 27, 1979.

† School of Administration, University of New Brunswick, Fredericton, New Brunswick E3B 5A3, Canada.

‡ School of Management and Administration, University of Texas at Dallas, Richardson, Texas 75080.

In this paper the DSRG is developed as a generalization of DMRDP, and therefore the analysis depends heavily on the results in [1]. Therefore, the relevant results from [1] are summarized in the next section.

**2. Summary of results from DMRDP.** In [1] we have considered a Markov ratio decision process in which the objective was to maximize the ratio of the total discounted rewards. The process is observed periodically at time points $n = 0, 1, 2, \cdots$, and at each time it will be in any state $i$, $i = 1, 2, \cdots, N$ of the set $S$. However, in state $i$ there is a finite set $C_i$ of $K_i$ alternatives of actions numbered $1, 2, \cdots, k, \cdots, K_i$. If action $k \in C_i$ is taken while in state $i$, the system generates two finite rewards $a_i^k$ and $b_i^k$ where $b_i^k > 0$ for all $k \in C_i$, and the system moves to state $j$ in the next step with probability $p_{ij}^k$. The rewards occurring in future periods are discounted by a factor $\beta$, $0 \leqq \beta < 1$, per period. In this process a stationary strategy is an $N$-tuple of probability vectors $x$ so that

$$x = (x_1, x_2, \cdots, x_i, \cdots, x_N),$$

where

$$x_i = (x_i^1, x_i^2, \cdots, x_i^k, \cdots, x_i^{K_i}), \tag{1}$$

and

$$\sum_{k \in C_i} x_i^k = 1, \qquad i \in S.$$

Here $x_i^k$ is the probability of using action $k \in C_i$ given that the process is in state $i$. A stationary strategy is pure if for each $x_i$, $x_i^{\hat{k}} = 1$ for exactly one $\hat{k} \in C_i$ and $x_i^k = 0$ for all $\hat{k} \neq k \in C_i$. We have shown that in the DMRDP one has a stationary and pure optimal strategy with a unique value, and these may depend on $\alpha$, the probability distribution over the initial state. Further, given $\alpha$ the DMRDP has a linear programming formulation. Given $\alpha$ and a maximization objective for the DMRDP, the dual program in the variables $r$ and $s = (s_1, s_2, \cdots, s_i, \cdots, s_N)$ is as follows:

$$\begin{aligned} &\text{Min } r \quad \text{subject to} \\ &\sum_{i=1}^{N} \alpha_i s_i \leqq 0, \\ &a_i^k - b_i^k r + \beta \sum_{j=1}^{N} P_{ij}^k s_j - s_i \leqq 0 \quad \text{for all } k \in C_i \text{ and } i \in S, \end{aligned} \tag{2}$$

where $r$ and $s_i (i = 1, 2, \cdots, N)$ are unrestricted.

THEOREM 1. *For a given $\alpha$, if $(\bar{r}, \bar{s})$ is optimal in* (2), *then it is the unique solution to*:

$$\max_{k \in C_i} (a_i^k - b_i^k r + \beta \sum_{j=1}^{N} p_{ij}^k s_j - s_i) = 0, \qquad i = 1, 2, \cdots, N, \tag{3}$$

*and*

$$\sum_{i=1}^{N} \alpha_i s_i = 0. \tag{4}$$

*Proof.* Theorem 6.1 of [1] shows that $(\bar{r}, \bar{s})$ satisfies (3) and (4). Therefore, what remains to be shown is the uniqueness of the solution to (3)–(4). Now suppose there are two solutions $(\bar{r}, \bar{s})$ and $(\hat{r}, \hat{s})$ attained, respectively, at $\bar{k} \in C_i$ and $\hat{k} \in$

$C_i (i = 1, 2, \cdots, N)$. Then,

$$a_i^{\bar{k}} - b_i^{\bar{k}} \bar{r} + \beta \sum_{j=1}^{N} p_{ij}^{\bar{k}} \bar{s}_j - \bar{s}_i = 0, \qquad 1, 2, \cdots, N, \tag{5}$$

$$\sum_{i=1}^{N} \alpha_i \bar{s}_i = 0, \tag{6}$$

$$a_i^{\hat{k}} - b_i^{\hat{k}} \hat{r} + \beta \sum_{j=1}^{N} p_{ij}^{\hat{k}} \hat{s}_j - \hat{s}_i = 0, \qquad i = 1, 2, \cdots, N, \tag{7}$$

and

$$\sum_{i=1}^{N} \alpha_i \hat{s}_i = 0. \tag{8}$$

Also we have

$$a_i^{\hat{k}} - b_i^{\hat{k}} \bar{r} + \beta \sum_{j=1}^{N} p_{ij}^{\hat{k}} \bar{s}_j - \bar{s}_i \leqq 0, \qquad i = 1, 2, \cdots, N, \tag{9}$$

$$a_i^{\bar{k}} - b_i^{\bar{k}} \hat{r} + \beta \sum_{j=1}^{N} p_{ij}^{\bar{k}} \hat{s}_j - \hat{s}_i \leqq 0, \qquad i = 1, 2, \cdots, N, \tag{10}$$

and

$$\sum_{i=1}^{N} \alpha_i (\bar{s}_i - \hat{s}_i) = 0. \tag{11}$$

Setting $\Delta s = \bar{s} - \hat{s}$ and $\Delta r = \bar{r} - \hat{r}$, subtraction of (5) from (10) gives

$$(I - \beta \bar{P}) \Delta s \leqq -\Delta r \bar{b}, \tag{12}$$

where

$$\bar{P} = \{p_{ij}^{\bar{k}}\} \quad \text{and} \quad \bar{b} = (b_i^{\bar{k}}, i = 1, 2, \cdots, N).$$

Similarly, if (9) is subtracted from (7) we get

$$(I - \beta \hat{P}) \Delta s \geqq -\Delta r \hat{b}. \tag{13}$$

Since $b_i^k > 0$ for all $i$ and $k$, clearly $\bar{b}, \hat{b} \gg 0$ where $\gg$ denotes strict inequality for each component. Further, $(I - \beta \bar{P})$ is invertible with $(I - \beta \bar{P})^{-1} = \sum_{i=0}^{\infty} (\beta \bar{P})^i \geqq I$. Therefore, if $\Delta r > 0$, from (12) we have $\Delta s \ll 0$. On the other hand, if $\Delta r < 0$, from (13) we obtain $\Delta s \gg 0$. But both these results violate (11), and therefore $\Delta r = 0$. Setting $\Delta r = 0$ in (12) and (13), respectively, we see that $\Delta s \leqq 0$ and $\Delta s \geqq 0$. Therefore $\Delta s = 0$, and thus the solution to the system of relations (3) and (4) is unique. We remark here that this property was, indeed, used in [1] for validating the algorithm; however, it has not been established there.

Now we consider another DMRDP in the same state space $S$ in which the objective is one of minimizing the ratio of two total discounted rewards. This will be helpful later in the paper. In this process let $D_i$ be a finite set of $L_i$ alternatives of actions numbered $1, 2, \cdots, l, \cdots L_i$, available in state $i \in S$. If action $l \in D_i$ is taken while in state $i$ the two rewards obtained are $a_i^l$ and $b_i^l$, $b_i^l > 0$ for all $l \in D_i$, and the system moves to state $j$ at the next step with probability $p_{ij}^l$. A stationary strategy in this process is an $N$-tuple of probability vectors $y$ so that

$$y = (y_1, y_2, \cdots, y_i, \cdots, y_N),$$

where

$$y_i = (y_i^1, y_i^2, \cdots, y_i^l, \cdots, y_i^{L_i}), \tag{14}$$

and

$$\sum_{l \in D_i} y_i^l = 1, \qquad i \in s.$$

For this decision process, the dual program in the variables $r$ and $s = (s_1, s_2, \cdots, s_i, \cdots, s_N)$ is as follows:

$$\begin{aligned} &\text{Max } r \quad \text{subject to} \\ &\sum_{i=1}^{N} \alpha_i s_i \geqq 0, \\ &a_i^l - b_i^l r + \beta \sum_{j=1}^{N} p_{ij}^l s_j - s_i \geqq 0, \qquad \text{for all } l \in D_i \text{ and } i \in S, \end{aligned} \tag{15}$$

where $r$ and $s_i$ $(i = 1, 2, \cdots N)$ are unrestricted. Now Theorem 2 below is readily obtained by analogous considerations.

THEOREM 2. *For given $\alpha$, if $(\underline{r}, \underline{s})$ is optimal in* (15), *then it is the unique solution to*:

$$\min_{l \in D_i} (a_i^l - b_i^l r + \beta \sum_{j=1}^{N} p_{ij}^l s_j - s_i) = 0, \qquad i = 1, 2, \cdots N, \tag{16}$$

*and*

$$\sum_{i=1}^{N} \alpha_i s_i = 0. \tag{17}$$

**3. Description of DSRG.** A finite state stochastic ratio game is a generalization of the finite state discounted Markov ratio decision process to a game context. There are two players in this generalized process, and a finite set of states $S$. At each time point the game is found in one of the states $i \in S$. While in state $i$, Players I and II, respectively, have the sets $C_i$ and $D_i$ of alternatives of actions described in § 2 above. If the players respectively choose actions $k \in C_i$ and $l \in D_i$ while in state $i$, there are two finite rewards $a_i^{kl}$ and $b_i^{kl}$, $b_i^{kl} > 0$, and the game moves to state $j$ at the next step with probability $p_{ij}^{kl}$. Since the DSRG is assumed to be a nonterminating one,

$$\sum_{j=1}^{N} p_{ij}^{kl} = 1, \qquad i \in S, \quad k \in C_i, \quad l \in D_i, \tag{18}$$

$$p_{ij}^{kl} \geqq 0, \qquad i, j \in S, \quad k \in C_i, \quad l \in D_i. \tag{19}$$

Both the rewards $a_i^{kl}$ and $b_i^{kl}$ occurring in future periods are discounted by a factor $\beta$, $0 \leqq \beta < 1$, per period. Let the reward matrices in state $i$ be denoted by $A_i$ and $B_i$ so that $A_i = \{a_i^{kl} | k \in C_i, l \in D_i\}$ and $B_i = \{b_i^{kl} | k \in C_i, l \in D_i\}$.

Now, given that the game starts in a specified state the evolution of DSRG may be represented by

$$\{i_n, k_n, l_n\}, \qquad n = 0, 1, 2, \cdots, \tag{20}$$

where $i_n$ is the state occupied at step or period $n$, and $k_n \in C_{i_n}$ and $l_n \in D_{i_n}$ are the moves chosen by the players respectively. Let the choices of moves $k_n \in C_{i_n}$ and $l_n \in D_{i_n}$ be, respectively, denoted by probability vectors $\Delta_n$ and $\Sigma_n$ defined over the sets $C_{i_n}$ and $D_{i_n}$. Clearly, in each of these vectors a single element would be 1 and the rest zeros. The

payoff function in the DSRG is defined as

$$r_{i_0} = \Big(\sum_{n=0}^{\infty} \beta^n \Delta_n A_{i_n} \Sigma_n\Big)/(\Sigma_n)/\Big(\sum_{n=0}^{\infty} \beta^n \Delta_n B_{i_n} \Sigma_n\Big). \tag{21}$$

In the DSRG for each player a pure strategy is defined as a sequence of moves, one in each period $n = 0, 1, 2, \cdots$, and a mixed strategy, in general, is a probability distribution over the set of all pure strategies. A behavioral strategy is a special type of mixed strategy and it is characterized by the property that the player makes an independent randomization at each move, rather than the entire sequence of moves by a single randomization. A stationary strategy is a behavioral strategy but the randomizations in the states are independent of the step or time period $n$. A strategy is optimal for Player I if and only if it guarantees at least the minimax value of the payoff function whatever be the strategy followed by Player II. Similarly, an optimal strategy for Player II may be defined.

Since the DSRG is a game of perfect recall we note from the work of Aumann [2] that, without loss of generality, the players can be restricted to behavioral strategies. Given $\alpha$, it will be shown that the DSRG has a minimax solution and the players have stationary optimal strategies.

**4. Existence of minimax solution.** Let $T$ and $U$, respectively, be the sets of all behavioral strategies for Players I and II. The set of all stationary strategies for the players are denoted by $X$ and $Y$ respectively. Clearly $X \subset T$ and $Y \subset U$. A stationary strategy $x \in X$ for Player I is a $N$-tuple of probability vectors having the same representation as specified by (1) above. Similarly, $y \in Y$ for Player II, is as shown above in (14).

THEOREM 3. *Given an initial probability vector, the DSRG has a minimax solution with a unique value and the players have stationary optimal strategies.*

*Proof.* Given an initial probability vector $\alpha$, let $r(t, u)$ be the payoff function if the players follow strategies $t \in T$ and $u \in U$ respectively. Now suppose Player II is fixed at a stationary strategy $y \in Y$ and Player I knows this information. Here Player I faces a DMRDP [1] in which the objective is one of maximization and the input data are

$$\begin{aligned} a_i^k &= a(y)_i^k \triangleq \sum_{l \in D_i} a_i^{kl} y_i^l, \\ b_i^k &= b(y)_i^k \triangleq \sum_{l \in D_i} b_i^{kl} y_i^l, \quad \text{and} \\ p_{ij}^k &= p(y)_{ij}^k \triangleq \sum_{l \in D_i} p_{ij}^{kl} y_i^l. \end{aligned} \tag{22}$$

The optimal value of this problem is

$$\bar{r}(y) = \operatorname*{Max}_{t \in T} r(t, y). \tag{23}$$

Since in the DMRDP [1] one has a stationary optimal strategy

$$\bar{r}(y) = \operatorname*{Max}_{x \in X} r(x, y), \tag{24}$$

and $\bar{r}(y)$ can be computed by solving linear program (2) with the input data given by (22) above. The unique optimal solution to this program will be denoted by $(\bar{r}(y), \bar{s}(y))$. Clearly $Y$ is compact and the input data depend continuously on $y_i^l$. Hence from Hoffman and Karp [4], it follows that $\bar{r}(y)$ is a continuous function of $y$ and it assumes a

minimum at least at one $y \in Y$. Thus

$$\bar{r}^* \triangleq \operatorname*{Min}_{y \in Y} \operatorname*{Max}_{x \in X} r(x, y) = \operatorname*{Min}_{y \in Y} \bar{r}(y) \tag{25}$$

is well defined. A similar consideration with Player I leads to

$$\underline{r}^* \triangleq \operatorname*{Max}_{x \in X} \operatorname*{Min}_{y \in Y} r(x, y) = \operatorname*{Max}_{x \in X} \underline{r}(x), \tag{26}$$

where $(\underline{r}(x), \underline{s}(x))$ is the unique solution to (15) with the appropriate input data. Now from (25) and (26) we have

$$\underline{r}^* = \operatorname*{Max}_{x \in X} \underline{r}(x) \leqq \operatorname*{Min}_{y \in Y} \bar{r}(y) = \bar{r}^*. \tag{27}$$

LEMMA 1. *If* $\bar{r}(y^*) = \bar{r}^*$, *then* $(\bar{r}(y^*), \bar{s}(y^*))$ *satisfies*

$$\operatorname*{Max}_{k \in C_i} \sum_{l \in D_i} (a_i^{kl} - b_i^{kl} r + \beta \sum_{j=1} p_{ij}^{kl} s_j - s_i) y_i^l \geqq 0, \qquad i = 1, 2, \cdots, N, \quad \forall y \in Y, \tag{28}$$

*and*

$$\sum_{i=1}^{N} \alpha_i s_i = 0, \tag{29}$$

*where for* $y = y^*$ *the inequality in* (28) *is an equality for all* $i$. *Similarly, if* $\underline{r}(x^*) = \underline{r}^*$, *then* $(\underline{r}(x^*), \underline{s}(x^*))$ *satisfies*

$$\operatorname*{Min}_{l \in D_i} \sum_{k \in C_i} (a_i^{kl} - b_i^{kl} r + \beta \sum_{j=1}^{N} p_{ij}^{kl} s_j - s_i) x_i^k \leqq 0, \qquad i = 1, 2, \cdots, N, \quad \forall x \in X, \tag{30}$$

*and*

$$\sum_{i=1}^{N} \alpha_i s_i = 0, \tag{31}$$

*where for* $x = x^*$ *the inequality in* (30) *is an equality for all* $i$.

*Proof.* We only prove the first part of the lemma, as the second part follows from similar arguments. From Theorem 1 it follows that for $y = y^*$ (28) and (29) are satisfied as equalities by $(\bar{r}(y^*), \bar{s}(y^*))$ uniquely. Therefore it remains to be shown that for each $y \neq y^*$ (28) holds for all $i$. This we prove by contradiction. Suppose for $\hat{y} \neq y^*$ (28) is violated at least for one $i$ when $\bar{r}(y^*)$ and $\bar{s}(y^*)$, respectively, are substituted for $r$ and $s$. Let $\tilde{y}_i = \hat{y}_i$ for $i$ for which (28) is violated and $\tilde{y}_j = y_j^*$ otherwise. Then from Theorem 1, it is clear that $(\bar{r}(y^*), \bar{s}(y^*))$ is feasible (but not optimal) in (2) with the input data corresponding to the case where Player II is fixed at $\tilde{y}$. Therefore $\bar{r}(y^*) > \bar{r}(\tilde{y})$ and this reveals the contradiction $\bar{r}(y^*) > \bar{r}^*$.

Let $\bar{X} = \{x | x \in X \text{ and } \underline{r}(x) = \underline{r}^*\}$ and $\bar{Y} = \{y | y \in Y \text{ and } \bar{r}(y) = \bar{r}^*\}$ so that Lemma 1 holds for each $x \in \bar{X}$ and $y \in \bar{Y}$. For each $y \in \bar{Y}$, define $\bar{G}(y)$ to be a set of $N$ two-person zero sum games such that the $i$th game in the set has its payoff matrix

$$\bar{Q}_i(y) = \{\bar{q}_i^{kl}(y) | k \in C_i, l \in D_i\}, \qquad i = 1, 2, \cdots, N \tag{32}$$

where

$$\bar{q}_i^{kl}(y) = a_i^{kl} - b_i^{kl} \bar{r}^* + \beta \sum_{j=1}^{N} p_{ij}^{kl} \bar{s}_j(y) - \bar{s}_i(y).$$

Similarly for each $x \in \bar{X}$ let $\underline{G}(x)$ be a set of $N$ two-person zero sum games with payoff matrices.

$$\underline{Q}_i(x) = \{\underline{q}_i^{kl}(x) | k \in C_i, l \in D_i\}, \qquad i = 1, 2, \cdots, N \tag{33}$$

where

$$\underline{q}_i^{kl}(x) = a_i^{kl} - b_i^{kl}\underline{r}^* + \beta \sum_{j=1}^{N} p_{ij}^{kl}\underline{s}_j(x) - \underline{s}_i(x).$$

LEMMA 2. *Any stationary strategy $y^* \in \bar{Y}$ is optimal for Player II in the corresponding set $\bar{G}(y^*)$, that is, $y_i^*$ is optimal in the i-th game whose payoff matrix is $\bar{Q}_i(y^*)(i = 1, 2, \cdots, N)$. Similarly, any $x^* \in \bar{X}$ is optimal for Player I in the corresponding set $\underline{G}(x^*)$. Furthermore, the value of each game in the sets $\bar{G}(y^*)$ and $\underline{G}(x^*)$ is zero.*

*Proof.* From (28) we immediately have

$$0 \leqq \operatorname*{Min}_{y \in Y} \operatorname*{Max}_{k \in C_i} [\bar{Q}_i(y^*)y_i] = \operatorname*{Min}_{y \in Y} \operatorname*{Max}_{x \in X} x_i\bar{Q}_i(y^*)y_i \tag{34}$$

$$\leqq \operatorname*{Max}_{x \in X} x_i\bar{Q}_i(y^*)y_i^* = \operatorname*{Max}_{k \in C_i} [\bar{Q}_i(y^*)y_i^*] = 0, \tag{35}$$

and these establish the required results for the games in $\bar{G}(y^*)$. A similar argument shows the corresponding results for the games in the set $\underline{G}(x^*)$.

LEMMA 3. *Given $x^* \in \bar{X}$ and $y^* \in \bar{Y}$, then $\bar{r}(y^*) = \underline{r}(x^*)$ or equivalently, $\bar{r}^* = \underline{r}^*$, and furthermore, $\bar{s}(y^*) = \underline{s}(x^*)$.*

*Proof.* By Lemma 2, each game in the set $\bar{G}(y^*)$ has the value zero. Therefore, there exists $\tilde{x} \in X$ optimal for Player I in the set so that

$$\tilde{x}_i\bar{Q}_i(y^*)y_i \geqq 0, \qquad i = 1, 2, \cdots, N, \quad \forall y \in Y. \tag{36}$$

Similarly there exists $\tilde{y} \in Y$ optimal for Player II in the set $\underline{G}(x^*)$ yielding

$$x_i\underline{Q}_i(x^*)\tilde{y}_i \leqq 0, \qquad i = 1, 2, \cdots, N, \forall x \in X. \tag{37}$$

Thus from (36) and (37) we have

$$\tilde{x}_i[\bar{Q}_i(y^*) - \underline{Q}_i(x^*)]\tilde{y}_i \geqq 0, \qquad i = 1, 2, \cdots, N. \tag{38}$$

Substituting for $\bar{q}_i^{kl}(y^*)$ and $\underline{q}_i^{kl}(x^*)$ from (32) and (33) respectively, and setting $\Delta r^* = (\bar{r}^* - \underline{r}^*)$ and $\Delta s^* = (\bar{s}(y^*) - \underline{s}(x^*))$, from (38) we obtain

$$(I - \beta\tilde{P})\Delta s^* \leqq -\Delta r^*\tilde{b}, \tag{39}$$

where

$$\tilde{P} = \sum_{l \in D_i} \sum_{k \in C_i} P_{ij}^{kl}\tilde{x}_i^k\tilde{y}_i^l \quad \text{and} \quad \tilde{b}_i = \sum_{l \in D_i} \sum_{k \in C_i} b_i^{kl}\tilde{x}_i^k\tilde{y}_i^l.$$

We note that $\tilde{b} \gg 0$ and $(I - \beta\tilde{P})^{-1} \geqq I$. Therefore, $\Delta s^* \leqq -(I - \beta\tilde{P})^{-1}\Delta r^*\tilde{b}$. Further we have that $\alpha\Delta s^* = 0$. So

$$0 = \alpha\Delta s^* \leqq -\Delta r^*\alpha(I - \beta\tilde{P})^{-1}\tilde{b}. \tag{40}$$

It is easy to verify that $\alpha(I - \beta\tilde{P})^{-1}\tilde{b} > 0$; so from (40) we have $\Delta r^* \leqq 0$. Since $\bar{r}^* \geqq r^*$, we conclude $\bar{r}^* = \underline{r}^*$. As a consequence of Lemma 2 we have

$$x_i^*\bar{Q}_i(y^*)y_i^* \leqq 0, \qquad i = 1, 2, \cdots, N \tag{41}$$

and

$$x_i^*\underline{Q}_i(x^*)y_i^* \geqq 0, \qquad i = 1, 2, \cdots, N. \tag{42}$$

Applying the same analysis as on (36) and (37) on (41) and (42) we obtain

$$(I-\beta P^*)\Delta s^* \geqq -\Delta r^* b^*, \tag{43}$$

where $P^*$ and $b^*$ are defined appropriately. Since $\Delta r^* = 0$, (39) and (43), respectively, show that $\Delta s^* \leqq 0$ and $\Delta s^* \geqq 0$. Therefore $\Delta s^* = 0$ and this completes the proof of the lemma.

In Lemma 3 we have shown also that $\Delta s^* = 0$. It may be noted that this result is not needed for proving Theorem 3, but is useful in the characterization of the solution given later.

Now $\bar{r}^* = \underline{r}^* = r(x^*, y^*)$. Using this in Lemma 1 and recalling (24) and (26) we can see that

$$r(x, y^*) \leqq r(x^*, y^*) \leqq r(x^*, y), \qquad \forall x \in X, \quad \forall y \in Y. \tag{44}$$

Because of the stationarity property in the DMRDP [1] used in arriving at (24) from (23), it is clear that

$$r(t, y^*) \leqq r(x^*, y^*) \leqq r(x^*, u), \qquad \forall t \in T, \quad \forall u \in U. \tag{45}$$

Relation (45) shows that $x^*$ and $y^*$ are optimal for Players I and II, respectively, in the DSRG among all the behavioral strategies $T$ and $U$. The dependency of the solution on $\alpha$ follows from the fact that in the DMRDP [1], which may be viewed as a special case of the DSRG in which one of the players is fixed at a stationary strategy, the solution in general is dependent on $\alpha$. This completes the proof of Theorem 3.

From Lemma 3 we have $\bar{r}^* = \underline{r}^*$ and let this be denoted by $r^*$. Further it follows that for all $x \in \bar{X}$, $\underline{s}(x) = \bar{s}(y^*)$ and for all $y \in \bar{Y}$, $\bar{s}(y) = \underline{s}(x^*)$. Therefore, all $\bar{s}(y)$ and $\underline{s}(x)$ for $y \in \bar{Y}$ and $x \in \bar{X}$ are identical and we denote this by $s^*$. Using these in (32) and (33) we see that for all $y \in \bar{Y}$ and $x \in \bar{X}$, $\bar{Q}_i(y)$ and $\underline{Q}_i(x)$ are identical and let this be denoted by $Q_i$. Thus the sets of games $\bar{G}(y)$ and $\underline{G}(x)$ for all $y \in \bar{Y}$ and $x \in \bar{X}$ are the same denoted by $G$. The $i$th game denoted by $G_i (i = 1, 2, \cdots, N)$ in the set $G$ has as its payoff matrix $Q_i = \{q_i^{kl} | k \in C_i, l \in D_i\}$ where

$$q_i^{kl} = a_i^{kl} - b_i^{kl} r^* + \beta \sum_{j=1}^{N} p_{ij}^{kl} s_j^* - s_i^*. \tag{46}$$

Using these results in Lemma 1 we readily obtain a characterization of the minimax solution to the DSRG. Now consider the $(N+1)$ equations as follows:

$$\text{Value } G_i = \text{Value } [a_i^{kl} - b_i^{kl} r^* + \beta \sum_{j=1}^{N} p_{ij}^{kl} s_j^* - s_i^*] = 0, \qquad i = 1, 2, \cdots, N \tag{47}$$

and

$$\sum_{i=1}^{N} \alpha_i s_i^* = 0. \tag{48}$$

Relations (47) state that the two-person zero-sum game $G_i (i = 1, 2, \cdots, N)$, whose payoff matrix $Q_i$ has as its element in the $(k, l)$ position that which is given within the brackets, has value zero. It follows from Lemma 1 and the results of Lemma 3 that the above $(N+1)$ equations characterize the minimax solution to the DSRG. The value is $r^*$ and the stationary optimal strategies $(x^*, y^*)$ are composed of the optimal strategies $(x_i^*, y_i^*)$ in the game $G_i (i = 1, 2, \cdots N)$. The above characterization leads to a convergent algorithm for computing the solution, and this is presented in the next section.

**5. Algorithm and convergence proof.**

ALGORITHM.

*Step* 1. Fixing Player II at stationary strategy $y^{(0)}$, find the unique solution $(r^{(0)}, s_i^{(0)}(i=1,2,\cdots,N))$ of the system

$$\text{(49)}\qquad \max_{k\in C_i} \sum_{l\in D_i} (a_i^{kl} - b_i^{kl} r^{(0)} + \beta \sum_{j=1}^{N} p_{ij}^{kl} s_j^{(0)} - s_i^{(0)}) y_i^{(0)l} = 0, \qquad i=1,2,\cdots,N,$$

and

$$\text{(50)}\qquad \sum_{i=1}^{N} \alpha_i s_i^{(0)} = 0,$$

using the algorithm to solve a Markov ratio decision process [1].

*Step* 2. Solve the $N$ two-person zero-sum games whose element in the $(k,l)$ position of the payoff matrix of $i$th game is

$$\text{(51)}\qquad [a_i^{kl} - b_i^{kl} r^{(0)} + \beta \sum_{j=1}^{N} p_{ij}^{kl} s_j^{(0)} - s_i^{(0)}]$$

to obtain the optimal strategies $x_i^{(1)}$ and $y_i^{(1)}$ and the unique value of $g_i^{(1)}$ for $i=1,2,\cdots,N$. Set $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, \cdots, x_i^{(1)}, \cdots, x_N^{(1)})$ and $y^{(1)} = (y_1^{(1)}, y_2^{(1)}, \cdots, y_i^{(1)}, \cdots, y_N^{(1)})$.

*Step* 3. If $g_i^{(1)} = 0$ for all $i$, then the solution of the DSRG is gotten as $x^* = x^{(1)}$, $y^* = y^{(1)}$, and $r^* = r^{(0)}$.

If $g_i^{(1)} \neq 0$ at least for one state, then go to Step 1 with the stationary strategy $y^{(1)}$ obtained in Step 2 above.

*Convergence Proof.* Firstly, we show that the sequence $r^{(m)}(m=0,1,2,\cdots)$ generated by the algorithm is a monotonically decreasing one bounded from below by the unique value $r^*$.

*Case* (i): $g_i^{(1)} = 0$ for all $i$. In this case $x^{(1)}$, $y^{(1)}$ and $r^{(0)}$ form the solution of the DSRG since the characterization given by (47) and (48) are satisfied.

*Case* (ii): $g_i^{(1)} \neq 0$ at least for one state.

In this case, clearly $g_i^{(1)} \leqq 0$ for all $i$ since (49) and (50) assure that the maximal payoff to Player I in each game is zero when Player II uses $y^{(0)}$. Therefore, $g_i^{(1)} < 0$ at least for one $i$. Now it remains to be shown that if Player II is fixed at $y^{(1)}$, the maximal solution of the resulting Markov ratio decision process has the property that $r^{(1)} < r^{(0)}$. For this purpose consider the dual linear programming formulation of the problem where Player II is fixed at $y^{(1)}$.

$$\text{Min } r \quad \text{subject to}$$

$$\text{(52)}\qquad \sum_{l\in D_i} (a_i^{kl} - b_i^{kl} r + \beta \sum_{j=1}^{N} p_{ij}^{kl} s_j - s_i) y_i^{(1)l} \leqq 0, \qquad k\in C_i, \quad i=1,2,\cdots,N;$$

$$\sum_{i=1}^{N} \alpha_i s_i = 0, \qquad r, s_i (i=1,2,\cdots,N) \text{ unrestricted.}$$

Now recalling the unique solution to this program as characterized by Theorem 1, we see that the solution $(r^{(0)}, s^{(0)})$ is feasible in the above program but not optimal. Feasibility follows from the fact that $g_i^{(1)} \leqq 0$ for all $i$ and nonoptimality from the fact that $g_i^{(1)} < 0$ for at least one $i$. Therefore the optimal value $r^{(1)}$ of the above program must be strictly less than $r^{(0)}$. Thus

$$\text{(53)}\qquad r^{(1)} < r^{(0)}.$$

Further, $(r^{(m)}, s^{(m)})(m = 0, 1, 2, \cdots)$ are all in compact set. Now following arguments similar to those given by Hoffman and Karp [4] we see that the sequence $r^{(m)}$ indeed converges to $r^*$. It is of interest to note that a similar proof will hold if the algorithm is started fixing Player I instead of Player II. But in this case, the sequence $r^{(m)}$ will be a monotonically increasing one bounded from above by $r^*$.

It is of interest to note that the algorithm is also capable of revealing approximate solutions in the following sense. The stationary strategy $y^{(m)}$ revealed at the $m$th iteration is capable of limiting the payoff to Player I to $r^{(m)}$. Similarly, if the algorithm is applied fixing Player I instead of Player II, one can see that $x^{(m)}$ will ensure a payoff of at least $r^{(m)}$ to Player I. Finally, we observe that the algorithm is structurally identical to the policy improvement algorithm given by Denardo [3] in the context of stochastic games of Shapley [7]. He shows how the contraction property applies to stochastic games and uses it in proving the convergence of his algorithm. We were not able to show the contraction property directly for our model. We therefore could not apply Denardo's proof of convergence.

**Acknowledgment.** The authors are grateful to Uri Rothblum for his comments and suggestions on earlier versions of this paper.

## REFERENCES

[1] V. AGGARWAL, R. CHANDRASEKARAN AND K. P. K. NAIR, *Markov ratio decision processes*, J. Optimization Theory Appl., 21 (1977), pp. 27–37.

[2] R. AUMANN, *Mixed and behavior strategies in infinite games*, Advances in Game Theory, Annals of Mathematics Studies, Number 52, Princeton, NJ, 1964, pp. 627–650.

[3] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Review, 9 (1967), pp. 165–177.

[4] A. J. HOFFMAN AND R. M. KARP, *On nonterminating stochastic games*, Management Sci., 12 (1966), pp. 359–370.

[5] W. S. JEWELL, *Markov-renewal programming. I: Formulations, finite return models*, Operations Res. 11 (1963), pp. 938–948.

[6] R. G. SCHROEDER, *Linear programming solutions to ratio games*, Operations Res. 18 (1970), pp. 300–305.

[7] L. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 1095–1100.

[8] J. VON NEUMANN, *A Model of general economic equilibrium*, Review of Economic Studies, 13 (1945), pp. 1–9.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0008 $01.00/0

# DILWORTH NUMBERS, INCIDENCE MAPS AND PRODUCT PARTIAL ORDERS*

MICHAEL SAKS†

**Abstract.** The Dilworth numbers of a finite partially ordered set $P$, $\{d_k(P) | k \geqq 0\}$, are the maximum sizes of the union of $k$ antichains. We give a characterization of the Dilworth numbers in terms of the dimensions of the kernels of certain linear maps on the complex vector space generated by the elements of $P$. This is applied to prove bounds on the Dilworth numbers of product partial orders in terms of those of its factors and to prove sufficient conditions for the product of two partial orders to have the Sperner property.

**1. Introduction.** A fundamental set of quantities associated with a finite partially ordered set $P$ are the Dilworth numbers $d_k(P)$, defined for each integer $k \geqq 0$ to be the size of the largest $k$-family (union of $k$ antichains) of $P$. Greene and Kleitman [1] and Greene [2] investigated these numbers in detail and proved several striking results about them. In this paper, we consider additional properties of the Dilworth numbers. In particular, we establish a connection between them and a class of linear maps associated with the poset. This enables us to prove theorems which relate the Dilworth numbers of a direct product order to those of its factors. This result is applied to strengthen a theorem due to Proctor, Saks and Sturtevant [7], which gives a sufficient condition for the product of two orders to have the Sperner property.

The use of linear algebra to study the Dilworth numbers was suggested by some recent work of Stanley [9], who used properties of linear maps to characterize partial orders with the Sperner property. The maps studied here are members of the incidence algebra of the poset, described by Rota [10], which has been studied extensively in connection with various enumeration problems.

**2. Preliminary definitions and results.** Let $(P, \leqq)$ be a finite partially ordered set. For $k \geqq 0$, a *$k$-family* of $P$ is a subset of $P$ containing no chain of cardinality greater than $k$; equivalently, it is a union of $k$ (possibly empty) antichains. If $\mathscr{C} = \{C_1, C_2, \cdots, C_m\}$ is any partition of $P$ into chains, then since every $k$-family intersects each chain at most $k$ times, we have $d_k(P) \leqq \sum_{i=1}^{m} \min(|C_i|, k)$. A chain partition $\mathscr{C}$ is said to be *$k$-saturated* if this holds as an equality. The following theorem was first proved by Greene and Kleitman.

THEOREM 2.1 [1]. *For any poset $P$ and positive integer $k$ there exists a $k$-saturated partition of $P$.*

For any sequence $f_0, f_1, \cdots$ let $\Delta f_i = f_i - f_{i-1}$ and $\delta f_i = \Delta f_i - \Delta f_{i+1}$. Greene and Kleitman also proved:

LEMMA 2.2 [1]. *For any poset $P$, $\delta d_k(P) \geqq 0$ for all $k \geqq 1$.*

If $P$ and $Q$ are posets their *direct product* $P \times Q$ is the set of pairs $(p, q)$ where $p \in P$ and $q \in Q$, with the order relation given by $(p_1, q_1) \geqq (p_2, q_2)$ if $p_1 \geqq p_2$ and $q_1 \geqq q_2$. Let $T_k$ denote the totally ordered set $t_1 \leqq t_2 \leqq \cdots \leqq t_k$. The following simple lemma was noted in [8]:

LEMMA 2.3. $d_k(P) = d_1(P \times T_k)$.

Let $\tilde{P}$ denote the vector space over $\mathbb{C}$ generated freely by the elements of $P$. An *incidence map* on $P$ is a linear map $\Phi: \tilde{P} \to \tilde{P}$ which sends each $a \in P$ to a linear

* Received by the editors August 8, 1979 and in revised form December 27, 1979.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This work was supported in part by the U.S. Office of Naval Research under Contract N00014-76-C-0366. The results in this paper are contained in the author's doctoral thesis, written under the direction of Professor Daniel J. Kleitman.

combination of elements less than or equal to $a$. Let $N(P)$ be the set of nilpotent incidence maps ($\Phi^k = 0$ for some finite $k$), then $N(P)$ consists of those maps which send each $a \in P$ to a sum of elements strictly less than $a$. For a product poset $P \times Q$, we have $\widetilde{P \times Q} = \tilde{P} \otimes \tilde{Q}$, and $N(P \times Q)$ is generated by maps $\Phi \otimes \tau$ where $\Phi$ and $\tau$ are incidence maps on $\tilde{P}$ and $\tilde{Q}$ and at least one is nilpotent.

For any space $\tilde{V}$ let $I_{\tilde{V}}$ denote the identity map. If $\tilde{W}$ is a subspace of $\tilde{V}$ and $\Phi: \tilde{V} \to \tilde{V}$, the restriction of $\Phi$ to $\tilde{W}$ is denoted $\Phi|\tilde{W}$. Let $D_k(\Phi)$ equal the dimension of the kernel of $\Phi^k$.

**3. Main results.**

THEOREM 3.1. *Let $P$ be a poset and $\Phi \in N(P)$. Then for all $k \geqq 0$,*

$$d_k(P) \leqq D_k(\Phi).$$

THEOREM 3.2. *For any poset $P$ there exists $\Phi \in N(P)$ such that*

$$d_k(P) = D_k(\Phi) \text{ for all } k \geqq 0.$$

THEOREM 3.3. *For any posets $P$ and $Q$,*

$$\text{(i)} \qquad d_1(P \times Q) \leqq \sum_{i \geqq 1} \Delta d_i(P)\, \Delta d_i(Q) \quad \text{and}$$

$$\text{(ii)} \qquad d_1(P \times Q) \leqq \sum_{i \geqq 1} \delta d_i(P)\, d_i(Q).$$

THEOREM 3.4. *For any posets $P$ and $Q$ and $k \geqq 1$,*

$$d_k(P \times Q) \leqq \sum_{i \geqq 1} \sum_{j \geqq 1} \delta d_i(P)\, \delta d_j(Q)\, d_1(T_i \times T_j \times T_k).$$

In order to prove these results we will need some results from linear algebra. The first lemma reviews the properties of the Jordan form of nilpotent maps. It is presented without proof; the reader is referred to any book on linear algebra (for example, [6]).

LEMMA 3.5 (*Jordan Decomposition for nilpotent maps*). *Let $\tilde{V}$ be a finite dimensional vector space over $\mathbb{C}$ and $\Phi: \tilde{V} \to \tilde{V}$ be a nilpotent map. Then*:

(i) *There exists a decomposition $\bigoplus_{j=1}^{n} \tilde{V}_j$ of $\tilde{V}$ into $\Phi$-invariant subspaces such that $D_1(\Phi|\tilde{V}_j) = 1$ for each $1 \leqq j \leqq n$.*

(ii) *The collection $\{\dim \tilde{V}_j\}$ is the same for all such decompositions of $\tilde{V}$.*

(iii) *Each $\tilde{V}_j$ has a basis $v_1^j, v_2^j, \cdots, v_{r_j}^j$ (where $r_j = \dim \tilde{V}_j$) such that $\Phi(v_1^j) = 0$ and $\Phi(v_i^j) = \Phi(v_{i-1}^j)$ for $i > 1$.*

LEMMA 3.6. *Let $\bigoplus_{i=1}^{n} \tilde{V}_i$ be a Jordan decomposition of $\tilde{V}$ with respect to a nilpotent map $\Phi$. Then*

$$\text{(i)} \quad D_k(\Phi) = \sum_{i=1}^{n} \min(k, \dim \tilde{V}_i) \text{ and}$$

$$\text{(ii)} \quad \Delta D_k(\Phi) = |\{\tilde{V}_i : \dim \tilde{V}_i \geqq k\}|.$$

*Proof.* By Lemma 3.5(iii), each time $\Phi$ acts on $\tilde{V}_i$ it reduces the dimension by 1 until $\tilde{V}_i$ is annihilated, so $D_k(\Phi|\tilde{V}_i) = \min(k, \dim \tilde{V}_i)$. Since each $\tilde{V}_i$ is $\Phi$-invariant, $D_k(\Phi) = \sum_{i=1}^{n} D_k(\Phi|\tilde{V}_i) = \sum_{i=1}^{n} \min(k, \dim \tilde{V}_i)$, proving (i). Now, by definition $\Delta D_k(\Phi|\tilde{V}_i) = D_k(\Phi|\tilde{V}_i) - D_{k-1}(\Phi|\tilde{V}_i)$. By the previous observation this equals 1 if $\dim \tilde{V}_i \geqq k$ and 0 otherwise, hence (ii) follows. □

The key lemma in the proofs of Theorems 3.1 and 3.3 is the following:

LEMMA 3.7. *Let $\Phi: \tilde{V} \to \tilde{V}$ and $\psi: \tilde{W} \to \tilde{W}$ be nilpotent maps. Define $\lambda: \tilde{V} \otimes \tilde{W} \to \tilde{V} \otimes \tilde{W}$ by $\lambda = (\Phi \otimes I_{\tilde{W}}) + (I_{\tilde{V}} \otimes \psi)$. Then*

$$D_1(\lambda) = \sum_{k \geqq 1} \Delta D_k(\Phi)\, \Delta D_k(\psi).$$

*Proof.* First consider the case in which the Jordan decompositions of $\tilde{V}$ and $\tilde{W}$ with respect to $\Phi$ and $\psi$ each consist of a single block; we show that $D_1(\lambda) = \min(\dim \tilde{V}, \dim \tilde{W})$. Let $m = \dim \tilde{V}$, $n = \dim \tilde{W}$ and, without loss of generality, assume $m \leqq n$. Let $v_1, \cdots, v_m$ and $w_1, \cdots, w_n$ be Jordan bases of $\tilde{V}$ and $\tilde{W}$, and set $\tilde{U} = \tilde{V} \otimes \tilde{W}$. (For $i$ not in the range 1 to $m$ and $j$ not in the range 1 to $n$, we set $v_i = w_j = 0$.) We can write $\tilde{U} = \bigoplus_{r=1}^{m+n-1} \tilde{U}_r$ where $\tilde{U}_r$ is the space generated by $\{v_i \otimes w_j : i + j = r+1\}$. For a general element $x = \sum_i c_i(v_i \otimes w_{r+1-i})$ of $\tilde{U}_r$ we have:

$$\begin{aligned} \lambda\Big(\sum_i c_i(v_i \otimes w_{r+1-i})\Big) &= \sum_i c_i(v_i \otimes w_{r-i}) + \sum_i c_i(v_{i-1} \otimes w_{r+1-i}) \\ &= \sum_i (c_i + c_{i+1})(v_i \otimes w_{r-i}) \end{aligned} \tag{3.1}$$

which is an element of $\tilde{U}_{r-1}$, so $\operatorname{Im}(\lambda|\tilde{U}_r) \subseteq \tilde{U}_{r-1}$. Thus the images of $\lambda|\tilde{U}_r$ are disjoint for each $r$ and we have $D_1(\lambda) = \sum_r D_1(\lambda|\tilde{U}_r)$. We now show that, for $r > m$, $\ker(\lambda|\tilde{U}_r)$ is trivial and, for each $r$ in the range 1 to $m$, $\ker(\lambda|\tilde{U}_r)$ is generated by the element $u_r = \sum_{i=1}^{m} (-1)^i v_i \otimes w_{r+1-i}$. Therefore $D_1(\lambda) = m$ as claimed.

For $r > m$, suppose that $x = \sum_i c_i(v_i \otimes w_{r+1-i})$ is in $\ker(\lambda|\tilde{U}_r)$ with $x \neq 0$. Let $s$ be the largest index such that $c_s \neq 0$. Since $r > m \geqq s > 0$, $v_s \otimes w_{r-s}$ is a nonzero vector and by (3.1) the coefficient of $v_s \otimes w_{r-s}$ in $\lambda(x)$ is $c_s + c_{s+1} = c_s \neq 0$. This contradicts $x \in \ker(\lambda|\tilde{U}_r)$.

For $r \leqq m$, it is easy to verify that $u_r \in \ker(\lambda|\tilde{U}_r)$. For any $x$ in the kernel, (3.1) implies that $c_i = c_1$ if $i$ is odd and $c_i = -c_1$ if $i$ is even, so $x = -c_1 u_r$ and $D_1(\lambda|\tilde{U}_r) = 1$. Hence $D_1(\lambda) = \min(\dim \tilde{V}, \dim \tilde{W})$.

For the general case, let $\bigoplus V_i$ and $\bigoplus W_j$ be Jordan decompositions of $V$ and $W$ with respect to $\Phi$ and $\psi$. Since each $V_i$ is $\Phi$-invariant and each $W_j$ is $\psi$-invariant, it follows that each $V_i \otimes W_j$ is $\lambda$-invariant. Thus $D_1(\lambda) = \sum_i \sum_j D_1(\lambda|V_i \otimes W_j)$. The result proved above applies to each term in the sum so we have $D_1(\lambda) = \sum_i \sum_j \min(\dim \tilde{V}_i, \dim \tilde{W}_j)$. Define $f(i, j, k) = 1$ if $\dim \tilde{V}_i \geqq k$ and $\dim \tilde{W}_j \geqq k$, and $f(i, j, k) = 0$ otherwise. Then

$$\begin{aligned} \sum_i \sum_j \min(\dim \tilde{V}_i, \dim \tilde{W}_j) &= \sum_i \sum_j \sum_k f(i, j, k) \\ &= \sum_k |\{\tilde{V}_i : \dim \tilde{V}_i \geqq k\}|\, |\{\tilde{W}_j : \dim \tilde{W}_j \geqq k\}|, \end{aligned}$$

which equals $\sum_k \Delta D_k(\Phi)\, \Delta D_k(\psi)$ by Lemma 3.6(ii). □

*Proof of Theorem* 3.1. We first prove the inequality for $k = 1$. Let $A$ be an antichain of maximum cardinality, let $I = \{x \in P | \exists a \in A \text{ such that } x \leqq a\}$, and $J = I \cap A^c$. Then $\operatorname{Im}(\Phi|\tilde{I}) \subseteq \tilde{J}$ and so:

$$D_1(\Phi) \geqq D_1(\Phi|\tilde{I}) \geqq |I| - |J| = |A| = d_1(P), \tag{3.2}$$

which proves the theorem for $k = 1$. To prove the theorem for any $k$, we define the map $\tau_k \in N(T_k)$ by $\tau_k(t_1) = 0$ and $\tau_k(t_i) = t_{i-1}$ if $i > 1$. The following sequence of relations now yields the desired result:

$$d_k(P) = d_1(P \times T_k) \leqq D_1(\Phi \otimes I_{\tilde{T}_k} + I_{\tilde{P}} \otimes \tau_k) = D_k(\Phi).$$

The first equality is Lemma 2.3. the inequality is (3.2) applied to $P \times T_k$ and the map $\Phi \otimes I + I \otimes \tau_k$ which is in $N(P \times T_k)$. The final equality is obtained from Lemma 3.7, by

noting $\Delta D_j(\tau_k) = 1$ if $j \leqq k$ and 0 otherwise, so that

$$D_1(\Phi \otimes I + I \otimes \tau_k) = \sum_i \Delta D_i(\Phi)\, \Delta D_i(\tau_k) = D_k(\Phi),$$

as required. □

*Proof of Theorem* 3.2. For each $1 \leqq k \leqq |P|$ let $\mathscr{C}_k$ be a $k$-saturated partition of $P$ into chains. Define maps $\chi_k : P \to P$ by

$$\chi_k(a) = \begin{cases} 0 & \text{if } a \text{ is minimal in its chain in } \mathscr{C}_k, \\ a' & \text{where } a' \text{ is covered by } a \text{ in its chain in } \mathscr{C}_k. \end{cases}$$

Note that the kernel of $\chi_k^k$ is the space generated by the $k$ smallest elements in each chain so, since $\mathscr{C}_k$ is $k$-saturated,

$$D_k(\chi_k) = \sum_{C \in \mathscr{C}_k} \min(k, |C|) = d_k(P). \tag{3.3}$$

Now let $t_1, t_2, \cdots, t_{|P|}$ be complex numbers which are algebraically independent over the rationals and set $\Phi = \sum_{j=1}^{|P|} t_j \chi_j$. We now show that $D_k(\Phi) \leqq d_k(P)$ for all $k$; Theorem 3.1 guarantees that $D_k(\Phi) \geqq d_k(P)$, so $\Phi$ is the desired map.

By (3.3), it now suffices to show that, for all $k$, the rank of $\Phi^k$ is greater than or equal to the rank of $\chi_k^k$, which we do by showing that a minor of $\Phi^k$ is singular only if the corresponding minor of $\chi_k^k$ is also. By algebraic independence a minor of $\Phi^k$ is singular only if its determinant, considered as a polynomial in the $t_i$, is identically zero. This implies that the corresponding minor in $\chi_k^k$, whose determinant is obtained by evaluating the polynomial at $t_k = 1$ and $t_i = 0$ for $i \neq k$, is also singular. □

*Proof of Theorem* 3.3. Let $\Phi : \tilde{P} \to \tilde{P}$ and $\psi : \tilde{Q} \to \tilde{Q}$ be maps satisfying Theorem 3.2. Since $d_k(P) = D_k(\Phi)$ and $d_k(Q) = D_k(\psi)$ for all $k$, we have

$$\sum_k \Delta d_k(P)\, \Delta d_k(Q) = \sum_k \Delta D_k(\Phi)\, \Delta D_k(\psi).$$

By Lemma 3.7, this equals $D_1(\Phi \otimes I_Q + I_P \otimes \psi)$, which, by Theorem 3.1, is greater than or equal to $d_1(P \times Q)$, since $\Phi \otimes I_Q + I_P \otimes \psi \in N(P \times Q)$. This proves (i).

To prove (ii), we note the identities

$$\begin{aligned} \sum_k \Delta d_k(P)\, \Delta d_k(Q) &= \sum_k \Delta d_k(P) d_k(Q) - \sum_k \Delta d_k(P) d_{k-1}(Q) \\ &= \sum_k (\Delta d_k(P) - \Delta d_{k+1}(P)) d_k(Q) \\ &= \sum_k \delta d_k(P) d_k(Q). \end{aligned}$$

□

*Proof of Theorem* 3.4. The proof of this theorem involves simple manipulation of sums using Theorem 3.3 and Lemmas 2.2 and 2.3.

$$\begin{aligned} d_k(P \times Q) &= d(P \times Q \times T_k) && \text{by Lemma 2.3,} \\ &\leqq \sum_j \delta d_j(P) d_j(Q \times T_k) && \text{by Theorem 3.3(ii),} \\ &= \sum_j \delta d_j(P) d_1(Q \times (T_k \times T_j)) && \text{by Lemma 2.3,} \\ &\leqq \sum_j \delta d_j(P) \sum_i \delta d_i(Q) d_i(T_k \times T_j) && \text{by Theorem 3.3(ii),} \\ & && \text{since } \delta d_j(P) \geqq 0\ \forall_j \text{ (Lemma 2.2),} \\ &= \sum_j \sum_i \delta d_j(P)\, \delta d_i(Q) d_1(T_i \times T_j \times T_k) && \text{by Lemma 2.3} \quad \square \end{aligned}$$

**4. Sperner product orders.** A *ranking* of a poset $P$ is a partition of $P$ into sets $P_0, P_1, \cdots, P_j$ such that for each $i$ every element of $P_i$ is covered only by elements in $P_{i+1}$. The set $P_i$ is called the $i$th rank of $P$. A ranked poset $P$ is said to be *Sperner* if the rank of largest size is an antichain of maximum size, *k-Sperner* if the union of the $k$ largest ranks is a maximum $k$-family, and *strongly Sperner* if it is $k$-Sperner for all $k \geqq 1$.

If $P$ and $Q$ are ranked, there is an induced ranking on $P \times Q$:

$$(P \times Q)_j = \bigcup_i P_i \times Q_{j-i}$$

where $P_i$ and $Q_{j-i}$ are taken to be empty if they are not defined. $P$ and $Q$ are said to be *compatible* if there exists an integer $t$ such that for all $i$ and $j$, $|P_i| < |P_j|$ only if $|Q_{t-i}| \leqq |Q_{t-j}|$. The integer $t$ (which need not be unique) is called the *compatibility index*. If $P$ and $Q$ are compatible with index $t$, then $(P \times Q)_t$ is the union over $i$ of the product of the $i$th largest rank in $P$ and the $i$th largest rank in $Q$.

THEOREM 4.1. *Let $P$ and $Q$ be ranked posets which are strongly Sperner and compatible. Then $P \times Q$ is Sperner.*

*Proof.* In a strongly Sperner poset $P$, $d_i(P)$ is the size of the largest $i$ ranks, so $\Delta d_i(P)$ is the size of the $i$th largest rank. If $t$ is the compatibility index of $P$ and $Q$, then $(P \times Q)_t$ has cardinality $\sum_i \Delta d_i(P) \Delta d_i(Q)$. By Theorem 3.3, no antichain is larger. □

In [7], examples were presented to show that if either order is not strongly Sperner or if they are not compatible, then $P \times Q$ need not be Sperner.

## REFERENCES


[1] C. GREENE AND D. J. KLEITMAN, *The structure of Sperner k-families*, J. Comb. Th. (A), 20 (1976), pp. 41–68.

[2] C. GREENE, *Some partitions associated with a partially ordered set*, J. Comb. Th. (A), 20 (1976), pp. 69–79.

[3] J. R. GRIGGS, *On chains and Sperner k-families in ranked posets*, J. Comb. Th. (A), to appear.

[4] J. R. GRIGGS, *Poset measure and saturated partitions*, preprint.

[5] L. H. HARPER, *The morphology of partially ordered sets*, J. Comb. Th., 17 (1974), pp. 44–58.

[6] K. HOFFMAN AND R. KUNZE, *Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[7] R. PROCTOR, M. SAKS AND D. STURTEVANT, *Product partial orders with the Sperner Property*, Discrete Math., to appear.

[8] M. SAKS, *A short proof of the existence of k-saturated partitions of a partially ordered set*, Advances in Mathematics, 33 (1979), pp. 207–211.

[9] R. STANLEY, *Weyl groups, the hard Lefschetz Theorem and the Sperner property*, this Journal, this issue, pp. 168–184.

[10] G.-C. ROTA, *On the foundations of combinatorial theory, I: Theory of Mobius functions*, Z. Wahrscheinlichkeitstheorie, 2 (1964), pp. 340–368.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0009 $01.00/0

# THE COMPLEXITY OF COLORING CIRCULAR ARCS AND CHORDS*

M. R. GAREY,† D. S. JOHNSON,† G. L. MILLER‡ AND C. H. PAPADIMITRIOU‡

**Abstract.** The word problem for products of symmetric groups, the circular arc graph coloring problem, and the circle graph coloring problem, as well as several related problems, are proved to be *NP*-complete. For any fixed number $K$ of colors, the problem of determining whether a given circular arc graph is $K$-colorable is shown to be solvable in polynomial time.

**1. Introduction.** The *NP*-completeness of many standard graph-theoretic problems for general graphs [4] has motivated the study of various special classes of graphs for which these problems might be less difficult. A variety of results, both positive (i.e., polynomial time algorithms) and negative (i.e., proofs of *NP*-completeness), have been obtained for such classes as planar graphs, comparability graphs, interval graphs, chordal graphs, circular arc graphs, and circle graphs (see [4]). However, a number of significant questions have remained open. In this paper we address two of these open questions, namely the questions of how difficult it is to *color* circular arc graphs and circle graphs.

A graph $G$ is called a *circular arc graph* if its vertices can be placed in one-to-one correspondence with a family $F$ of arcs of a circle in such a way that two vertices of $G$ are joined by an edge if and only if the corresponding two arcs in $F$ intersect one another. For example, the graph in Fig. 1(a) is a circular arc graph because it has the circular arc model shown in Fig. 1(b). Circular arc graphs were first discussed in [8] as a natural

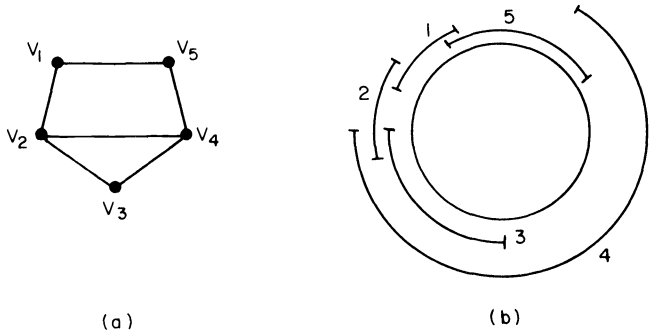



FIG. 1. *A circular arc graph and its circular arc model.*

generalization of *interval graphs* (defined analogously, but using intervals on a line instead of arcs of a circle), and they have since been studied extensively [6], [10], [11], [12], [13]. Tucker [13] has recently given a polynomial time algorithm for recognizing circular arc graphs. Gavril [6] has shown that the problems of finding a maximum independent set, a maximum clique, and a minimum covering by cliques, all of which are *NP*-complete for general graphs, can be solved in polynomial time for circular arc graphs.

* Received by the editors November 19, 1979.
† Bell Laboratories, Murray Hill, New Jersey 07974.
‡ Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

A graph $G$ is called a *circle graph* if its vertices can be placed in one-to-one correspondence with a family of *chords* of a circle in such a way that two vertices are joined by an edge $G$ if and only if the corresponding chords intersect. Fig. 2 shows a circle graph and its chord model. Although no polynomial time recognition algorithm is

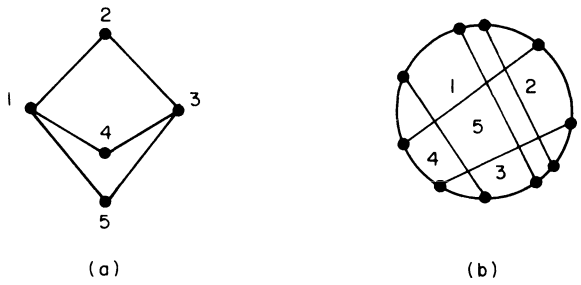



FIG. 2. *A circle graph and its chord model.*

known for circle graphs, [5] shows that, if the graph is described by giving its chord model, then both the maximum independent set problem and the maximum clique problem can be solved in polynomial time.

We shall study the coloring problems for these two classes in the following form: Given a family $F$ of circular arcs (chords) and a positive integer $K$, can the arcs (chords) in $F$ be colored with $K$ or fewer colors so that no two intersecting arcs (chords) have the same color?

A number of partial results about the arc coloring problem (coloring circular arc graphs) can be found in [12], which also notes the potential applicability of circular arc coloring to the following register allocation problem. Consider a loop in a computer program, and regard the flow of control around the loop as being described by a circle. For each assignment of a value to a variable within the loop, the *lifetime* of that assignment consists of the portion of the loop that begins where the assignment is made and that ends where that value is used for the last time. Each such lifetime thus corresponds to an arc of the circle. Furthermore, a $K$-coloring of this set of arcs can be regarded as assigning one of $K$ registers to each lifetime, in such a way that, if the value corresponding to that lifetime is stored in the associated register, then no value will ever have to be recomputed or stored elsewhere. The minimum value of $K$ for which the circular arc graph can be colored therefore gives the minimum number of registers needed for doing this.

The chord coloring problem is discussed in [2], where it is shown to model a problem of realizing a given permutation using a minimum number of parallel stacks.

In this paper we provide strong evidence that neither of these coloring problems can be solved in polynomial time, by showing that they are both *NP*-complete. (Readers unfamiliar with the central notions and terminology pertaining to the theory of *NP*-completeness can consult [1] or [4].) We begin by concentrating on the circular arc coloring problem. In § 2 we show that this problem is equivalent to the word problem for products of symmetric groups[1] and use this equivalence to derive an

[1] The equivalence, at least in one direction, was apparently known to Tucker [12].

$O(n \cdot K! \cdot K \cdot \log K)$ algorithm for coloring $n$ circular arcs with $K$ colors (whenever possible). This implies that there is a sense in which circular arc coloring is easier than general graph coloring. The algorithm will run in polynomial time for any *fixed* value of $K$, whereas for general graphs the coloring problem is *NP*-complete for *every* fixed value of $K > 2$ [3]. However, if $K$ is not fixed, the circular arc coloring problem loses its advantage and becomes, like the general problem, *NP*-complete. We prove this in § 3 by showing that the word problem for products of symmetric groups is itself *NP*-complete. In § 4 the *NP*-completeness of the chord coloring problem is derived by a direct transformation from the circular arc coloring problem. Finally, in § 5, we discuss the implications of our results and some of the remaining open problems and directions for further research.

**2. Circular arc coloring as a permutation problem.** In this section, we formalize the circular arc coloring problem (in a manner suitable for computation), introduce the word problem for products of symmetric groups, and prove that these two problems are equivalent with respect to polynomial time solvability. We then use this equivalence to give an $O(n \cdot K! \cdot K \cdot \log K)$ algorithm for coloring $n$ circular arcs with $K$ colors whenever such a coloring is possible.

We formalize the circular arc coloring problem as follows: A *family* $F$ of *circular arcs* is a set $\{A_1, A_2, \cdots, A_n\}$, where each $A_i$ is an ordered pair $(a_i, b_i)$ of positive integers, with $a_i \neq b_i$. Let $m$ denote the largest integer among all the $a_i$'s and $b_i$'s. Then we can regard the circle as being divided into $m$ parts by $m$ equally spaced points, numbered clockwise as $1, 2, \cdots, m$, and each $A_i = (a_i, b_i)$ can be regarded as representing the circular arc from point $a_i$ to point $b_i$, again in the clockwise direction. Notice that we might have either $a_i < b_i$ or $b_i < a_i$ for any $A_i$.

The *span* $sp(A_i)$ of an arc $A_i = (a_i, b_i)$ is the set $\{a_i + 1, a_i + 2, \cdots, b_i\}$ if $a_i < b_i$ or $\{a_i + 1, \cdots, m, 1, 2, \cdots, b_i\}$ if $b_i < a_i$. We say that two arcs $A_i$ and $A_j$ *intersect* if $sp(A_i) \cap sp(A_j)$ is not empty. Notice that two arcs do not intersect if they share only common endpoints. The *circular arc graph* corresponding to the family $F$ is the graph $G = (F, E)$, where $\{A_i, A_j\} \in E$ if and only if $A_i$ and $A_j$ intersect.

Notice that, since we are only interested in the intersection pattern among arcs in $F$, there is no loss in generality in assuming that all the integers appearing in the pairs $(a_i, b_i)$ are bounded above by $2n$, where $n$ is the numbers of arcs in $F$. (If not, we can simply sort the $a_i$'s and $b_i$'s and replace each by its rank in the sorted sequence.) Henceforth we shall restrict our attention to families $F$ satisfying this property. The arc coloring problem can now be defined as follows:

*ARC COLORING.* Given a family $F$ of circular arcs and a positive integer $K$, can $F$ be partitioned into $K$ classes so that no two arcs in the same class intersect? (Or, equivalently, can the circular arc graph $G = (F, E)$ be colored with $K$ colors?)

To define the word problem for products of symmetric groups, let $S_K$ denote the symmetric group of all permutations on $\{1, 2, \cdots, K\}$ (i.e., the set of all one-to-one functions from $\{1, 2, \cdots, K\}$ onto itself). For $X \subseteq \{1, 2, \cdots, K\}$, let $S_X$ denote the subgroup of $S_K$ consisting of exactly those permutations that leave all elements outside of $X$ fixed. If $P_1$ and $P_2$ are subsets of $S_K$, then their product $P_1 \cdot P_2$ is the set of all permutations $\pi \in S_K$ that can be written as $\pi = \pi_1 \cdot \pi_2$ (with $\pi_1 \cdot \pi_2$ interpreted as first applying $\pi_1$ and then applying $\pi_2$), where $\pi_1 \in P_1$ and $\pi_2 \in P_2$. The word problem for products of symmetric groups (WPPSG) is defined as follows:

*WPPSG.* Given $K$, subsets $X_1, X_2, \cdots, X_m \subseteq \{1, 2, \cdots, K\}$, and a permutation $\pi \in S_K$, does $\pi$ belong to the set $P = S_{X_1} \cdot S_{X_2} \cdot S_{X_3} \cdots S_{X_m}$, i.e., can $\pi$ be written as $\pi = \pi_1 \cdot \pi_2 \cdot \pi_3 \cdots \pi_m$ where $\pi_i \in S_{X_i}$ for $1 \leqq i \leqq m$?

The main result of this section is then given by the following theorem:

THEOREM 1. *WPPSG is polynomially equivalent to ARC COLORING.*

*Proof.* We describe the two transformations. First, given an instance $K, X_1, X_2, \cdots, X_m, \pi$ of WPPSG, we shall show how to construct in polynomial time a family $F$ of circular arcs such that $F$ is $K$-colorable if and only if $\pi \in P$.

Without loss of generality, we may assume that each integer $i \in \{1, 2, \cdots, K\}$ occurs in at least one set $S_j$; for if $i$ occurs in no such set, then either $\pi(i) \neq i$ and the answer is trivially "no" for this instance, or $\pi(i) = i$ and we can simply delete $i$ from the instance (decreasing all integers larger than $i$ by 1) to obtain an equivalent instance. The family $F$ will be formed using the points $1, 2, \cdots, K+m$. For each $i \in \{1, 2, \cdots, K\}$, it contains a set $F_i$ of arcs determined by the sets $X_j$ that contain $i$ and a single arc $C_i$ that depends on $\pi^{-1}(i)$. Each $F_i$ is constructed as follows: Let $l_i[1], l_i[2], \cdots, l_i[k(i)]$ denote the indices of the sets $X_j$ that contain $i$, listed in increasing order. Then $F_i$ consists of the $k(i)$ arcs

$$\begin{aligned} A_{i1} &= (i, K + l_i[1]), \\ A_{i2} &= (K + l_i[1], K + l_i[2]), \\ A_{i3} &= (K + l_i[2], K + l_i[3]), \\ &\vdots \\ A_{i,k(i)} &= (K + l_i[k(i)-1], K + l_i[k(i)]). \end{aligned}$$

Notice that the spans of the arcs in $F_i$ are pairwise disjoint and that the union of the spans includes exactly the points from $i+1$ up to $K + l_i[k(i)]$. The arc $C_i$ simply spans the region from the end of the last arc in $F_i$ to the beginning of the first arc in $F_{\pi^{-1}(i)}$:

$$C_i = (K + l_i[k(i)], \pi^{-1}(i)).$$

Letting $C = \{C_1, C_2, \cdots, C_K\}$, the family $F$ is defined by

$$F = \bigcup_{i=1}^{K} F_i \cup C.$$

An example of the construction is shown in Fig. 3.

It is easy to see that the family $F$ can be constructed in polynomial time. It remains for us to show that $F$ is $K$-colorable if and only if $\pi \in P$.

To do this, we first consider all possible ways of $K$-coloring the alternative family $F'$, which uses the points $1, 2, \cdots, K+m+1$ and which is derived from $F$ by replacing each arc $C_i = (K + l_i[k(i)], \pi^{-1}(i)) \in C$ by the two arcs $(K + l_i[k(i)], K+m+1)$ and $(K+m+1, \pi^{-1}(i))$. Let $F'_i$, $1 \leqq i \leqq K$, denote the subset of $F'$ that consists of all arcs in $F_i$, the arc $(K + l_i[k(i)], K+m+1)$, and the arc $(K+m+1, i)$. Then the sets $F'_i$ form a partition of $F'$, and each set $F'_i$ is made up of a collection of pairwise disjoint arcs that together span all the points $p$, $1 \leqq p \leqq K+m+1$. It follows that at each such point $p$ all $K$ colors must be distributed among the $K$ arcs (one from each $F'_i$) that span $p$.

Any $K$-coloring of $F'$ can be described by a collection of functions $\sigma_p$, $1 \leqq p \leqq K+m+1$, where $\sigma_p(j)$ denotes that index $i \in \{1, 2, \cdots, K\}$ such that, among all the arcs spanning point $p$, color $j$ is assigned to the one from $F'_i$. Thus each $\sigma_p$ is a permutation of $\{1, 2, \cdots, K\}$. Without loss of generality we can assume that $\sigma_1(j) = j$ for all $j$, i.e., that

each arc of the form $(K+m+1, i)$ is assigned color $i$. Furthermore, we observe that in each set $F'_i$ the two arcs $(K+m+1, i)$ and $(i, K+l_i[1])$ both intersect the $K-1$ arcs $(K+m+1, k)$, $i+1 \leqq k \leqq K$, and $(k, K+l_k[1])$, $1 \leqq k \leqq i-1$, so that both these arcs

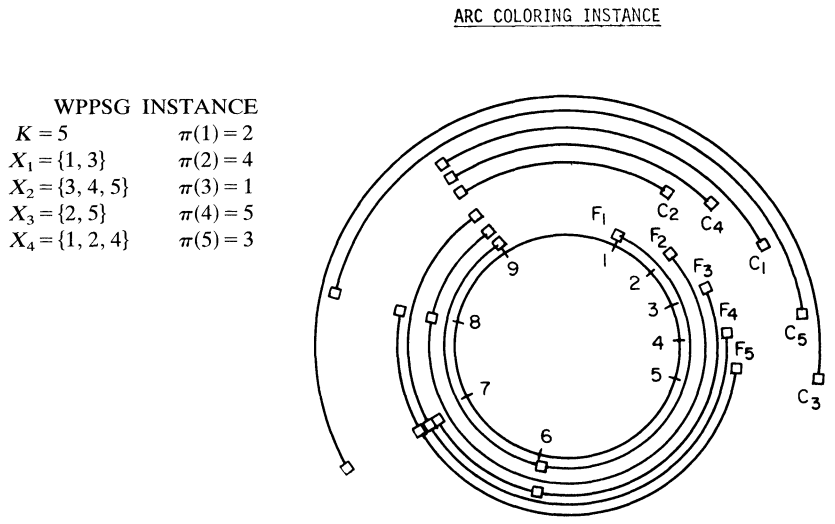



FIG. 3. *An instance of WPPSG and the corresponding instance of ARC COLORING constructed from it.*

must have the same color. Thus, in any $K$-coloring of $F'$, we have

$$\sigma_1 = \sigma_2 = \cdots = \sigma_{K+1}.$$

We next examine how $\sigma_{p+1}$ can be formed from $\sigma_p$, $K+1 \leqq p < K+m+1$. If $\sigma_p(j) = i$ and the arc from $F'_i$ spanning point $p$ also spans point $p+1$, then we necessarily must have $\sigma_p(j) = \sigma_{p+1}(j) = i$. Thus the only cases in which $\sigma_{p+1}(j)$ can differ from $\sigma_p(j)$ are those in which $F'_{\sigma_p(j)}$ contains an arc that ends (and, by construction, another arc that starts) at the point $p$. The colors assigned to the sets having this property by $\sigma_p$ can be arbitrarily redistributed in forming $\sigma_{p+1}$. However, by our construction, these are exactly the sets $F'_i$ such that $i \in X_{p-K}$. Therefore, we can write $\sigma_{p+1} = \sigma_p \cdot \pi_{p-K}$ where $\pi_{p-K} \in S_{X_{p-K}}$. Furthermore, any such choice of $\pi_{p-K}$ provides a legal way of redistributing colors at this point.

Thus the possible "final" permutations $\sigma_{K+m+1}$ that can be obtained by $K$-colorings of $F'$ have a particularly simple structure. They are exactly those permutations that can be written as $\pi_1 \cdot \pi_2 \cdot \cdots \cdot \pi_m$ where each $\pi_i$ belongs to $S_{X_i}$, i.e., they comprise the set $P = S_{X_1} \cdot S_{X_2} \cdot \cdots \cdot S_{X_m}$.

Recalling that $F'$ was obtained from $F$ by "splitting" each arc $C_i \in C$ into two parts, we observe that the $K$-colorings of $F'$ that correspond to $K$-colorings of $F$ are exactly those in which both parts of each $C_i$ are assigned the same color. To interpret this in terms of the $\sigma_p$, notice that one "part" of $C_i$, the arc $(K+l_i[k(i)], K+m+1)$, was placed in the set $F'_i$, whereas the other "part", the arc $(K+m+1, \pi^{-1}(i))$, was placed in the set $F'_{\pi^{-1}(i)}$. Thus, in order for both parts of $C_i$ to have the same color, we must have

$$\sigma^{-1}_{K+m+1}(i) = \sigma_1^{-1}(\pi^{-1}(i)).$$

Since this equality must hold for all $i \in \{1, 2, \cdots, K\}$, and since $\sigma_1$ was assumed to be the identity permutation, this implies that a $K$-coloring of $F'$ corresponds to a $K$-coloring of $F$ if and only if $\sigma_{K+m+1}^{-1} = \pi^{-1}$, or $\sigma_{K+m+1} = \pi$. But the set of possible values for $\sigma_{K+m+1}$ is exactly the set $P$, so $F$ is $K$-colorable if and only if $\pi$ belongs to $P$, which is what we set out to prove.

For the transformation in the other direction, suppose that we are given a family $F$ of circular arcs and a number $K$ of colors. Let $m$ be the largest integer used in the description of the arcs in $F$. Without loss of generality, we may assume that each point $p$, $1 \leqq p \leqq m$, is spanned by exactly $K$ arcs from $F$. If some point $p$ is spanned by more than $K$ arcs, then this can easily be discovered in polynomial time, and it implies that the answer in this instance must be "no." If some point $p$ is spanned by $k < K$ arcs, then we can add $K - k$ arcs of the form $(p-1, p)$ (or $(m, 1)$ if $p = 1$) to $F$ without changing its $K$-colorability.

Given an $F$ of the above form, we first modify it to form an equivalent family $F^*$ on the points $1, 2, \cdots, K+m$. Let $D_1, D_2, \cdots, D_K$ be any ordering of the arcs in $F$ that span the point 1. Then we replace each arc $(a, b) \in F - \{D_1, D_2, \cdots, D_K\}$ by the arc $(K+a, K+b) \in F^*$, and we replace each arc $D_i = (a, b)$ by the two arcs $(K+a, i)$ and $(i, K+b)$. Since the two arcs replacing each $D_i$ must necessarily have the same color in any $K$-coloring of $F^*$, it follows immediately that $F^*$ is $K$-colorable if and only if $F$ is $K$-colorable.

The gist of the argument from this point on is that $F^*$ has the same type of structure as the family $F$ constructed in the first half of the proof, so all we need to do is to invert the transformation used there. In order to bring out the structure of $F^*$, we shall partition it into sets $F_i$, $1 \leqq i \leqq K$, and $C$. The set $C$ consists of exactly those arcs in $F^*$ that contain the point 1 in their spans. The sets $F_i$ will be constructed in the order $F_1, F_2, \cdots, F_K$, with a particular $F_i$ being formed by selecting certain arcs from the set

$$R(i) = F^* - C - \bigcup_{j=1}^{i-1} F_j$$

as follows: The first arc selected to be in $F_i$ is the single arc in $F^*$ that has $i$ as its left endpoint. Then, so long as there exists an arc in $R(i)$ whose left endpoint is the same as the right endpoint of the last arc added to $F_i$, we choose one such arc and add it to $R(i)$. Thus each $F_i$ will consist of a collection of disjoint arcs that span all points from $i+1$ up to some point $P_i$ (and no others). We also index the arcs in $C$ as $C_1, C_2, \cdots, C_K$ in such a way that the left endpoint of arc $C_i$ is the same as the right endpoint $P_i$ of the last arc added to $F_i$. The fact that every point is spanned by exactly $K$ arcs from $F^*$ enables all of this to be done.

Now we are in a position to construct the sets $X_1, X_2, \cdots, X_m$ and permutation $\pi$ for the corresponding WPPSG instance. The set $X_j$ consists of those integers $i \in \{1, 2, \cdots, K\}$ such that $F_i$ contains an arc with right endpoint $K+j$. The permutation $\pi$ has $\pi(i) = j$ if and only if the arc $C_j$ has right endpoint $i$.

It is not difficult to see that this transformation can be performed in polynomial time. It is also straightforward to verify that if the transformation from the first half of the proof is applied to the WPPSG instance, the resulting ARC COLORING instance is exactly $F^*$. Hence exactly the same argument as used for that transformation suffices to show that $\pi \in S_{X_1} \cdot S_{X_2} \cdot \cdots \cdot S_{X_m}$ if and only if $F^*$ is $K$-colorable, and our proof is complete. □

There is an obvious algorithm for solving the WPPSG problem—and, therefore, the ARC COLORING problem, via the transformation of Theorem 1. Given $X_1, X_2, \cdots, X_m$, one simply computes all elements of the set $L = S_{X_1} \cdot S_{X_2} \cdot \cdots \cdot S_{X_m}$ by

starting with the set of permutations $P_0 = \{e\}$ and successively constructing $P_{j+1} = \{\pi_1 \cdot \pi_2 : \pi_1 \in P_j \text{ and } \pi_2 \in S_{X_{j+1}}\}$, $j = 0, 1, \cdots, m-1$. The set $P$ is then given by $P_m$, and we can easily check whether $\pi$ belongs to it.

To analyze this algorithm, we observe that multiplication of two permutations over $\{1, 2, \cdots, K\}$ can be done in $O(K)$ operations, and, in order to store sets of permutations, we may assume that each permutation $\sigma \in S_K$ is associated with a distinct integer $I(\sigma)$, $1 \leqq I(\sigma) \leqq K!$, in such a way that $\sigma$ can be computed from $I(\sigma)$ and $I(\sigma)$ computed from $\sigma$ in $O(K \log K)$ operations (e.g., see [9, pp. 19, 579]). Then the space required for the algorithm is $O(K!)$, and the time required is $O(m \cdot (K!)^2 \cdot K \cdot \log K)$.

This time complexity can be improved, however, by making use of the fact that, if $\pi_1 \in P_j$, $\pi_2 \in S_{X_{j+1}}$, and $\pi_1 \cdot \pi_2 = \pi_1' \in P_j$, then $\pi_1 \cdot S_{X_{j+1}} = \pi_1' \cdot S_{X_{j+1}}$, so we need not compute any of the products involving $\pi_1'$. Hence we can compute $P_{j+1}$ from $P_j$ as follows:

*Step* 1. Select a permutation $\pi_1$ from $P_j$ and remove it from $P_j$.

*Step* 2. For each permutation $\pi_2 \in S_{X_{j+1}}$,

(a) add $\pi_1 \cdot \pi_2$ to $P_{j+1}$;

(b) remove $\pi_1 \cdot \pi_2$ from $P_j$ (if it's there).

*Step* 3. If $P_j$ is nonempty, return to Step 1.

This method for computing $P_{j+1}$ from $P_j$ has the property that each product gives us a new member of $P_{j+1}$. Thus it requires at most $K!$ products and at most $O(K!)$ conversions between a permutation $\sigma$ and its index $I(\sigma)$. Using this method, the time for the overall algorithm therefore becomes $O(m \cdot K! \cdot K \cdot \log K)$.

The transformation from ARC COLORING to WPPSG given in the proof of Theorem 1 can be implemented easily to run in time $O(K \cdot n)$. Thus we have the following corollary:

COROLLARY. *Deciding whether a family of n circular arcs is K-colorable can be done in* $O(n \cdot K! \cdot K \cdot \log K)$ *time.*

The same time complexity suffices for *constructing* a $K$-coloring, since in solving the WPPSG instance we can easily save enough information to allow us to reconstruct a sequence of permutations whose product is $\pi$. Thus, for any fixed value of $K$, the circular arc coloring problem can be solved in linear time, and for small values of $K$ the algorithm might actually be practical.

**3. ARC COLORING and WPPSG are *NP*-complete.** In this section we show that WPPSG is *NP*-complete. By the results of the preceding section, this will imply that ARC COLORING is *NP*-complete. The latter result will in turn imply that CHORD COLORING is *NP*-complete, as we shall see in the next section. In all three cases, we leave to the reader the straightforward verification that the problem in question is in *NP*.

THEOREM 2. *WPPSG is NP-complete.*

*Proof.* The known *NP*-complete problem that we transform to WPPSG is the following:

*DIRECTED DISJOINT CONNECTING PATHS* (*DDCP*). Given a directed acyclic graph $G = (V, A)$, an ordering $s_1, s_2, \cdots, s_n$ of the vertices with in-degree 0, and an ordering $t_1, t_2, \cdots, t_n$ of the vertices with out-degree 0 (we may assume that the two sets have the same size), does $G$ contain $n$ mutually vertex-disjoint paths, each going from a distinct $s_i$ to the corresponding $t_i$, $1 \leqq i \leqq n$?

The undirected version of this problem was proved *NP*-complete by Knuth (see [7]), and the directed acyclic version can be proved *NP*-complete by a trivial modification of his proof.

Suppose we are given an instance $G=(V, A)$, $s_1, s_2, \cdots, s_n$, and $t_1, t_2, \cdots, t_n$ of the DDCP problem. The first step of the transformation is to replace each arc $a=(u, v)\in A$ by two arcs $(u, w_a)$ and $(w_a, v)$, where $w_a$ is a new vertex involved in only these two arcs. This certainly has no effect of the existence of the desired paths. Let $G'=(V', A')$ denote the resulting directed graph.

Now, let $v_1, v_2, \cdots, v_q$ be any topological sorting of the vertices of $G'$, i.e., any ordering such that, for each $i$, all vertices $x$ for which $(v_i, x)\in A'$ come after $v_i$ in the sequence. Such an ordering can be constructed in time linear in $|A'|$. (See, e.g., [Knuth, Vol. 1]). Furthermore, without loss of generality, we may assume that $v_i=s_i$ for $1\leqq i\leqq n$ and that $v_{q-n+i}=t_i$ for $1\leqq i\leqq n$.

For each vertex $v_i$, let

$$B(i)=\{j:(v_j, v_i)\in A'\}.$$

The sets $X_j$, $1\leqq j\leqq m=q-n+1$, for the corresponding WPPSG instance are then defined as follows:

$$X_j=\{n+j\}\cup B(n+j), \qquad 1\leqq j\leqq q-n$$

$$X_{q-n+1}=\{1, 2, \cdots, q-n\}.$$

The permutation $\pi$ is defined by:

$$\pi(i)=q-n+i, \qquad 1\leqq i\leqq n$$

$$\pi(i)=i-n, \qquad n+1\leqq i\leqq q.$$

This transformation is easily performed in polynomial time. It remains for us to show that $\pi\in P=S_{X_1}\cdot S_{X_2}\cdot\cdots\cdot S_{X_m}$ if and only if the desired paths from each $s_i$ to each $t_i$ exist in $G'$.

First, let us examine how the WPPSG instance can be interpreted in terms of the graph $G'$. Each position in a permutation corresponds to a vertex of $G'$. Initially, each such position/vertex is labeled by its own index. When we apply a permutation $\pi_i$ from some $S_{X_i}$ we move the labels around on some subset of vertices, specifically on some subset of the vertices whose indices belong to the set $X_i$. Furthermore, the set $X_i$ contains precisely the indices of vertex $v_{n+i}$ and its immediate predecessors in $G'$. Thus the process of choosing a sequence of permutations $\pi_1, \pi_2, \cdots, \pi_m$, each $\pi_i\in S_{X_i}$, corresponds exactly to choosing a sequence of label rearrangements, first among $v_{n+1}$ and its immediate predecessors, then among $v_{n+2}$ and its immediate predecessors, and so on, until finally we are allowed to rearrange the labels on all vertices in $V'-\{t_1, t_2, \cdots, t_n\}$. Our goal is to move each label $i$, $1\leqq i\leqq n$, all the way from vertex $s_i=v_i$ to the corresponding vertex $t_i=v_{q-n+i}$. Once this has been done, the final permutation can be chosen to arbitrarily rearrange the labels on the vertices outside of $\{t_1, t_2, \cdots, t_n\}$. (In essence, we don't really care what labels end up on these vertices, but the WPPSG problem requires that the entire permutation (i.e., the complete final labeling) be specified.) Thus the permutation $\pi$ belongs to $P$ if and only if the above relabeling process can be performed in such a way that the label $i$ ends up on vertex $v_{\pi(i)}$, $1\leqq i\leqq q$.

Given this interpretation, it is not difficult to see that the transformation works as required. Suppose that $G'$ does contain a set of vertex-disjoint paths, one from each $s_i$ to the corresponding $t_i$, $1\leqq i\leqq n$. Let $A^*\subset A'$ denote the set of all arcs that occur in these $n$ paths. Notice that, since the paths are disjoint, no vertex will appear more than once as right endpoint of an arc in $A^*$. The $j$th step of the corresponding relabeling process, $1\leqq j\leqq q-n$, is performed as follows: At the $j$th step we are allowed to rearrange the labels that occur on vertex $v_{n+j}$ and its immediate predecessors. If there is some arc of

the form $(u, v_{n+j}) \in A^*$, then we simply interchange the labels on $u$ and $v_{n+j}$, leaving all other labels where they were. If there is no such arc in $A^*$, then no labels at all are moved. Since the given paths are disjoint and since $V'$ is indexed in topological order, it is straightforward to verify that this process will succeed in moving each label $i$, $1 \leqq i \leqq n$, from $s_i$ to $t_i$ by the end of step $q-n$. Step $q-n+1$ then can rearrange the labels on the vertices in $V'-\{t_1, t_2, \cdots, t_n\}$ from where they have been left by the preceding steps to where they are required to be. Thus the existence of the specified disjoint paths implies the existence of the required relabeling sequence, which in turn implies that $\pi \in P = S_{X_1} \cdot S_{X_2} \cdot \cdots \cdot S_{X_m}$.

For the other direction, suppose there exist $\pi_i \in S_{X_i}$, $1 \leqq i \leqq m$, such that $\pi = \pi_1 \cdot \pi_2 \cdot \cdots \cdot \pi_m$, and consider the corresponding relabeling process on $G'$. For $1 \leqq i \leqq n$, we know that label $i$ starts out on $s_i$ and ends up on the corresponding $t_i$. What we need to show is that each label $i$ moves only along a path in $G'$ and that the paths for two such labels never intersect at a vertex. For the first of these, suppose that at the $j$th relabeling step label $i$ is moved, but not along an arc of $G'$ (or not in the proper direction). The topological ordering of $V'$ insures that $X_j$ is the first set to contain $v_{n+j}$, so label $i$ could not have appeared on $v_{n+j}$ at the beginning of this step. Thus step $j$ must move label $i$ from one immediate predecessor of $v_{n+j}$ to another such immediate predecessor. This implies that $v_{n+j}$ must be one of the original vertices of $G$, because each of the vertices added to $G$ in forming $G'$ has only *one* immediate predecessor. In this case, however, we know that each immediate predecessor of $v_{n+j}$ has only one arc leaving it in $G'$, the one to $v_{n+j}$, so no immediate predecessors of $v_{n+j}$ occur in any sets after $X_j$. Thus such a "parallel move" of label $i$ would prevent it from ever reaching $t_i$, a contradiction which proves that the labels $1, 2, \cdots, n$ move only along paths in $G'$. To see that two such paths cannot intersect, we simply need to observe that the only time a label $i$, $1 \leqq i \leqq n$, can move to a vertex $v_{n+j}$ by moving along an arc of $G'$ is at step $j$, and only one such label can be moved to $v_{n+j}$ during that step. Thus the paths followed by these labels must be disjoint, and the proof is complete. □

As a consequence of Theorems 1 and 2, we immediately have the following:

COROLLARY. *ARC COLORING is NP-complete.*

We can also make a remark about an interesting special case of WPPSG, that in which each set $X_i$ contains only two elements. This is simply the problem of determining, given a permutation $\pi$ and a sequence of pairwise interchanges, whether $\pi$ can be realized by performing some subsequence of the given interchanges. Let us call this problem WPPSG2. We can transform any instance of WPPSG to an equivalent instance of WPPSG2 by replacing each set $X_i = \{a_1, a_2, \cdots, a_l\}$ by the following sequence of $\binom{l}{2}$ two-sets:

$$\begin{aligned} &\{a_1, a_2\}, \{a_1, a_3\}, \quad \cdots \quad, \{a_1, a_l\}, \\ &\qquad \{a_2, a_3\}, \quad \cdots \quad, \{a_2, a_l\}, \\ &\qquad\qquad \vdots \\ &\{a_{l-2}, a_{l-1}\}, \{a_{l-2}, a_l\}, \\ &\qquad\qquad \{a_{l-1}, a_l\} \end{aligned}$$

Thus we have as a corollary of Theorem 2:

COROLLARY. *WPPSG2 is NP-complete.*

**4. Chord coloring.** We model the chord coloring problem as follows: A *chord*, like a circular arc, is a pair $(a, b)$ of integers. The difference between an arc and a chord lies in the way we interpret such a pair. If the integers occurring in a family of chords (or arcs) are arranged in clockwise order around a circle, the chord $(a, b)$ is viewed as the straight line connecting $a$ and $b$, whereas the arc $(a, b)$ is viewed as the arc of the circle (in the clockwise direction) from $a$ to $b$. Note that, in this interpretation, the chords $(a, b)$ and $(b, a)$ are identical, but the arcs $(a, b)$ and $(b, a)$ are different (and complements of each other). We can, however, identify the chord $(a, b)$ with the *shorter* of the two arcs $(a, b)$ and $(b, a)$, i.e., the one with smaller span (as defined in § 2), breaking ties arbitrarily. Then it is easy to see that two chords intersect if and only if the corresponding two arcs $A_i$ and $A_j$ *overlap*, in the sense that their spans intersect and, in addition, neither $sp(A_i) \subseteq sp(A_j)$ nor $sp(A_j) \subseteq sp(A_i)$. In order to avoid any confusion when we use this identification in what follows, we shall always refer to chords as "overlapping" rather than intersecting.

The *circle graph* corresponding to a family $F = \{A_1, A_2, \cdots, A_n\}$ of chords is the graph $G = (F, E)$ where $\{A_i, A_j\} \in E$ if and only if the chords $A_i$ and $A_j$ overlap. The chord coloring problem is then defined as follows:

*CHORD COLORING.* Given a family $F$ of chords and a positive integer $K$, can $F$ be partitioned into $K$ classes so that no two chords in the same class overlap? (Or, equivalently, can the circle graph $G = (F, E)$ be colored with $K$ colors?)

The main result of this section shows that CHORD COLORING is at least as hard as ARC COLORING.

THEOREM 3. *CHORD COLORING is NP-complete.*

*Proof.* We derive this result by showing that ARC COLORING is polynomially transformable to CHORD COLORING. Given an instance, $F, K$ of ARC COLORING, we shall show how to construct in polynomial time a family $F'$ of chords such that the chords in $F'$ are $K$-colorable if and only if the arcs in $F$ are $K$-colorable. The idea behind the construction is quite simple and can be summarized as follows: If we view chords in terms of their corresponding arcs, arc coloring and chord coloring are almost identical problems, differing only in cases where one arc is contained in another (see Fig. 4(a)). We are going to *remove* all such occurrences of containment from $F$ by replacing each arc by a sequence of small chords (Fig. 4(b)). However, we must ensure that all small chords replacing a particular original arc behave like a single arc, in the sense that they all must be given the same color. We do this by adding a "clique" of $K - 1$ chords at each of the junction points (Fig. 4(c) shows the details around the junction points circled in Fig. 4(b)).

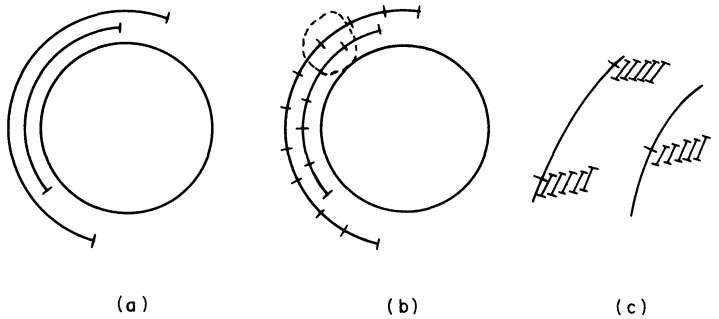



FIG. 4. *An instance of arc containment (a), the result of replacing each arc by a sequence of small "chords" (b), and a "closeup" showing how "cliques" are added at junction points (c).*

Formally, let $F$ be the given family of $n$ arcs, with $m \leqq 2n$ denoting the largest integer used in their descriptions, and let $K$ be the specified number of colors. For each arc $A_i = (a_i, b_i) \in F$ and each point $p \in sp(A_i)$, $F'$ contains the chords

$$(2K(2pn-(n+i)), 2K(2(p+1)n-i)), \qquad p = a_i + 1,$$

$$(2K(2pn-i), 2K(2(p+1)n-i)), \qquad a_i + 1 < p < b_i.$$

Furthermore, it is easy to verify that two original arcs intersect if and only if there are two chords derived from them that overlap. Now consider each pair of chords, derived from the same original arc, that share a common endpoint. By the construction, that common endpoint has the form $2Kx$ for some integer $x$. We then add the following "clique" chords, all containing the point $2Kx$ in their spans:

$$\begin{aligned}
&(2Kx-1, 2Kx+K-1),\\
&(2Kx-2, 2Kx+K-2),\\
&\quad\vdots\\
&(2Kx-(K-1), 2Kx+1).
\end{aligned}$$

Observe that these $K-1$ "clique" chords all overlap one another and, in addition, they all overlap the two chords that share endpoint $2Kx$. Furthermore, these are the *only* chords that they overlap.

Since each $A_i$ satisfies $|sp(A_i)| \leqq m \leqq 2n$, the above construction clearly can be performed in polynomial time. By sorting all the chord endpoints and replacing each endpoint by its rank in the sorted order, all of which can be done in polynomial time, we also can convert the set of chords into one having the same intersection pattern and having a description using no integer larger than twice the total number of chords. For convenience, however, we shall continue to work with the "un-condensed" version in the remainder of the proof.

We claim that the arcs of $F$ are $K$-colorable if and only if the chords of $F'$ (using "overlap" instead of "intersect") are $K$-colorable. Given any $K$-coloring of $F$, let $C(A_i)$ denote the color used for arc $A_i$. Then, for each $A_i$, we color all the chords in $F'$ derived from $A_i$ with color $C(A_i)$. Since two chords derived from the same arc $A_i$ do not overlap and since two chords derived from different arcs $A_i$ and $A_j$ do not overlap unless $A_i$ and $A_j$ intersect (in which case we know that $C(A_i) \neq C(A_j)$), this "partial" coloring correctly assigns different colors to overlapping chords. All that remains is to color the various "clique" chords. Consider the clique chords surrounding some point $2Kx$ that is a common endpoint of two chords derived from a particular arc $A_i$. Since these $K-1$ clique chords overlap only one another and two chords already colored with color $C(A_i)$, we may color each of them with a different one of the remaining $K-1$ colors. Doing this for each such set of clique chords, we finally obtain a $K$-coloring for the chords in $F'$.

On the other hand, suppose that we have a $K$-coloring for the chords in $F'$. Consider any two chords that share a common endpoint and that are derived from the same original arc $A_i$. These two chords must be assigned the same color, since both overlap all the $K-1$ clique chords surrounding their common endpoint and $K-1$ distinct colors must be used on those clique chords. It follows that, for each original arc $A_i$, *all* chords in $F'$ derived from $A'$ must have the same color. Thus, we can obtain a $K$-coloring for the arcs in $F$ by assigning to each arc $A_i$ the same color that is assigned to all the chords derived from $A_i$. This is a legal $K$-coloring, because two arcs $A_i$ and $A_j$ in

$F$ intersect only if two chords in $F'$ derived from them overlap, and because the given coloring for $F'$ assigned different colors to any two chords that overlap. □

**5. Conclusion.** In this paper we have shown that the word problem for products of symmetric groups is *NP*-complete, and from this have derived the *NP*-completeness of graph coloring, even when restricted to circular arc graphs or circle graphs. Although we have not given formal definitions for the register allocation problem and the problem of realizing a permutation with parallel stacks, which were claimed to be equivalent to circular arc graph and circle graph coloring in § 1, the *NP*-completeness of these problems also follows from our results. (The reader may fill in the details by looking up the formal definitions in [2], [12].)

A number of open questions remain. In § 2 we were able to present an algorithm which, for any fixed $K$, ran in polynomial time and produced a $K$-coloring of a family of circular arcs if one existed. Does a similar algorithm exist for the chord coloring problem, or is there, as with general graph coloring, some fixed $K$ for which the chord coloring problem is *NP*-complete? What is the complexity of the coloring problem for *proper* circular arc graphs (graphs representable by families of arcs which intersect if and only if they overlap)?

More basically, is there a polynomial time algorithm for recognizing circle graphs and constructing their representations in terms of chords (or arcs)? Such algorithms have been found by Tucker for circular arc graphs [13] and proper circular arc graphs [10]. A similar algorithm for circle graphs might well widen the usefulness of the algorithms in [5], as these assume that the representation of the circle graph is known.

## REFERENCES


[1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] S. EVEN AND A. ITAI, *Queues, stacks, and graphs*, Theory of Machines and Computations, Z. Kohavi and A. Paz, eds., Academic Press, New York, 1971, pp. 71–86.

[3] M. R. GAREY, D. S. JOHNSON AND L. J. STOCKMEYER, *Some simplified NP-complete graph problems*, Theor. Comput. Sci., 1 (1976), pp. 237–267.

[4] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, CA, 1979.

[5] F. GAVRIL, *Algorithms for a maximum clique and a maximum independent set of a circle graph*, Networks, 3 (1973), pp. 261–273.

[6] ———, *Algorithms on circular-arc graphs*, Networks, 4 (1974), pp. 357–369.

[7] R. M. KARP, *On the complexity of combinatorial problems*, Networks, 5 (1975), pp. 45–68.

[8] V. KLEE, *What are the intersection graphs of arcs in a circle?*, Amer. Math. Monthly, 76 (1969), pp. 810–813.

[9] D. E. KNUTH, *The Art of Computer Programming*, Vol. 3, Addison-Wesley, Reading, MA, 1973.

[10] A. TUCKER, *Matrix characterizations of circular-arc graphs*, Pacific J. Math., 39 (1971), pp. 535–545.

[11] ———, *Structure theorems for some circular-arc graphs*, Discrete Math. 7 (1974), pp. 167–195.

[12] ———, *Coloring a family of circular arcs*, SIAM J. Appl. Math., 29 (1975), pp. 493–502.

[13] ———, *An efficient test for circular-arc graphs*, SIAM J. Comput., 9 (1980), pp. 1–24.

SIAM J. ALG. DISC. METH.
Vol. 1. No. 2. June 1980


0196-5212/80/0102-0010 $01.00/0

# RECONSTRUCTION OF A PAIR OF GRAPHS FROM THEIR CONCATENATIONS*

SUKHAMAY KUNDU†, E. SAMPATHKUMAR‡, JAMES SHEARER¶ AND DEAN STURTEVANT§

**Abstract.** A concatenation of two vertex disjoint graphs is defined to be the graph obtained by identifying a vertex of one graph with a vertex of the other graph. We show that an arbitrary pair of connected graphs can be uniquely reconstructed from the set of all their distinct concatenations, with only trivial exceptions.

**1. The reconstruction problem.** Let $G_1$ and $G_2$ be two vertex disjoint graphs, and let $x_i (i = 1, 2)$ be an arbitrary vertex in $G_i$. The graph $G(x_1x_2)$ obtained by merging the vertices $x_1$ and $x_2$ into a single vertex is called the *concatenation* of $G_1$ and $G_2$ at the points $x_1$ and $x_2$. The reconstruction problem after concatenation is to determine the original pair of graphs $G_1$ and $G_2$ given the family of all concatenations $G(x_1x_2)$ obtained by varying the vertices $x_1$ and $x_2$ in the respective graphs. As in other graph reconstruction problems, it is understood that the vertices in $G(x_1x_2)$ are unlabeled, and that the concatenated vertex in $G(x_1x_2)$ is not distinguishable as such from other nodes.

We show that with only trivial exceptions any pair of connected graphs is uniquely determined from its concatenations. The reconstruction theorem given here requires only that we know the distinct nonisomorphic $G(x_1x_2)$'s, and not how many times each concatenation appears among all concatenations.

The original graph reconstruction problem [12] states that, for all graphs with three or more vertices, $G$ can be uniquely reconstructed from the family of graphs $\{G\backslash x\}$, $G\backslash x$ being the subgraph of $G$ obtained by deleting the vertex $x$ and all edges incident with $x$. This particular reconstruction property has been established for a large set of graphs including trees, outer planar graphs, unicyclic graphs and disconnected graphs [1]–[7]. Graph reconstructions from elementary contractions and elementary partitions were considered in [8]–[10]. In all cases, the proofs depend heavily on the particular structural properties assumed for the graphs. No unified technique is yet available for general graph reconstruction problems.

**2. Preliminaries.** Let $G = (V, E)$ be a connected graph (unless otherwise specified, all graphs discussed in this paper are assumed to be connected). For $v \in v$, define the *cutdegree* of $v$, denoted $\mathrm{cd}(v)$, to be the number of connected components in the induced subgraph $G\backslash v$ on $V\backslash\{v\}$. $v$ is a *cutvertex* if $\mathrm{cd}(v) > 1$. (Note that if $G$ is a tree, then the cutdegree of a vertex is the same as its ordinary degree.) A *block* of $G$ is a maximal subgraph without cutvertices. A *limb* (at a vertex $v$) is a component of $G\backslash v$ together with $v$ and all the edges in $E$ joining $v$ to that component.

A *rooted graph* is a pair $(G, v)$, where $G = (V, E)$ is a graph and $v \in V$. Two rooted graphs $(G_1, v_1)$, $(G_2, v_2)$ are *isomorphic* if there is a graph isomorphism $\phi: G_1 \to G_2$ such that $\phi(v_1) = v_2$. The *concatenation* of two rooted graphs $(G_1, v_1)$ and $(G_2, v_2)$, denoted

* Received by the editors July 28, 1978, and in final revised form November 15, 1979.

† Logicon, Inc., Lexington, Massachusetts 02173. Now at Bell Laboratories, Murray Hill, New Jersey 07974.

‡ Department of Mathematics, Karnataka University, Dharwar Karnataka 580003, India.

¶ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author 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. The work of this author was supported in part by a National Science Foundation Graduate Fellowship.

by $(G_1, v_1) \circ (G_2, v_2)$ is the rooted graph $(G(v_1v_2), v_1v_2)$, the concatenated vertex being denoted by $v_1v_2$.

For a graph $G$, define the sequence $\underset{\sim}{\alpha}(G) = (\alpha_1, \alpha_2, \cdots)$ by $\alpha_i = \alpha_i(G) =$ number of vertices of cutdegree $i$ (if $G$ is trivial, $\alpha(G) = (0, 0, \cdots)$). Define $\Delta(G) = \max\{i: \alpha_i \neq 0\}$; $\Delta(G) = 1$ if $G$ is a block. Thus, if $G = (G_1, v_1) \circ (G_2, v_2)$, we have $\underset{\sim}{\alpha}(G) = \underset{\sim}{\alpha}(G_1) + \underset{\sim}{\alpha}(G_2) + \underset{\sim}{\varepsilon}(\mathrm{cd}\,(v_1) + \mathrm{cd}\,(v_2)) - \underset{\sim}{\varepsilon}(\mathrm{cd}\,(v_1)) - \underset{\sim}{\varepsilon}(\mathrm{cd}\,(v_2))$. Here addition is componentwise, and $\underset{\sim}{\varepsilon}(i)$ is the sequence with $i$th entry equal to one, the rest zero.

LEMMA 1. *In any limb $L$ at a vertex $x$ of a connected nontrivial graph $G$, there is a vertex $v \neq x$ of cutdegree one in $G$.*

*Proof.* Let a spanning tree of $G$ be given. The subtree which spans $L$ will have a vertex $v \neq x$ of degree one; $v$ is the desired vertex. □

LEMMA 2. *Suppose $G_1$ and $G_2$ are known to be nontrivial. Then $\underset{\sim}{\alpha}(G_1) + \underset{\sim}{\alpha}(G_2)$ and $\{\Delta(G_1), \Delta(G_2)\}$ can be determined from the set of nonisomorphic concatenations of $G_1$ and $G_2$.*

*Proof.* Let $C$ be a concatenation with lexicographically least $\underset{\sim}{\alpha}(C)$. By Lemma 1, this must arise when two vertices of cutdegree one are identified, so $\underset{\sim}{\alpha}(C) = \underset{\sim}{\alpha}(G_1) + \underset{\sim}{\alpha}(G_2) + (-2, 1, 0, \cdots)$ and $\underset{\sim}{\alpha}(G_1) + \underset{\sim}{\alpha}(G_2) = \underset{\sim}{\alpha}(C) + (2, -1, 0, \cdots)$. Furthermore, $\max\{\Delta(G_1), \Delta(G_2)\} = \max\{i: \alpha_i(G_1) + \alpha_i(G_2) \neq 0\}$.

Let $C'$ be a concatenation with $\Delta(C')$ maximum. This must arise when two vertices of maximum cutdegree in $G_1$ and $G_2$ are identified. Hence, $\Delta(C') = \Delta(G_1) + \Delta(G_2)$ and $\min\{\Delta(G_1), \Delta(G_2)\} = \Delta(C') - \max\{\Delta(G_1), \Delta(G_2)\}$. □

**3. Main theorems.**

THEOREM 3. *Suppose the graphs $G_1$ and $G_2$ are known to be nontrivial. Then they can be determined, up to isomorphism, by the set of their (nonisomorphic) concatenations.*

*Proof.* By the preceding lemma, we may distinguish four cases.

*Case* 1. $\Delta(G_1) > \Delta(G_2) > 1$. Let $C$ be a concatenation with $\Delta(C)$ maximum. The concatenation point is the unique vertex of cutdegree $\Delta(C)$. Define $\mathscr{L}(C) = \{L: L$ is a limb of $C$ at a vertex of cutdegree $\geqq \Delta(G_1)$ and $L$ has no vertices of cutdegree $\Delta(G_1)\}$. Notice that $\mathscr{L}(C)$ is $\mathscr{L}(G_1)$ together with at least two proper subgraphs of $G_2$, since $\Delta(G_2) > 1$.

Let $C'$ be a concatenation with $\alpha_{\Delta(G_1)}(G_1) - 1$ vertices of cutdegree $\Delta(G_1)$ and one vertex of cutdegree $\Delta(G_1) + 1$. This must be the result of the identification of a vertex of cutdegree $\Delta(G_1)$ in $G_1$ with a vertex of cutdegree 1 in $G_2$ (existence guaranteed by Lemma 1). $\mathscr{L}(C') = \mathscr{L}(G_1) \cup \{G_2\}$, so $\{G_2\} = \mathscr{L}(C') \backslash \mathscr{L}(C)$. "Plucking" $G_2$ from the (uniquely determined) concatenation vertex in $C'$ leaves $G_1$.

*Case* 2. $\Delta(G_1) = \Delta(G_2) = \Delta > 1$. Let $C$ be a concatenation with $\alpha_\Delta(G_1) + \alpha_\Delta(G_2) - 1$ vertices of degree $\Delta$ and one vertex of degree $\Delta + 1$ with a limb at that vertex of maximum size (over all such concatenations). This limb will be the largest (in number of vertices) of $G_1$ and $G_2$.

*Case* 3. $\Delta(G_1) > \Delta(G_2) = 1$. The *block-cutvertex tree* $T(G)$ associated with a graph $G$ is defined as follows. The vertices of $T(G)$ are the blocks and cutvertices of $G$. Two vertices in $T(G)$ are joined by an edge if and only if one is a cutvertex and the other is a block (in $G$) containing that vertex. It is easily verified that $T(G)$ is a tree.

We consider separately the cases $T(G_1)$ is a path and is not a path. These cases may be distinguished given the set of concatenations, since if $T(G_1)$ is a path, then using Lemma 1, there is a concatenation whose block-cutvertex tree is a path. On the other hand, $T(G_1)$ is a sub-graph of $T(C)$ for every concatenation $C$; so if $T(G_1)$ is not a path, neither is $T(C)$.

If $T(G_1)$ is not a path, there is a vertex in $T(G_1)$ of degree at least three. Let $C$ be a concatenation having the longest path in $T(C)$ from a leaf to a vertex of degree at least

three, subject to $T(C)$ having the minimum number of leaves. The leaf on this path will be the block $G_2$.

When $T(G_1)$ is a path, there are three cases.

*Case* 3a. $T(G_1)$ has 3 vertices, that is, every concatenation has three blocks. (This is the case where $G_1$ consists of 2 blocks with a common cutvertex.) If there is a unique block which appears as a limb in every concatenation, it is $G_2$. $G_1$ can then be determined as follows. Suppose the blocks of $G_1$ are not both $G_2$. Then some concatenation contains a unique limb isomorphic to $G_2$ which may then be plucked from the graph leaving $G_1$. In the remaining case, consider those concatenations in which $G_2$ is joined to the cutvertex of $G_1$. The automorphism group of the graph $G_2$ induces a partition of its vertices into equivalence classes. If the cutpoint of $G_1$ lies in the same class in each block, then, in some concatenation, the cutvertex lies in the same class of all 3 blocks uniquely determining $G_1$. If this case does not occur, then in some concatenation the cutvertex will occur in one class in one block, and in another class in the other two blocks. This again uniquely determines $G_1$. If there are two blocks which appear as limbs in every concatenation, then $G_1$ consists of the concatenation of two identical blocks while $G_2$ is a distinct block. Hence in any concatenation, two blocks are identical and the other is $G_2$.

*Case* 3b. Every concatenation has four blocks. If there is a unique block which appears as vertex $v$ or $\nu$ in every concatenation $C$ with $T(C)$ as in Fig. 1, then this block is $G_2$. This again uniquely determines $G_1$ except when both endblocks of $G_1$ are isomorphic to $G_2$. But in this case, it is easy to see that there will be concatenations in which the blocks corresponding to $v$ and $\nu$ are joined at points in the same equivalence class of $G_2$ uniquely determining $G_1$. Otherwise, the two endblocks of $G_1$ are identical and $G_2$ is the "odd man out" among the endblocks of any concatenation for which $T(C)$ is not a path.

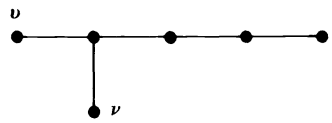



FIG. 1

*Case* 3c. Every concatenation has at least five blocks. There will be a concatenation with a vertex $\nu$ of cutdegree 3 and a unique limb of $\nu$ which is also a block. This block is $G_2$. Plucking this $G_2$ from $\nu$ leaves $G_1$.

*Case* 4. $\Delta(G_1) = \Delta(G_2) = 1$. In any concatenation there will be a unique cutvertex. The limbs at this cutvertex will be the graphs $G_1$ and $G_2$. □

If the restriction that the graphs be nontrivial is relaxed then, in general, it is not true that $\Delta(G_1)$ and $\Delta(G_2)$ may be calculated (if $G_2$ is trivial, then $\alpha(G_1) + \alpha(G_2) = \alpha(C)$ for the unique concatenation $C$, which is why the proof of Lemma 2 breaks down). In fact, any time there is a unique concatenation $C$, it is possible that $G_1 = C$ and $G_2$ is trivial. The cases in which this is not the only possibility is dealt with in the following theorem.

THEOREM 4. *Let $G_1$ and $G_2$ be nontrivial graphs. They have a unique concatenation if and only if they are both vertex transitive.*

*Proof. If.* Clear. *Only if.* Suppose $G_1$ is not vertex transitive.

If $G_1$ has a cutvertex $x$, let $y$ be a vertex of cutdegree one in $G_1$ and $z$ any vertex in $G_2$. Then $\alpha((G_1, x) \circ (G_z, z)) \neq \alpha((G_1, y) \circ (G_z, z))$ and, hence, these two concatenations are different.

If $G_1$ is a block, let $x_1$ and $x_2$ be two inequivalent vertices in it. Let $z$ be a vertex of maximum cutdegree in $G_2$. Then in the concatenations $C_i = (G_1, x_1) \circ (G_2, z)(i = 1, 2)$, the concatenation points are the unique vertices $v_i$ of maximum cutdegree $\Delta(G_2)+1$. Hence, $C_1$ is isomorphic with $C_2$ if and only if the rooted graphs $(C_1, v_1)$ and $(C_2, v_2)$ are isomorphic which is true if and only if the rooted limbs at $v_i$ are (pairwise) isomorphic. But the rooted limbs at $v_i$ are the rooted limbs at $z$ in $G_2$ together with $(G_1, x_i)$. As $(G_1, x_1)$ and $(G_1, x_2)$ are not isomorphic, we conclude $C_1$ and $C_2$ are nonisomorphic concatenations. □

Combining the two theorems, we have

COROLLARY 5. *Two connected graphs may be recovered from the set of their nonisomorphic concatenations except when there is a unique concatenation C, and the limbs at the unique cutvertex in C are vertex transitive.*

**4. Concluding remarks.** When the graphs are not connected, however, reconstruction is not so easy. For example, suppose the components of $G_1$ are $A$, $A$, $B$ and the components of $G_2$ are $A$, $C$. The set of concatenations is the same as that obtained when $G_1$ has components $A$, $B$, and $G_2$ has components $A$, $A$, $C$. However, the authors believe that if the multiset of concatenations is given (i.e., we know the multiplicity of each concatenation), then the graphs may be uniquely reconstructed in all cases.

**Acknowledgment.** The authors would like to thank Jim Walker for helpful comments relating to the proof of these theorems.

REFERENCES

[1] J. A. BONDY, *On Ulam's conjecture for separable graphs*, Pacific J. Math., 31 (1969), pp. 281–288.
[2] W. B. GILES, *The reconstruction of outer planar graphs*, J. Combinatorial Theory Ser. B, 16 (1974), pp. 215–226.
[3] F. HARARY, *On the reconstruction of a graph from a collection of subgraphs*, Theory of Graphs and its Applications, M. Fielder, ed., Prague, 1964, pp. 47–52.
[4] F. HARARY AND E. PALMER, *The reconstruction of a tree from its maximum subtrees*, Canad. J. Math., 18 (1966), pp. 803–811.
[5] P. J. KELLEY, *A congruence theorem for trees*, Pacific J. Math., 7 (1957), pp. 961–968.
[6] B. MANVEL, *Reconstruction of trees*, Canad. J. Math., 22 (1970), pp. 55–60.
[7] ———, *Reconstruction of unicyclic graph*, Proof Techniques in Graph Theory, F. Harary, ed., Academic Press, New York, 1969.
[8] S. KUNDU, E. SAMPATHKUMAR AND V. H. BHAVE, *Reconstruction of a tree from its homomorphic images and other related transforms*, J. Combinatorial Theory Ser. B, 20 (1976), pp. 117–123.
[9] S. KUNDU, *Reconstruction of a unicyclic graph from its elementary contractions*, to appear.
[10] E. SAMPATHKUMAR AND V. H. BHAVE, *Reconstruction of a graph from its elementary partition graphs*, to appear.
[11] F. HARARY, *Theory of Graphs*, Addison-Wesley, Reading, MA, 1969.
[12] S. M. ULAM, *A Collection of Mathematical Problems*, Wiley Interscience, New York, 1960.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 2, June 1980


0196-5212/80/0102-0014 $01.00/0

# CONTROLLING OVERLOAD IN A DIGITAL SYSTEM*

E. ARTHURS† AND B. W. STUCK†

**Abstract.** Service requests arrive to an input buffer in a digital system. The buffer capacity and work discipline are to be chosen such that the maximum rate at which tasks are processed through the buffer while still meeting delay criteria is not above some maximum level. This input buffer is intended to be a control valve to prevent internal system overload. A processor visits the buffer at random time epochs, serves up to a given maximum number of tasks, leaves, and repeats this process. The equilibrium probability of buffer overflow, and the equilibrium waiting time distribution of a task until it is processed are studied as a function of model parameters.

**1. Introduction.** The mathematical model studied here is as follows: tasks arrive at a finite capacity buffer according to a Markov counting process conditioned on the number of tasks in the buffer. When a task arrives and finds the buffer full, one task (either the arrival or one in the buffer) is rejected or lost and buffer overflow is said to have occurred. A processor visits the buffer at random time intervals, processes up to a given maximum number of tasks according to a fixed work scheduling algorithm (called the buffer discipline), and then leaves. Each task occupies the processor for a time interval called the service time of the task. The service times are unknown, assumed for simplicity to be independent identically distributed random variables. Our goal is to study, as a function of model parameters, the probability of buffer overflow and the waiting time distribution of a task until its completion. Such a system has been called a loss-delay system [9].

The primary motivation for this study came from attempting to determine performance limitations in the so-called overload control mechanisms currently widely used in various electronic telephone switching systems. In such systems, our buffer would be the input buffer for the entire call processing system; by regulating how much work is let into the system through this buffer for further processing, time fluctuations in the stream of arriving calls can be smoothed; if there were no regulation, system performance would be significantly and adversely affected.

In addition, this study was motivated by the following secondary considerations: (i) Digital systems in which several buffers share a single processor are quite common in applications outside of telephone call switching [3], [8], [12], (ii) the cost of information storage is currently a significant economic factor in the design specification of a digital system, and it is therefore important to analyze the effects of a finite storage system on performance.

Finally, our problem is a highly simplified description of a computer disk information retrieval and storage system, where the buffer is an annulus on one level of the disk, the processor is the read/write head, and the processor intervisit time interval lasts from when the head leaves a given annulus until it next returns [12].

Although in most systems the intervisit distribution to a buffer is determined endogenously as the result of a multiple buffer scheduling rule [7], we feel it is important to first understand the performance limitations of a one (finite) buffer system with exogenous specified intervisit distribution, and will present generalizations of this work (to a set of buffers) elsewhere.

The next section presents a summary of the main results of this study. The third section states the mathematical problem and fixes notation. Section 4 presents a

* Received by the editors May 4, 1978, and in revised form September 27, 1979.
† Bell Laboratories, Murray Hill, New Jersey 07974.

detailed analysis of our model for one particular buffer discipline. Sections 5, 6 and 7 present brief sketches of how to modify the § 4 analysis for other buffer disciplines.

**2. Summary of results.** The maximum throughput rate of any digital system, denoted $\lambda_C$, is defined to be the maximum task completion rate for which all delay requirements can be met. When the mean rate of work arriving exceeds $\lambda_C$, an *external* overload is said to occur. If there is no means for limiting the number of tasks accepted, then an external overload can result in an overcommitment of the system resources (e.g., processing, storage, and so forth) with the subsequent loss of processing efficiency and violation of one or more delay requirements; this phenomenon is called *internal* overload.

Requests for service appear at the input queue as the first step in processing. Internal overload is prevented by limiting the maximum number of requests processed per processor visit to a suitable fixed value, $S$. For normal traffic loads $S$ will appear large in the sense that most attempts will be processed during the next processor visit after their arrival. When external overload occurs $S$ will appear small in that many attempts will wait more than one processor visit after their arrival to be processed. In actually implemented overload controls, $S$ may be adjusted depending on an estimate of the mean intervisit interval, for example, with the longer the intervisit time the smaller the value of $S$. Here we choose to fix $S$ because we feel it illustrates the essential features of controlling overload without undue analytic complications. Our analysis can in principle be extended to allow $S$ to depend on such complications.

It is also felt necessary to limit the size of the input queue to a suitable value $Q$ in order to maintain satisfactory waiting time performance in the presence of external overload. For normal traffic loads $Q$ will appear big in the sense that the probability an arriving task finds the input queue full is small (e.g., $<10^{-4}$), while when there is an external overload $Q$ will appear small in that there is a significant chance that an arriving task will be rejected from a full input queue.

Four techniques of operating the input queue were analyzed:

1. Nongated latest arrival cleared (NGLAC),
2. Nongated earliest arrival cleared (NGEAC),
3. Gated latest arrival cleared (GLAC),
4. Gated earliest arrival cleared (GEAC).

A precise definition of each of these buffer disciplines is given in § 4. Roughly speaking, a gated discipline involves shutting a gate on arrivals at the instant the processor arrives to work on the tasks in the input queue, and doing up to a maximum of $S$ jobs until the gate is reached, while a nongated discipline has no gate. All call requests are processed in order of arrival within the input queue, first come, first served.

The design problem is to choose $S$, $Q$ and the intervisit time probability distribution (i.e., specify the speed and work schedule of the central processor), such that

1. the maximum rate of accepting and processing calls through the input queue, $\lambda_A$, does not exceed the maximum rate at which the central processor services work, which is called the system mean capacity, denoted $\lambda_C$;

2. the fraction of incoming tasks which are rejected is sufficiently small when the mean arrival rate of work requests does not exceed the mean capacity;

3. the probability distribution of flow time (the time from when a work request arrives until the processor finishes processing it) satisfies the service requirements.

We implicitly assume from this point on that internal overload can be avoided if $\lambda_A$ does not exceed $\lambda_C$.

The principal findings were that

(i) visiting the input queue at constantly spaced time intervals vs. exponentially distributed time intervals can *significantly* reduce loss;

(ii) visiting the input queue more frequently, processing fewer work requests each visit vs. emptying the queue each visit, can *significantly* reduce loss.

This may imply that smaller storage capacity may be adequate if the intervisit intervals are deterministic and $S$ is small, vs. the larger storage capacity required for irregular intervisit intervals and emptying the buffer of all work.

The major theoretical contribution of this work is to extend the existing analyses of finite queues to include:

(a) the server leaving the queue after completing up to a given maximum number of tasks;

(b) a general class of Markov arrival processes with both state-dependence and group arrival.

A systematic method is given for determining the probability of buffer overflow and the Laplace-Stieltjes transform of the distribution of waiting time for such systems in equilibrium. Our approach is a natural generalization of the method of Riordan [9] for the M/G/1 finite queue.

**Closing comments.** Many qualifications are needed to be able to apply this work to an actual application in an intelligent manner. Here we touch on only two. First, we have chosen to ignore retry phenomena, such as studied by Cohen [14], where work attempts that time out will try at a later point in time to reenter the queue. Second, we have chosen to assume the intervisit time intervals were independent identically distributed random variables, when in fact the amount of work processed on a given visit to the input queue in an actual system may significantly correlate with the length of successive intervisit intervals (processing many work attempts will lengthen it, while processing few work attempts will shorten it). We do not believe that it is impossible to incorporate either of these phenomena into the analysis presented here, but it would have obscured the essential elements that were brought forth here. We leave these extensions for future work.

**3. Problem statement.** Figure 1 shows a block diagram of the queuing model. The model treated here is as follows:

**3.1. Buffer discipline.** All tasks are served in order of arrival, first come, first served, in each of the four buffer disciplines now described. Four buffer disciplines will be treated:

**3.1.1. Nongated latest arrival cleared (NGLAC).**

(a) The maximum number of jobs processed each processor visit is limited to $S$. If the buffer is emptied before $S$ requests are processed, the processor leaves.

(b) If a job arrives and finds the buffer full ($Q$ jobs already present), it is cleared from the system.

**3.1.2. Gated latest arrival cleared (GLAC).**

(a) Those jobs processed during a processor visit are limited to those already in the buffer at the time epoch that the processor arrives. The maximum number of tasks processed each processor visit is $S$.

(b) Same as (b) in NGLAC.

**3.1.3. Nongated earliest arrival cleared (NGEAC).**

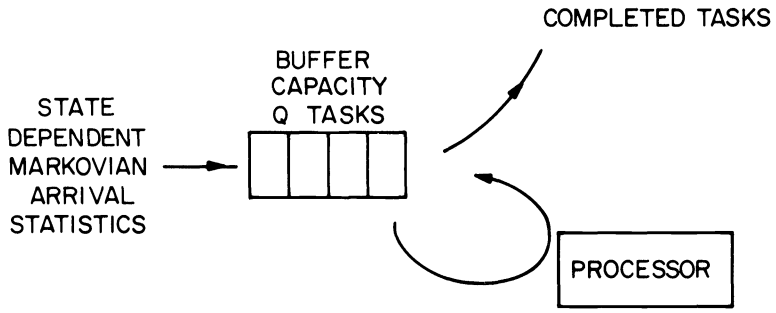
(a) Same as (a) in NGLAC.

(b) If a job arrives and finds the buffer ($Q$ jobs already present), it is inserted at the end of the queue and the task with the earliest arrival epoch is cleared from the system.

**3.1.4. Gated earliest arrival cleared (GEAC).**

(a) Same as (a) for GLAC.

(b) Same as (b) for NGEAC.




$\{V_i\}$ – INTERVISIT TIME INTERVALS

$\{P_i\}$ – PROCESSING TIME INTERVALS

S – MAXIMUM NUMBER OF TASKS EXECUTED EACH PROCESSOR VISIT

OVERLOAD CONTROL BLOCK DIAGRAM

FIG. 1

**3.2. Processor description.** The processing times $\{P_i\}$ for the jobs ($i = 1, 2, \cdots$) are assumed to be independent identically distributed (i.i.d.) random variables drawn from a common probability distribution $G_P(t)$, $G_P(0) = 0$, with associated Laplace-Stieltjes transform denoted $\gamma_P(s)$. The mean processing time $E(P)$ is assumed finite and nonnegative.

The intervisit intervals $\{V_i\}$ are the time intervals from when the processor leaves the buffer until it next returns. From this point on, the intervisit intervals are assumed to be i.i.d. random variables drawn from a common probability distribution $G_V(t)$, $G_V(0) = 0$, with associated Laplace-Stieltjes transform denoted $\gamma_V(s)$. We assume the mean intervisit interval, $E(V)$, is finite and strictly positive.

For ease of exposition we further assume that the distributions $G_P(t)$ and $G_V(t)$ are not jointly arithmetic on the same span. That is, there does not exist a positive real number $\delta$ such that $G_P(t)$ and $G_V(t)$ have their common support concentrated on the set of points $(0, \delta, 2\delta, 3\delta, \cdots)$.

**3.3. Arrival processes.** The arrival process is a Markov counting process conditioned on the number of tasks in the buffer. More specifically, given that the number of

tasks in the buffer at an arrival epoch time instant $t$ is equal to $m$, and that no completion of service occurs in the interval $(t, t+\delta]$

(i) the probability of an arrival epoch in $(t, t+\delta]$ is given by

$$C(m)\delta + o(\delta);$$

(ii) the probability of a group size equal to $K$, given that there was an arrival epoch in $(t, t+\delta]$ is equal to

$$R_B(K|m) \quad \text{for } K = 1, 2, \cdots;$$

(iii) the probability of more than one arrival epoch in $(t, t+\delta]$ is $o(\delta)$.

(iv) To avoid discussing physically uninteresting situations, we further assume that $C(m)$ is bounded and nonincreasing with $C(Q-1)>0$, and the distribution of the group size $R_B(K|m)$ has a finite mean for all $m$.

We now define some functionals of the arrival process which will be used later on. Let $\tilde{m}$ denote the event that the buffer contains $m$ tasks at time $T$ and no service completions occur in $(T, T+t]$. Then we can write

$$R_A(K, t|\tilde{m}) \equiv \Pr[K \text{ arrivals in } (T, T+t]|\tilde{m}].$$

The forward Kolmogoroff equations for $R_A(K, t|\tilde{m})$ can be set up in the usual manner from basic postulates (i)–(iv), taking note that since no completions occur in $(T, T+t]$, the buffer contains the minimum of $Q$ and $(m+K)$ at any time $t$, where $K$ is the number of arrivals in $(T, T+t]$. We then have

$$\frac{d}{dt}R_A(K, t|\tilde{m}) = -C(\min[m+K, Q])R_A(K, t|\tilde{m}) + \sum_{j=0}^{K-1} C(\min[m+j, Q])R_B(K-j|\min[m+j, Q])R_A(j, t|\tilde{m})$$

for $m = 0, 1, \cdots, Q$ and $K = 0, 1, 2, \cdots$, with initial conditions

$$R_A(K, 0) = \begin{cases} 1, & K = 0, \\ 0, & K > 0. \end{cases}$$

The Kolmogoroff equations can be solved explicitly in two special cases. In the special case of the finite source of size $Q$, $R_A(K, t|\tilde{m})$ is given by

$$R_A(K, t|\tilde{m}) = \begin{cases} \binom{Q-m}{K}(1-e^{-\lambda t})^K e^{-\lambda t(Q-m-K)}, & 0 \leqq K \leqq Q-m, \\ 0, & K > Q-m. \end{cases}$$

When the arrival process is a compound Poisson counting process, it can be shown that

$$R_A(K, t|\tilde{m}) = R_A(K, t) \quad \text{independent of } m$$

and

$$\rho_A(x, t) = \sum_{K=0}^{\infty} R_A(K, t)x^K = \exp[ct(\rho_B(x)-1)]$$

where $1/c$ is the mean time interval between arrival epochs, and $\rho_B(x)$ is the moment generating function of the arrival batch size compounding distribution.

Let $A_m$ denote the random variable for the number of arrivals during an intervisit interval, given that there are $m$ tasks in the buffer at the beginning of the interval. The probability distribution of $A_m$, denoted $R_V(K|\tilde{m})$, is given by

$$R_V(K|\tilde{m}) = \int_0^\infty R_A(K, t|\tilde{m})\, dG_V(t).$$

Let $B_m$ denote the random variable for the number of tasks arriving during a service interval given that there are $m$ tasks in the buffer at the beginning of the interval. The probability distribution of $B_m$, denoted $R_P(K|\tilde{m})$, is then given

$$R_P(K|\tilde{m}) = \int_0^\infty R_A(k, t|\tilde{m})\, dG_P(t).$$

**4. Analysis of the NGLAC discipline.** Three cases are to be distinguished in the analysis of the NGLAC buffer discipline:

(A) $E(P) > 0$ and $S$ finite,

(B) $E(P) = 0$,

(C) $E(P) > 0$ and $S$ infinite.

From a mathematical point of view the latter two cases are degenerate versions of (A) where the number of processing phases collapse to zero and one respectively. (The case $E(P) = 0$ is referred to as a bulk service in the literature, e.g., [9].)

**4.1. Case (A).** We define the Markov process $\nu(t)$ which takes values in the system state space $\tilde{\Sigma}$. Elements of the set $\tilde{\Sigma}$ are triples $(m, n, x)$ where $m$ is the number of tasks in the buffer ($m = 0, \cdots, Q$), $n$ is the processor phase (which will be defined shortly; $n = 0, \cdots, S$), and $x$ is the elapsed time of the phase, $x \in (0, \infty]$. The phase $n = 0$ corresponds to the processor not processing work in the buffer, while $n > 0$ corresponds to the processor doing work at the buffer and there have been $(S - n)$ completions during the current processor visit (we choose to number the phases this way solely for convenience in the analysis to follow; naturally, other numbering schemes may be used). More precisely, we choose to define the set $\tilde{\Sigma} = \Omega \times (0, \infty]$, where

$$\Omega = \begin{cases} \{(m, 0) \text{ for } n = 0 \text{ and } m = 0, \cdots, Q\}, \\ \{(m, n) \text{ for } n = 1, \cdots, S \text{ and } m = 1, \cdots, Q\}. \end{cases}$$

For later use, we note that at the beginning of a phase, the possible state values belong to $\tilde{\Omega} \times \{x = 0+\}$, with

$$\tilde{\Omega} = \bigcup_{n=0}^{S} \Omega_n,$$

$$\Omega_0 = \{(m, n = 0): m = 0, \cdots, Q-1\}, \quad \Omega_n = \{(m, n): m = 1, \cdots, Q-1\}|_{n=1,\cdots,S-1},$$

$$\Omega_S = \{(m, n = S): m = 1, \cdots, Q\}.$$

Figure 2 shows one possible arrival pattern and the resulting state changes for $(Q = 5, S = 2)$.

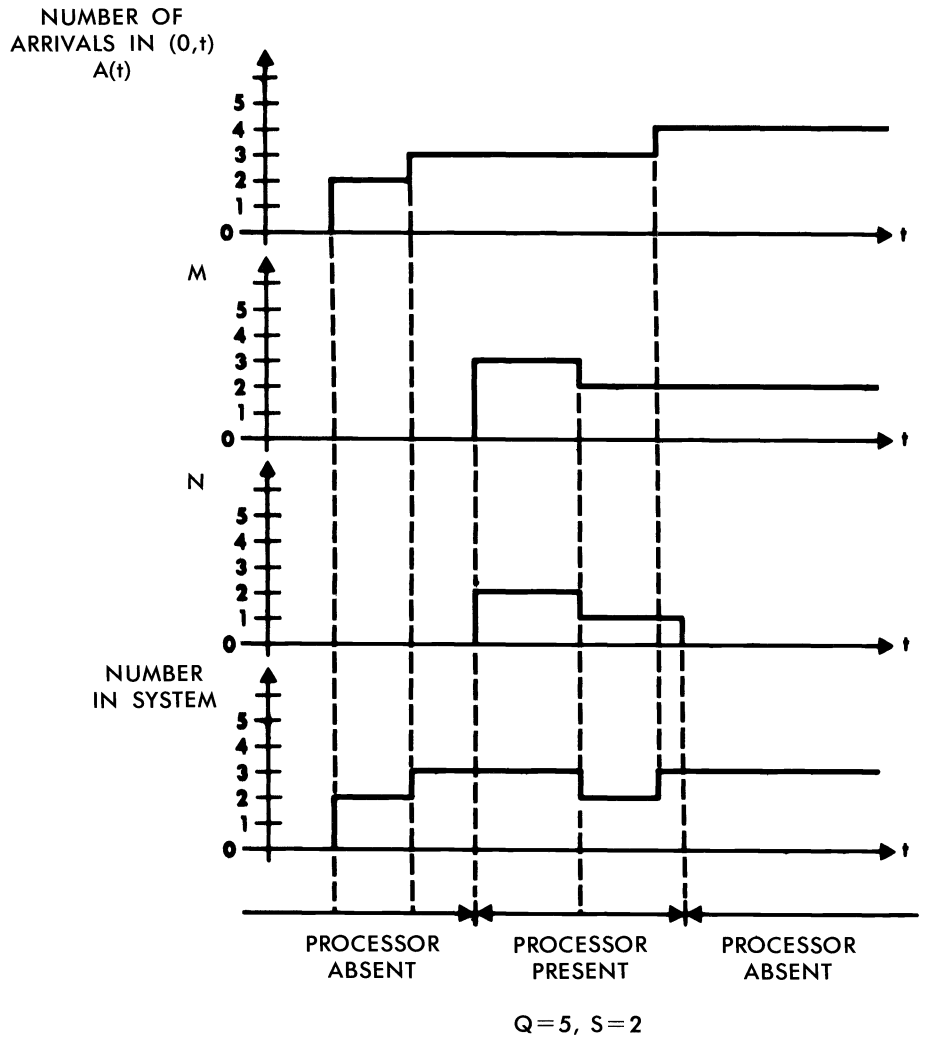



FIG. 2

Next, we define $P[\nu(t)]\,dx = P(m, n, x, t)\,dx$ as the probability that $\nu(t)$ belongs to the set $(m, n, y)$: $(m, n) \in \Omega$, $x \leqq y < x + dx$, at time $t$. For convenience, we assume a state space transition occurs at $t = 0$. Rather than analyze the probabilistic behavior of $\nu(t)$, we choose to approach the analysis via an imbedded semi-Markov process $\nu_I(t)$.[1] $\nu_I(t)$ also takes values in $\tilde{\Omega} \times (0, \infty]$. Let $\{t_i\}$ be the set of time epochs, $t_0 = 0$, $T_{i+1} > t_i$, of phase transitions in the $\nu(T)$ process. If $\nu(t_i) \equiv (m_i, n_i, 0)$, then we define

$$\nu_I(t) \equiv (m_i, n_i, x = t - t_i) \quad \text{for } t_i < t \leqq t_{i+1}.$$

We define $\pi[\nu_I(t)]\,dx = \pi(m, n, x, t)\,dx$ as the probability that $\nu_I(t)$ belongs to the set $(m, n, x)$: $(m, n) \in \tilde{\Omega}$, $x \leqq y < x + dx$ at time $t$. From this discussion, it follows that for

[1] The use of semi-Markov processes in the analysis of queuing systems was introduced by A. Fabens [6]; one of the first systematic treatments of queuing theory problems using semi-Markov chains was due to D. G. Kendall [13].

$x>0$, $(m, n)\in\Omega$,

$$(1)\qquad P(m,n,x,t)=\begin{cases}\sum_{j=0}^{m}\pi(j,0,x,t)R_A(m-j,x|j) & \text{for } 0\leqq m<Q,\ n=0,\\ \sum_{j=0}^{Q-1}\pi(j,0,x,t)\sum_{k=Q-j}^{\infty}R_A(k,x|j) & \text{for } m=Q,\ n=0,\\ \sum_{j=1}^{m}\pi(j,n,x,t)R_A(m-j,x|j) & \text{for } 0<m<Q,\ 0<n<S,\\ \sum_{j=1}^{Q-1}\pi(j,n,x,t)\sum_{k=Q-j}^{\infty}R_A(k,x|j) & \text{for } m=Q,\ 0<n<S,\\ \sum_{j=1}^{m}\pi(j,S,x,t)R_A(m-j,x|j) & \text{for } 0<m<Q,\ n=S,\\ \sum_{j=1}^{Q}\pi(j,S,x,t)\sum_{k=Q-j}^{\infty}R_A(k,x|j) & \text{for } m=Q,\ n=S,\end{cases}$$

where $R_A(k, x|j)=\Pr[A(x)=k|j]$. For simplicity of exposition, we deal with the steady state case from this point on. Our goals are to determine the probability of buffer overflow and the probability distribution for the waiting time of an accepted task, i.e., the time interval from when a task arrives and is accepted into the buffer until the processor starts processing the task.

PROPOSITION 4.1. *Given the preceding assumptions, there exists a unique equilibrium probability distribution denoted* $\pi_e$,

$$\lim_{t\to\infty}\pi(m,n,x,t)=\pi_e(m,n,x).$$

$\pi_e$ *is independent of* $\pi(m, n, x, t=0)$ *and is given by*

$$(2)\qquad \pi_e(m,n,x)=\begin{cases}\hat{\pi}_e(m,0)[1-G_V(x)]/E(V), & n=0,\\ \hat{\pi}_e(m,n)[1-G_P(x)]/E(P), & n>0,\end{cases}$$

*for* $(m, n)\in\tilde{\Omega}$. $\hat{\pi}_e$ *is given by*

$$(3)\qquad \hat{\pi}_e(m,n)=\begin{cases}\tilde{\pi}(m,0)E(V)/N, & n=0,\\ \tilde{\pi}(m,n)E(P)/N, & n>0,\end{cases}$$

$$(4)\qquad N=\sum_{(m,0)\in\tilde{\Omega}}\tilde{\pi}(m,0)E(V)+\sum_{(m,n>0)\in\tilde{\Omega}}\tilde{\pi}(m,n)E(P)$$

*for* $(m, n)\in\tilde{\Omega}$, *where* $\tilde{\pi}(m, n)$ *is the invariant measure associated with the generator* $g(m', n'; m, n)$ *of the semi-Markov process* $v_I(t)$,

$$g(m', n'; m, n)=\Pr[\nu_I(t_{i+1})=(m', n', x=0)|\nu_I(t_i)=(m, n, x=0)].$$

The proof of this theorem is broken down into two parts. First we show that *g is the generator of an irreducible aperiodic Markov chain.*

From earlier assumptions about the arrival process, $R_V(k)$ and $R_P(k)$ are positive for all negative $k$. It is clear that all states $(m, n)$ lead to $(0, 0)$ in a finite number of transitions; simply assume no arrivals during the successive phases until $(0, 0)$ is reached. The states $(m, S)$ can be reached from $(0, 0)$ in one transition. The states

$(m, n)$, where $(m, n) \in \tilde{\Omega}$ and $0 < n < S$ can be reached from $(0, 0)$ by the path $(0, 0)$, $(m, S)$, $(m, S-1), \cdots, (m, n)$. The states $(m, 0)$, $1 \leqq m \leqq Q-1$ can be reached from $(0, 0)$ by the path $(0, 0), (1, S), (1, S-1), \cdots, (1, 1), (m, 0)$. Finally, $(0, 0)$ can be reached from $(0, 0)$ in one step. Hence the chain is aperiodic and irreducible.

We now turn to the second part of the proof of the theorem: To specify the behavior of $\nu_I(t)$ between the transition epochs $\{t_i\}$, define the sojourn time distribution $H(t; m, n)$,

$$H(t; m, n) = \Pr[t_{i+1} - t_i < t | \nu_I(t_i) = (m, n, x = 0)]$$
$$= \begin{cases} G_V(t), & n = 0, \\ G_P(t), & n > 0. \end{cases}$$

The generator $g$ is given explicitly, for $(m, n)$ and $(m', n') \in \tilde{\Omega}$, by

$$g(0, 0; 0, 0) = R_V(0|0);$$

$$g(m', S; m, 0) = R_V(m' - m|m), \qquad m = 0, \cdots, Q-1; \quad 0 < m' < Q; \quad m \leqq m';$$

$$g(Q, S; m, 0) = \sum_{k=Q-m}^{\infty} R_V(k|m), \qquad m = 0, \cdots, Q-1;$$

$$g(m', S-1; m, S) = R_P(m' - m + 1|m), \qquad m = 1, \cdots, Q-1; \quad 0 < m' < Q-1;$$
$$m - 1 < m';$$

$$g(Q-1, S-1; m, S) = \sum_{k=Q-m}^{\infty} R_P(k|m), \qquad m = 1, \cdots, Q;$$

$$g(0, 0; 1, n) = R_P(0|1), \qquad n = 1, \cdots, S;$$

$$g(Q-1, n-1; m, n) = \sum_{k=Q-m}^{\infty} R_P(k|m), \qquad m = 1, \cdots, Q-1; \quad 0 < n < S;$$

$g(m', n'; m, n) = 0$ otherwise.

Since the Markov chain is irreducible and aperiodic, and the state space is compact, all states are positive recurrent, so there exists a unique invariant measure $\tilde{\pi}(m, n)$ [5, p. 183], where

$$\tilde{\pi}(m, n) = \sum_{(m,n) \in \tilde{\Omega}} g(m, n; m', n') \tilde{\pi}(m', n'),$$

$$\tilde{\pi}(m, n) \geqq 0 \quad \text{and} \quad \sum_{(m,n) \in \tilde{\Omega}} \tilde{\pi}(m, n) = 1,$$

for $(m, n) \in \Omega$. Since $\nu_I(t)$ has been shown to be an irreducible aperiodic semi-Markov process with all states positive recurrent, the desired result now follows [10, p. 109, Th. 5.17]. Q.E.D.

PROPOSITION 4.2. *Given the preceding assumptions, there exists a unique equilibrium probability distribution, $P_e$,*

$$\lim_{t \to \infty} P(m, n, x, t) = P_e(m, n, x).$$

*$P_e$ is independent of $P(m, n, x, t=0)$ and is given by*

$$
\begin{cases}
\sum_{k=0}^{m} \hat{\pi}_e(m-k, 0)[1-G_V(x)]R_A(k, x|m-k)/E(V) & \text{for } 0 \leqq m < Q, \quad n = 0, \\
\sum_{k=0}^{m-1} \hat{\pi}_e(m-k, n)[1-G_P(x)]R_A(k, x|m-k)/E(P) & \text{for } 1 \leqq m < Q, \quad 0 < n < S, \\
\sum_{k=1}^{Q} \hat{\pi}_e(Q-k, 0)[1-G_V(x)] \sum_{l=k}^{\infty} R_A(l, x|Q-k)/E(V) & \text{for } m = Q, \quad n = 0, \\
\sum_{k=1}^{Q-1} \hat{\pi}_e(Q-k, 0)[1-G_P(x)] \sum_{l=k}^{\infty} R_A(l, x|Q-k)/E(P) & \text{for } m = Q, \quad 0 < n < S, \\
\sum_{k=0}^{m-1} \hat{\pi}_e(m-k, S)[1-G_P(x)]R_A(k, x)/E(P) & \text{for } 1 \leqq m < Q, \quad n = S, \\
\sum_{k=0}^{Q-1} \hat{\pi}_e(Q-k, S)[1-G_P(x)] \sum_{l=k}^{\infty} R_A(l, x|Q-k)/E(P) & \text{for } m = Q, \quad n = S.
\end{cases} \tag{5}
$$

*Proof.* This result follows from the arguments given in the previous proposition plus the relationship between $P(m, n, x, t)$ and $\pi(m, n, x, t)$ given in (1). Q.E.D.

Next we wish to determine the probability of loss, $L$, the equilibrium probability of buffer overflow. If we define Pr [Accept] as the probability that an arrival will be accepted, we have Pr [Accept] equals $(1-L)$, and if we define $N_A(T)$ as the number of tasks accepted for processing in an interval of duration $T$, and $N(T)$ as the total number of tasks arriving in an interval of duration $T$, then [10]

$$
\begin{aligned}
\Pr[\text{Accept}] &= \lim_{T\to\infty} \frac{N_A(T)}{N(T)} \quad \text{a.s.} \\
&= \frac{\text{average number accepted per arrival epoch}}{\text{average number of arrivals per arrival epoch}}.
\end{aligned} \tag{6}
$$

Therefore, we can write Pr [Accept] as $\hat{N}_A/\hat{N}$, where

$$
\hat{N}_A = \int_0^{\infty} dx \sum_{(m,n)\in\tilde{\Omega}} P_e(m, n, x)C(\tilde{m}) \left[ \sum_{k=1}^{Q-m} kR_B(k|\tilde{m}) + \sum_{k=Q-m+1}^{\infty} (Q-m)R_B(k|\tilde{m}) \right] \tag{7}
$$

$$
\hat{N} = \int_0^{\infty} dx \sum_{(m,n)\in\Omega} P_e(m, n, x)C(\tilde{m}) \sum_{k=1}^{\infty} kR_B(k|\tilde{m}). \tag{8}
$$

If we denote the mean number of completions per unit time by $\lambda_A$, then we can write

$$
\lambda_A = \frac{1}{E(P)} \int_0^{\infty} dx \sum_{n=1}^{S} \sum_{m=1}^{Q} P_e(m, n, x). \tag{9}
$$

Let $\lambda$ denote the mean number of arrivals per unit time. In terms of the quantities just defined, it is straightforward to show that

$$
\lambda = \frac{\lambda_A}{(1-L)}. \tag{10}
$$

Our next goal is to determine an explicit expression for the Laplace-Stieltjes transform of the waiting time distribution. Because tasks are served in order of their

arrival, and because it was assumed the processing times of the tasks were independent, we can write

$$\gamma_F(z) = \gamma_W(z)\gamma_P(z) \tag{11}$$

where $\gamma_F(z)$ is the Laplace-Stieltjes transform of the equilibrium flow time distribution; hence, once we find $\gamma_W(z)$, we will know $\gamma_F(z)$. From elementary arguments it can be seen that

$$\gamma_W(z) = \int_0^\infty dx \Bigg[ \sum_{m=0}^{Q-1} P_a(m, 0, x)\gamma_W(z|m, 0, x) + \sum_{n=1}^{S} \sum_{m=1}^{Q-1} P_a(m, n, x)\gamma_W(z|m, n, x) \Bigg] \tag{12}$$

where $P_a(m, n, x)\,dx$ is the probability that an accepted task finds the system in $\{(m, n, y)\colon (m, n) \in \Omega,\ x \leqq y < x + dx\}$, and $\gamma_W(z|m, n, x)$ is the Laplace-Stieltjes transform of the equilibrium waiting time distribution of an accepted task, given the system state.

We can write $P_a(m, n, x) = \hat{P}(m, n, x)/\hat{N}_A$,

$$P_a(m, n, x) = \begin{cases} \hat{P}_0/\hat{N}, & 0 \leqq m < Q,\ n = 0, \\ \hat{P}_n/\hat{N}, & 0 < m < Q,\ n > 0. \end{cases}$$

$$\hat{P}_0 = \sum_{j=0}^{m} P_e(m-j, 0, x) \sum_{k=j+1}^{\infty} C(m-j)R_B(k|m-j) \quad \text{for } 0 \leqq m < Q, \quad n = 0, \tag{13}$$

$$\hat{P}_n = \sum_{j=0}^{m-1} P_e(m-j, n, x) \sum_{k=j+1}^{\infty} C(m-j)R_B(k|m-j) \quad \text{for } 0 < m < Q, \quad 0 < n \leqq S.$$

It is then straightforward to show that

$$\gamma_W(z|m, n, x) = \begin{cases} \gamma_P^m(z)\gamma_V^{\lceil (m+1)/S \rceil} I_0, & n = 0, \\ \gamma_P^{m-1}(z)\gamma_V^{\lceil (m+1-n)/S \rceil} I_n, & m+1 \geqq n > 0, \end{cases}$$

$$I_0 = \int_0^\infty \frac{d_y G_V(x+y)\, e^{-zy}}{[1 - G_V(x)]}, \qquad I_n = \int_0^\infty \frac{d_y G_P(x+y)\, e^{-zy}}{[1 - G_P(x)]}, \tag{14}$$

$\lceil y \rceil$ = smallest nonnegative integer greater than or equal $y$.

Thus, from (6)–(14) we can calculate the Laplace–Stieltjes transform of the equilibrium waiting time distribution, and find the moments of the distribution; numerical methods are currently being investigated for calculating the approximate distribution by inverting this transform, in order to examine the asymptotic behavior of the distribution.

**4.2. Case (B).** In this case, since $E(P) = 0$, there is only one phase, the intervisit phase $n = 0$. The generator must be modified in the obvious way to reflect the sudden completion of up to $S$ tasks at the end of the intervisit period. The determination of the waiting time distribution and loss are carried out in the same manner as for case (A) with the difference that there is only one phase and $\gamma_P(z)$ is replaced by unity in (14).

**4.3. Case (C).** In this case, $S = \infty$, the buffer is always emptied before the processor leaves the buffer. Thus there are only two phases, the intervisit phase, $n = 0$, and a single processing phase, since all processing phases are identical. The generator must be modified in the obvious manner to return to the processing phase as long as there are tasks in the buffer. The determination of the loss and waiting time distribution

is carried out in the same manner as for case (A), except that there are only two phases and the terms $\gamma^{[k/S]}(z)$ are replaced by unity.

**4.4. An illustrative example.** Now we illustrate our analysis with some machine calculations. In many problems of practical interest (e.g., [3]), the mean intervisit interval is much greater than the mean processing time per task; in these applications, it is often reasonable to employ the approximation $E(P)=0$. We do so here because it reduces the state space dimension from $Q(S+2)-S$ to $Q-S+1$ and hence reduces the computation cost involved in calculating $\pi$. Figures 3 and 4 (respectively 5 and 6) plot probability of loss (respectively mean waiting time) vs. mean number of arrivals per intervisit period for a simple Poisson process with $E(P)=0$, for a deterministic ($D$) and exponential ($E_1$) intervisit distribution, a holding $\lambda_C$, the maximum mean completion rate, constant, because in most applications $\lambda_C$ will be a given design parameter. The invariant measure $\pi$ was calculated using EISPACK [11, § 2.1.10]. Note that for a fixed mean number of arrivals per intervisit interval, (a) the loss can be significantly smaller for the deterministic vs. exponential intervisit distribution, and (b) visiting or polling the buffer more frequently (smaller $S$) reduces the probability of buffer overflow. An additional observation is that as the mean number of arrivals per intervisit interval becomes infinite, the mean waiting time $E(W)$ approachs $Q/\lambda_C$. Finally, note that as the mean number of arrivals per intervisit interval approaches zero, the mean waiting time approaches $SE(V_1^2)/2E(V_1)$, where $V_S$ is the random variable denoting the intervisit interval for a given $S$, $E(V_1)=1/\lambda_C$. As the mean arrival rate approaches infinity, the mean rate of accepting work is

$$\lim_{\lambda\to\infty}\lambda_A=\frac{S}{SE(P)+E(V)}\equiv\lambda_C, \tag{15}$$

since as the arrival rate becomes infinite, there will be $S$ tasks processed each visit. Roughly speaking, to first order $S$ controls the maximum rate of accepting work (i.e., $S$ is the overload control mechanism), while $Q$ controls buffer overflow probability, for fixed $E(P)$ and $E(V_S)$. Using the algorithms which lead to Figs. 3–6, the designer can choose a particular $(Q, S)$ pair and determine performance over a range of arrival rates.

**5. Analysis of the GLAC discipline.** There are only two distinguishable cases for the GLAC discipline:

(A) $E(P)<0$,

(B) $E(P)=0$.

The case $S=\infty$ does not arise because the largest effective value for $S$ is $Q$. Case (B) is identical for the GLAC and NGLAC disciplines, so it only remains to discuss case (A).

**5.1. Case (A).** The analysis in this section is quite brief, since it is identical with the overall analysis in the previous section. The only difference in the two cases is that the generator associated with $\nu_l(t)$ is different here from that in the previous section.

Throughout this section, we assume $Q\geqq S$ because the $S=Q$ case is identical with the $S>Q$ case. The sets $\Omega$ and $\tilde{\Omega}$ are

$$\begin{aligned}\Omega&=\bigcup_{n=0}^{S}\bigcup_{m=n}^{Q}(m,n),\\ \tilde{\Omega}&=\left[\bigcup_{n=0}^{S-1}\bigcup_{m=n}^{Q-1}(m,n)\right]\cup\left[\bigcup_{m=S}^{Q}(m,n=S)\right].\end{aligned} \tag{16}$$

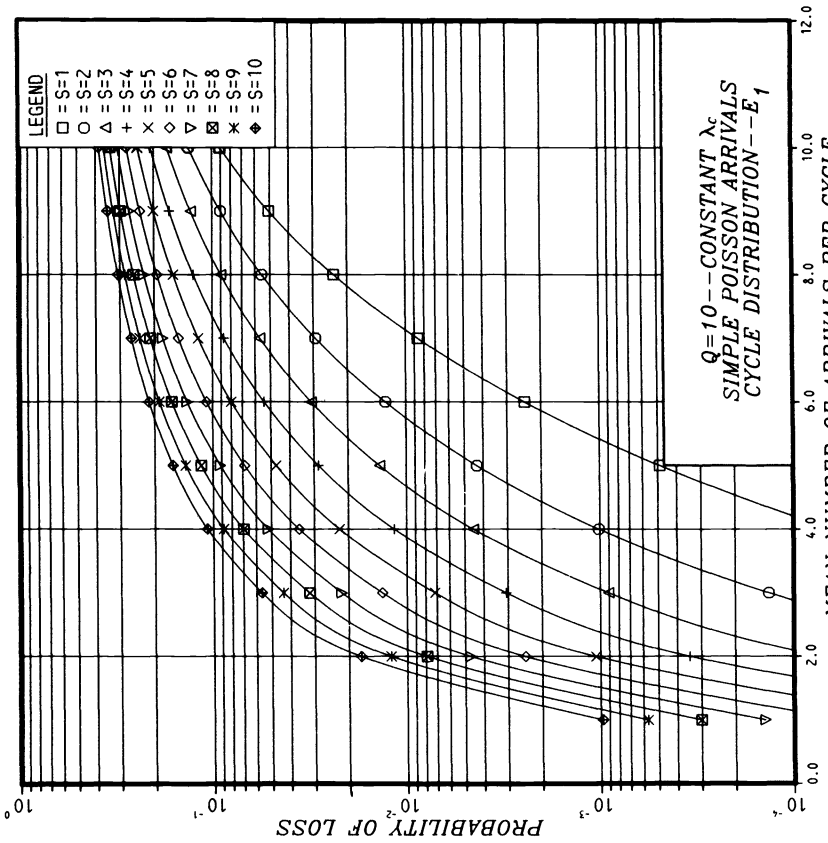

LEGEND
□ = S=1
○ = S=2
△ = S=3
+ = S=4
× = S=5
◇ = S=6
▽ = S=7
⊠ = S=8
✳ = S=9
◆ = S=10
Q=10--CONSTANT λc
SIMPLE POISSON ARRIVALS
CYCLE DISTRIBUTION--E1
PROBABILITY OF LOSS
MEAN NUMBER OF ARRIVALS PER CYCLE

FIG. 4

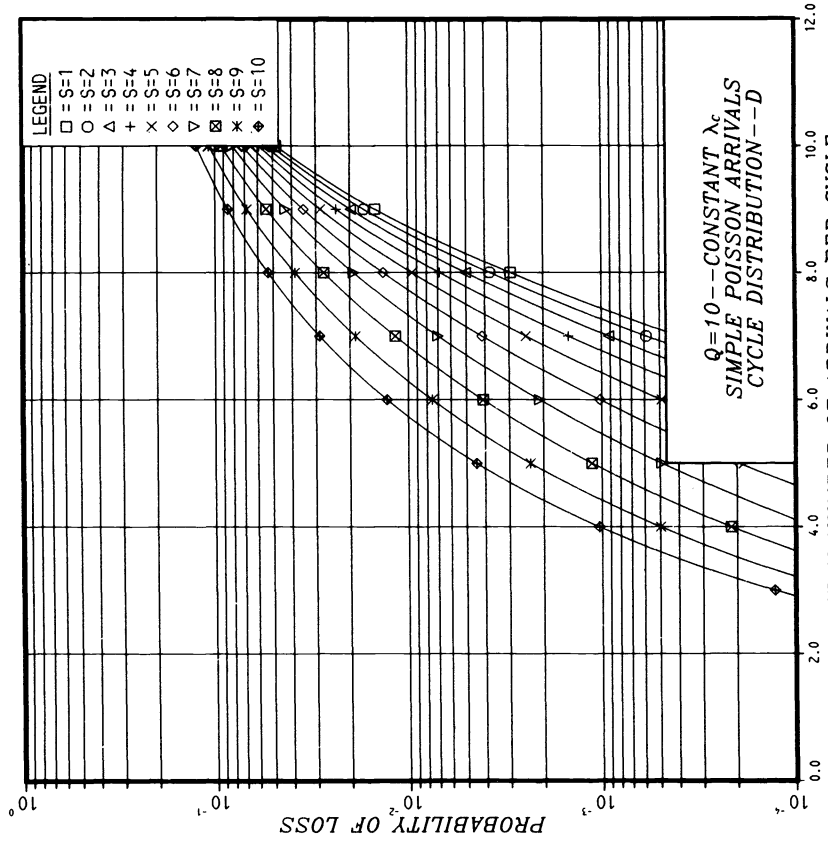

LEGEND
□ = S=1
○ = S=2
△ = S=3
+ = S=4
× = S=5
◇ = S=6
▽ = S=7
⊠ = S=8
✳ = S=9
◆ = S=10
Q=10--CONSTANT λc
SIMPLE POISSON ARRIVALS
CYCLE DISTRIBUTION--D
PROBABILITY OF LOSS
MEAN NUMBER OF ARRIVALS PER CYCLE

FIG. 3

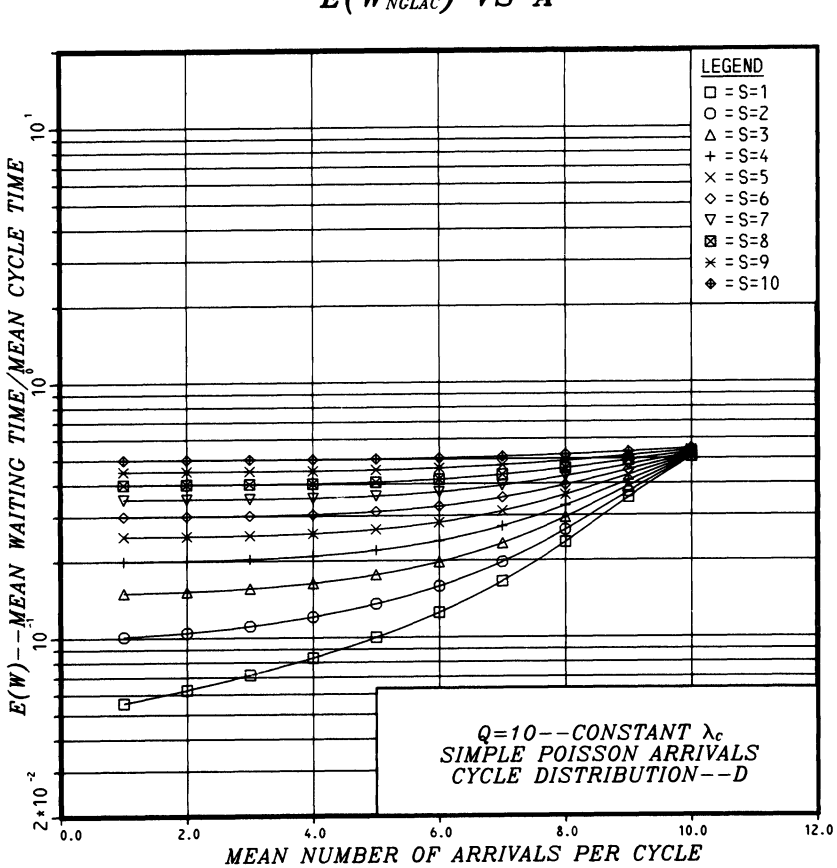

E(W_NGLAC) VS A
LEGEND
□ = S=1
○ = S=2
△ = S=3
+ = S=4
× = S=5
◇ = S=6
▽ = S=7
⊠ = S=8
✳ = S=9
⊕ = S=10
Q=10--CONSTANT λc
SIMPLE POISSON ARRIVALS
CYCLE DISTRIBUTION--D
E(W)--MEAN WAITING TIME/MEAN CYCLE TIME
MEAN NUMBER OF ARRIVALS PER CYCLE
0.0
2.0
4.0
6.0
8.0
10.0
12.0

FIG. 5

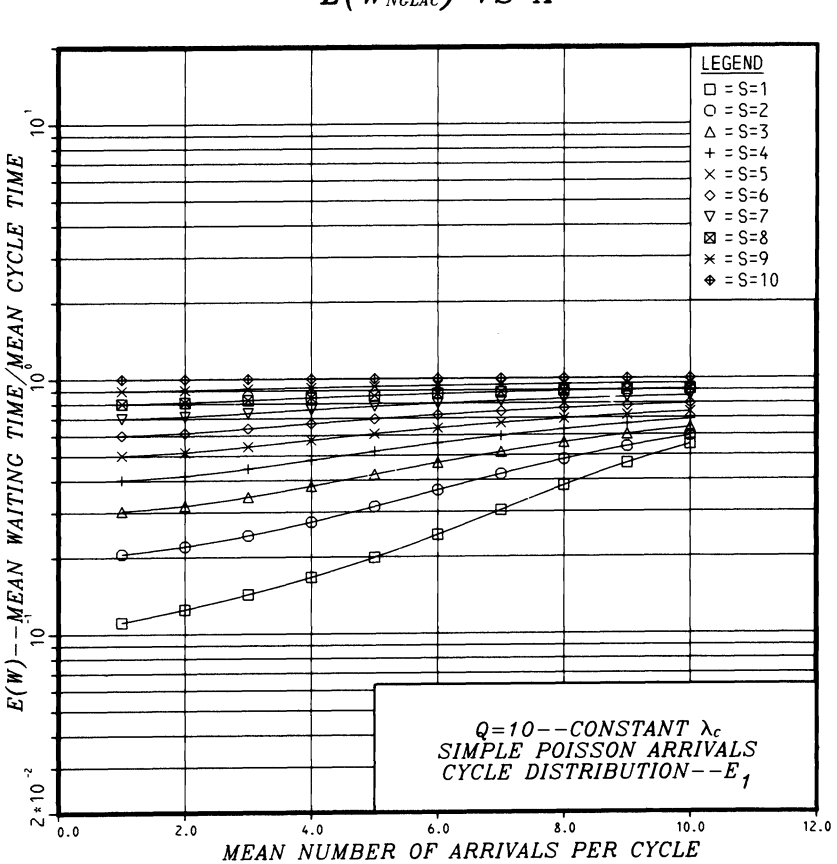

E(W_NGLAC) VS A
LEGEND
□ = S=1
○ = S=2
△ = S=3
+ = S=4
× = S=5
◇ = S=6
▽ = S=7
⊠ = S=8
✳ = S=9
⊕ = S=10
Q=10--CONSTANT λc
SIMPLE POISSON ARRIVALS
CYCLE DISTRIBUTION--E1
E(W)--MEAN WAITING TIME/MEAN CYCLE TIME
MEAN NUMBER OF ARRIVALS PER CYCLE
0.0
2.0
4.0
6.0
8.0
10.0
12.0

FIG. 6

The generator $g(m', n'; m, n)$ for $(m, n), (m', n') \in \tilde{\Omega}$, is

$$
\begin{aligned}
&g(0, 0; 0, 0) = R_V(0|0);\\
&g(m', \min(m', S); m, 0) = R_V(m' - m|m) \quad \text{for } 0 < m' < Q, \quad 0 \leqq m \leqq m';\\
&g(Q, S; m, 0) = \sum_{k=Q-m}^{\infty} R_V(k|m) \quad \text{for } 0 \leqq m < Q;\\
&g(m', S-1; m, S) = R_P(m' - m + 1|m) \quad \text{for } 0 < m' < Q-1, \quad S-1 < m \leqq m';\\
&g(Q-1, S-1; m, S) = \sum_{k=Q-m+1}^{\infty} R_P(k|m) \quad \text{for } S \leqq m;\\
&g(m', n-1; m, n) = R_P(m' - m + 1|m) \quad \text{for } 1 \leqq n \leqq S-1, \quad 0 < m' < Q-1;\\
&g(Q-1, n-1; m, n) = \sum_{k=Q-m+1}^{\infty} R_P(k|m)\\
&g(m', n'; m, n) = 0 \quad \text{otherwise.}
\end{aligned} \tag{17}
$$

The analysis is identical with that for the NGLAC case, with the proviso that the invariant measures associated with the two generators will in general be different (for $S = 1$ the two are identical).

**6. The NGEAC case.** The probability of buffer overflow is the same for the NGEAC and NGLAC cases, but the waiting time distribution in general is different. The main reason for studying the NGEAC discipline is that the waiting time of an accepted task can be significantly smaller at high arrival rates for NGEAC vs. NGLAC discipline.

**6.1. Summary of results for the NGEAC case.** Before proceeding with the analysis, we wish to gain insight into the differences in the mean waiting time for an accepted task in the NGEAC vs. NGLAC buffer discipline; we do so by appealing to Little's theorem [15].

We define $D$ to be the random variable denoting the duration of time that a given cleared task spends in the buffer. Clearly, $D_{\text{NGLAC}} = 0$ while $D_{\text{NGEAC}} \geqq 0$. From Little's theorem, since the probability of buffer overflow is the same for the two buffer disciplines, we see that

$$LE(D_{\text{NGEAC}}) + (1-L)E(W_{\text{NGEAC}}) = (1-L)E(W_{\text{NGLAC}}), \tag{18}$$

and rearranging, we find that

$$E(W_{\text{NGEAC}}) = E(W_{\text{NGLAC}}) - \frac{L}{1-L}E(D_{\text{NGEAC}}), \tag{19}$$

so $E(W_{\text{NGEAC}}) < E(W_{\text{NGLAC}})$. Elementary arguments show that $E(D_{\text{NGEAC}}) \sim O(Q/\lambda)$, i.e., $\lim_{\lambda\to\infty} E(D_{\text{NGEAC}}) = 0$. To calculate $E(D_{\text{NGEAC}})$ is as difficult as calculating $E/(W_{\text{NGEAC}})$, apparently. Here we content ourselves with approximating $E(D_{\text{NGEAC}})$ by $Q/\lambda$ for all $\lambda$. Figures 7 and 8 plot this approximation versus the mean number of arrivals per intervisit interval with $E(P) = 0$ as in Figs. 3, 4, with $\lambda_C$ fixed. Note that the approximation is much less than $E(W_{\text{NGEAC}})$ for $\lambda \gg 1$, and hence the actual mean waiting time for the NGEAC buffer discipline may be significantly smaller than for the NGLAC buffer discipline with all other factors the same. At low arrival rates ($L \ll 1$), it is plausible that the NGEAC and NGLAC buffer disciplines should have approximately the same mean waiting time.

**6.2. Analysis of the NGEAC case.** We choose to analyze only the case $E(P) > 0$ and $S$ finite, and leave the other two cases as exercises.

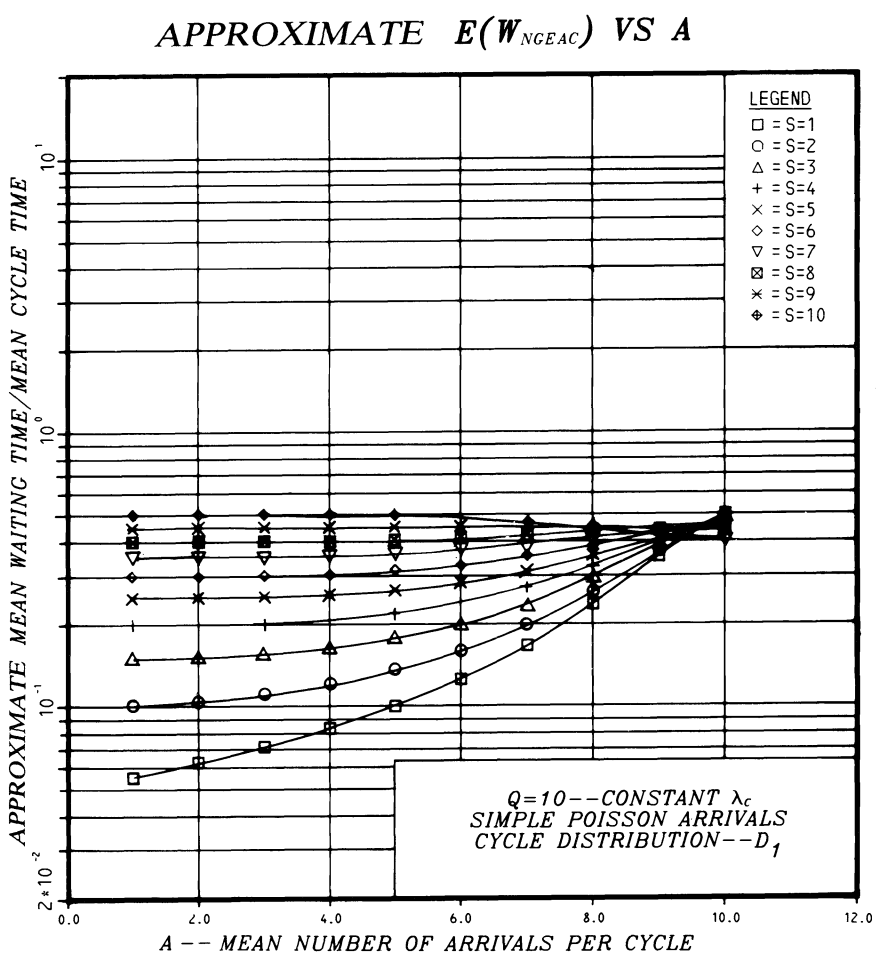

APPROXIMATE $E(W_{NGEAC})$ VS A
LEGEND
□ = S=1
○ = S=2
△ = S=3
+ = S=4
× = S=5
◇ = S=6
▽ = S=7
⊠ = S=8
✳ = S=9
⊕ = S=10
Q=10--CONSTANT $\lambda_c$
SIMPLE POISSON ARRIVALS
CYCLE DISTRIBUTION--$D_1$
APPROXIMATE MEAN WAITING TIME/MEAN CYCLE TIME
$10^1$
$10^0$
$10^{-1}$
$2*10^{-2}$
0.0
2.0
4.0
6.0
8.0
10.0
12.0
A -- MEAN NUMBER OF ARRIVALS PER CYCLE

FIG. 7

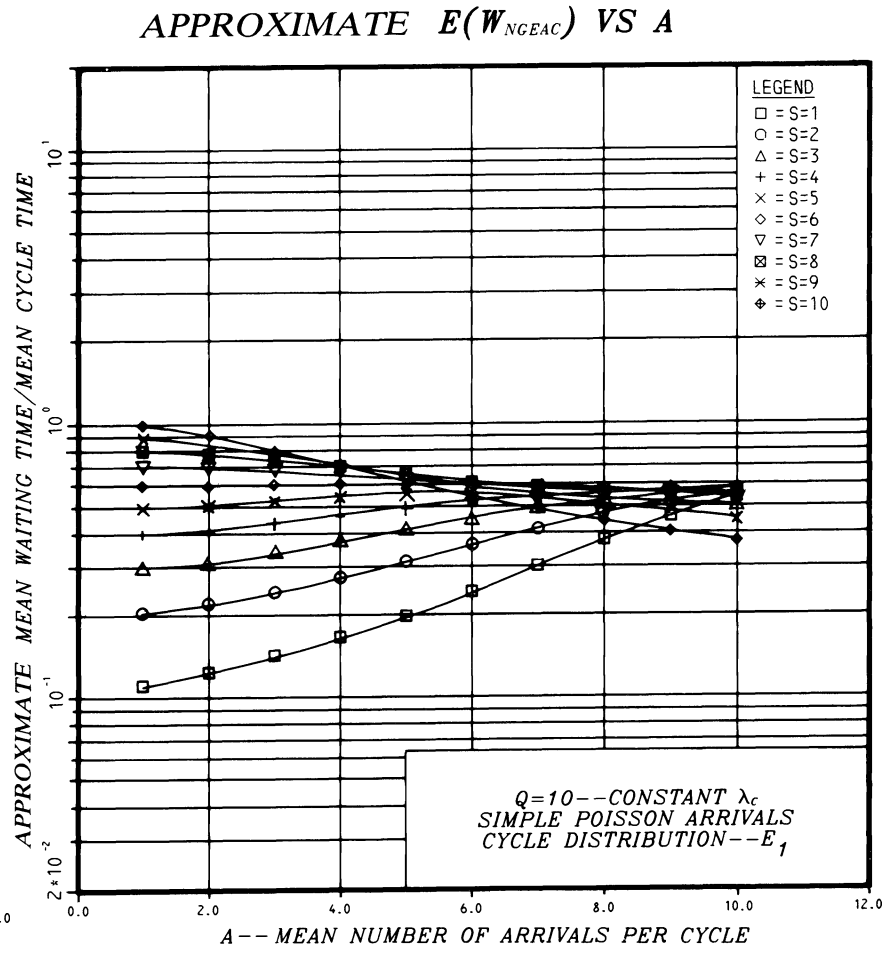

APPROXIMATE $E(W_{NGEAC})$ VS A
LEGEND
□ = S=1
○ = S=2
△ = S=3
+ = S=4
× = S=5
◇ = S=6
▽ = S=7
⊠ = S=8
✳ = S=9
⊕ = S=10
Q=10--CONSTANT $\lambda_c$
SIMPLE POISSON ARRIVALS
CYCLE DISTRIBUTION--$E_1$
APPROXIMATE MEAN WAITING TIME/MEAN CYCLE TIME
$10^1$
$10^0$
$10^{-1}$
$2*10^{-2}$
0.0
2.0
4.0
6.0
8.0
10.0
12.0
A -- MEAN NUMBER OF ARRIVALS PER CYCLE

FIG. 8

The definition of $\nu(t)$ is identical for the NGEAC and NGLAC buffer disciplines, and its analysis will not be repeated in this section.

Let $\Pr[W \leqq t, \text{acceptance} \mid m, n, x, l]$ be the equilibrium probability that a task will be accepted with a waiting time $W$ less than or equal to $t$, given that immediately after the arrival epoch the task is in position 1 of a queue of size $m$ with the system in phase $n$, and the elapsed time of the current phase is $x$. It then follows that the transform of the equilibrium distribution for the waiting time of an accepted task, $\gamma_W(z)$, is given by

$$\gamma_W(z) = \frac{1}{(1-L)\rho_b'(1)} \int_0^\infty dx[\Phi_0 + \Phi_n], \quad \text{where}$$

$$\Phi_0 = \sum_{m=0}^{Q} P_e(m, 0, x) \sum_{k=1}^{\infty} R_V(k|m)$$

$$\cdot \sum_{j=1}^{\min(k,Q)} \gamma_W(z|\min(m+k, Q), 0, x, \sigma(m, 0, j, k)), \tag{20}$$

$$\Phi_n = \sum_{n=1}^{S} \sum_{m=1}^{Q} P_e(m, n, x) \sum_{k=1}^{\infty} R_P(k|m)$$

$$\cdot \sum_{j=1}^{\min(k,Q-1)} \gamma_W(z|\min(m+k, Q), n, x, \sigma(m, n, j, k))$$

where $L$ and $P_e(m, n, x)$ are given in the previous section, and

$$\sigma(m, n, j, k) = \begin{cases} (m+j) - \max(m+k-Q, 0), & n = 0;\ 0 < k < Q;\ 0 \leqq m \leqq Q;\ 0 < j \leqq k, \\ j & n = 0;\ Q \leqq k;\ 0 \leqq m \leqq Q;\ 0 < j \leqq Q, \\ (m+j) - \max(m+k-Q, 0), & 1 \leqq n \leqq S;\ 0 < k < Q;\ 1 \leqq m \leqq Q;\ 0 \leqq j < k, \\ j+1, & 1 \leqq n \leqq S;\ Q \leqq k;\ 1 \leqq m \leqq Q;\ 0 < j \leqq Q-1, \end{cases} \tag{21}$$

$$\gamma_W(z|m, n, x, l) = \int_0^\infty e^{-zt} d_t \Pr[W \leqq t, \text{acceptance} \mid m, n, x, l]. \tag{22}$$

In order to find $\gamma_W(z|m, n, x, l)$, we define a new function, $\alpha = \alpha(z|m, n, l)$, where

$$\alpha(z|m, n, l) = \int_0^\infty e^{-zt} d_t \Pr[W \leqq t, \text{acceptance} \mid m, n, l] \tag{23}$$

where $\Pr[W \leqq t, \text{acceptance} \mid m, n, l]$ is the conditional probability that a task which is in position $l$ of a queue of size $m$ at the beginning of a phase of type $n$ will be accepted with the remaining waiting time $W$ less than or equal to $t$. Elementary considerations lead to the following explicit formula for $\gamma_W(z|m, n, x, l)$:

$$\gamma_W(z|m, n, x, l) = \begin{cases} \sum_{i=0}^{Q-m+l-1} \gamma_V(z, i|x)\alpha(z|\min(Q, m+i), S, l - \max(0, m+i-Q)) \\ \qquad \text{for } n = 0;\ 0 \leqq m \leqq Q;\ l \leqq m, \\ \sum_{i=0}^{Q-m+l-2} \gamma_P(z, i|x)\alpha(z|\min(Q, m+i) - 1, n-1, l-1-\max(0, m+i-Q+1)) \\ \qquad \text{for } 0 < n;\ 1 \leqq m \leqq Q;\ 1 < l \leqq m, \\ \int_0^\infty e^{-zt}\, d_t G_P(x+t)/[1 - G_P(x)] \quad \text{for } 0 < n;\ 1 \leqq m \leqq Q;\ l = 1, \end{cases} \tag{24}$$

where

$$\gamma_V(z, i|x) = \int_0^\infty e^{-zt} R_A(i, t) d_t G_V(x+t)/[1-G_V(x)],$$

$$\gamma_P(z, i|x) = \int_0^\infty e^{-zt} R_A(i, t) d_t G_P(x+t)/[1-G_P(x)]$$

and

$$\alpha(z|m, n, l)$$

$$= \begin{cases} 1 & \text{for } 1 \leqq n \leqq S;\ 1 \leqq m \leqq Q;\ l = 1, \\ \sum_{i=0}^{Q-m+l-1} \gamma_V(z, i|0)\alpha(z|\min(Q, m+i), S, l-\max(0, m+i-Q)) & \\ & \text{for } 1 \leqq n \leqq S;\ 1 \leqq m \leqq Q;\ 1 < l \leqq m, \\ \sum_{i=0}^{Q-m+l-2} \gamma_P(z, i|0)\alpha(z|\min(Q, m+i)-1, n-1, l-1-\max(0, m+i-Q+1)) & \\ & \text{for } 1 \leqq n \leqq S;\ 2 \leqq m \leqq Q;\ 1 < l \leqq m. \end{cases}$$

Numerical methods can be employed for approximating the distribution function that is the inverse of this Laplace-Stieltjes transform, as well as determining moments of the equilibrium waiting time distribution of an accepted task.

**7. Analysis of the GEAC case.** The overall analysis of this final case is identical with that of the NGEAC case, except that the generator there must be replaced by the generator for the GLAC case. The main reason for including this case is the mean waiting time using GEAC may be significantly smaller than for GLAC for arrival rates approaching infinity.

## REFERENCES


[1] T. P. BAGCHI AND J. G. C. TEMPLETON, *Some finite waiting space bulk queueing systems*, Lectures Notes in Economics and Mathematical Systems, Vol. 98, Springer-Verlag, New York, 1973, pp. 133–138.

[2] U. N. BHAT, *Some problems in finite queues*, Lectures Notes in Economics and Mathematical Systems, 98, Springer-Verlag, New York, 1973, pp. 139–156.

[3] D. H. CARBAUGH, G. G. DREW, H. GHIRON AND MRS. E. S. HOOVER, *No. 1 ESS call processing*, Bell System Tech. J., 43 (1964), pp. 2483–2531.

[4] R. W. CONWAY, W. L. MAXWELL AND L. W. MILLER, *Theory of Scheduling*, Addison-Wesley, Reading, MA, 1967.

[5] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[6] A. FABENS, *The solution of queueing and inventory models by semi-Markov processes*, J. Roy. Statist. Soc., 23B (1961), pp. 113–127.

[7] M. EISENBERG, *Queues with periodic service and changeover times*, Operations Res., 20 (1972), pp. 440–451.

[8] J. F. HAYES AND D. N. SHERMAN, *A study of data multiplexing techniques and delay performance*, Bell System Tech. J., 51 (1972), pp. 1983–2011.

[9] J. RIORDAN, *Stochastic Service Systems*, John Wiley, New York, 1962.

[10] S. M. ROSS, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, CA, 1970.

[11] B. T. SMITH ET AL., *Matrix Eigensystem Routines—EISPACK Guide*, Lecture Notes on Computer Science, No. 6, Springer-Verlag, New York, 1974.

[12] T. J. TEOREY AND J. B. PINKERTON, *A comparative analysis of disk scheduling policies*, Proc. 3rd Symp. Operating Systems Principles (October 1971), pp. 114–121.

[13] D. G. KENDALL, *Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain*, Ann. Math. Statist., 24 (1952), pp. 338–354.

[14] J. W. COHEN, *Basic problems of telephone traffic theory and the influence of repeated calls*, Telecommunication Rev., 18 (1957), pp. 49–100.

[15] J. D. C. LITTLE, *An elementary proof of the queueing formula* $L=\lambda W$, Operations Res., 9 (1960), pp. 383–387.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0001 $01.00/0

# SOME MONOTONICITY PROPERTIES OF PARTIAL ORDERS*

R. L. GRAHAM†, A. C. YAO‡ AND F. F. YAO§

**Abstract.** A fundamental quantity which arises in the sorting of $n$ numbers $a_1, a_2, \cdots, a_n$ is $\Pr(a_i < a_j|P)$, the probability that $a_i < a_j$ assuming that all linear extensions of the partial order $P$ are equally likely. In this paper, we establish various properties of $\Pr(a_i < a_j|P)$ and related quantities. In particular, it is shown that $\Pr(a_i < b_j|P') \geqq \Pr(a_i < b_j|P)$ if the partial order $P$ consists of two disjoint linearly ordered sets $A = \{a_1 < a_2 < \cdots < a_m\}$, $B = \{b_1 < b_2 < \cdots < b_n\}$ and $P' = P \cup \{\text{any relations of the form } a_k < b_l\}$. These inequalities have applications in determining the complexity of certain sorting-like computations.

**1. Introduction.** Many algorithms for sorting $n$ numbers $\{a_1, a_2, \cdots, a_n\}$ proceed by using binary comparisons $a_i : a_j$ to build successively stronger partial orders $P$ on $\{a_i\}$ until a linear order emerges (see, e.g., Knuth [2]). A fundamental quantity in deciding the expected efficiency of such algorithms is $\Pr(a_i < a_j|P)$, the probability that the result of $a_i : a_j$ is $a_i < a_j$ when all linear orders consistent with $P$ are equally likely. In this paper, we prove some intuitive but nontrivial properties of $\Pr(a_i < a_j|P)$ and related quantities. These results are important, for example, in establishing the complexity of selecting the $k$th largest number [4].

We begin with a motivating example. Suppose that tennis skill can be represented by a number, so that player $x$ will lose to player $y$ in a tennis match if $x < y$. Imagine a contest between two teams $A = \{a_1, a_2, \cdots, a_m\}$ and $B = \{b_1, b_2, \cdots, b_n\}$, where within each team the players are already ranked as $a_1 < a_2 < \cdots < a_m$ and $b_1 < b_2 < \cdots < b_n$. If the first match of the contest is between $a_1$ and $b_1$, what is the probability $p$ that $a_1$ will lose? Supposing that the two teams have never met before, it is reasonable to assume that all relative rankings among players of $A \cup B$ are equally likely, provided they are consistent with $a_1 < a_2 < \cdots < a_m$ and $b_1 < b_2 < \cdots < b_n$. It is easy to show by a simple calculation that $p = m/(m+n)$. Consider now a different situation when the two teams did compete before with results $a_{i_1} < b_{j_1}$, $a_{i_2} < b_{j_2}$, $\cdots$, $a_{i_t} < b_{j_t}$; in other words, the team $B$ players always won. Let $p'$ be the probability for $a_1 < b_1$ assuming that all orderings of elements in $A \cup B$, consistent with the known constraints, are equally likely. One would certainly expect that $p' \geqq p$, as the additional information indicates that the players on team $B$ are better than those on team $A$. However, the proof of this does not seem to be so trivial. The purpose of this paper is to establish several general theorems concerning such monotone properties.

We now give a proof[1] that $p' \geqq p$ in the preceding example. It establishes the result even when $A$ and $B$ are themselves only partially ordered, provided that $a_1$ and $b_1$ are the unique minimum elements in $A$ and $B$, respectively. Let us denote by $P'$ the partial order obtained by adding the relations $\{a_{i_1} < b_{j_1}, a_{i_2} < b_{j_2}, \cdots, a_{i_t} < b_{j_t}\}$ to $P = A \cup B$. We will show that $\Pr(a_1 < b_1|P')/\Pr(b_1 < a_1|P') \geqq m/n$, from which it follows that $\Pr(a_1 < b_1|P') \geqq m/(m+n) = \Pr(a_1 < b_1|P)$.

* Received by the editors October 5, 1979, and in final revised form November 12, 1979.

† Bell Laboratories, Murray Hill, New Jersey 07974.

‡ Computer Science Department, Stanford University, Stanford, California 94305. The work of this author was supported in part by the National Science Foundation under Grant MCS-77-05313.

§ Xerox Palo Alto Research Center, Palo Alto, California 94304.

[1] The proof given here, simplifying our original proof, is due to D. Knuth.

Consider the sets $S_0$ of all $(m+n-1)!/(m-1)!(n-1)!$ possible sequences of 0's and 1's with one element "underlined", where

(i) the sequence is of length $m+n$, with $m$ 0's and $n$ 1's,

(ii) the first character is 0,

(iii) one of the 1's is underlined.

Define the set $S_1$ similarly but with first character 1 and with one of the 0's underlined. We get a 1–1 correspondence between $S_0$ and $S_1$ by complementing both the first character and the underlined character. If $x_0 \in S_0$ corresponds to $x_1 \in S_1$, then $x_0 < x_1$ in the partial order $<$ defined on (0, 1)-sequences as follows: Say that $x < y$ if we can transform $x$ into $y$ by one or more replacements of '01' by '10'; or, equivalently, $x < y$ if $x$ and $y$ have the same number of 0's and for all $k$, the position of the $k$th 0 of $x$ is no further to the right than the $k$th 0 of $y$. List all the pairs of the correspondence as $x_0 \leftrightarrow x_1$, $y_0 \leftrightarrow y_1, \cdots$.

For a partial order $Q$ on a set $X$, we say that a 1–1 mapping $\lambda : X \to \{1, 2, \cdots, n\}$ is a *linear extension* of $Q$ if $\lambda(x) < \lambda(y)$ whenever $x < y$ in $Q$. Let $\lambda_{x_1}$ be a linear extension of $P'$ which places elements of $A$ into the positions where $x_1$ has a 0, and elements of $B$ into the positions where $x_1$ has a 1. The correspondence $x_0 \leftrightarrow x_1$ naturally associates to $\lambda_{x_1}$ a linear extension $\lambda_{x_0}$ of $P'$ in which the relative order of the $a_i$ and also the relative order of the $b_j$ are both unchanged. We therefore obtain a list of inequalities $N(x_1) \leqq N(x_0)$, $N(y_1) \leqq N(y_0), \cdots$, where $N(x_i)$ denotes the number of all linear extensions $\lambda_{x_i}$ defined above. (For some $x_i$, $N(x_i)$ may be 0.) Summing all the inequalities gives

$$m \cdot (\text{number of linear extensions of } P' \cup (b_1 < a_1)$$

$$\leqq n \cdot (\text{number of linear extensions of } P' \cup (a_1 < b_1),$$

which is what we wanted to show.

The preceding example suggests the following conjecture. Let $A = \{a_1, a_2, \cdots, a_m\}$, $B = \{b_1, b_2, \cdots, b_n\}$, $X = A \cup B$ and $(P, <)$ be a partial order on $X$ which contains no relation of the form $b_j < a_i$ or $a_i < b_j$.

Suppose $E = E_1 \cup \cdots \cup E_r$ and $E' = E'_1 \cup \cdots \cup E'_s$ where $E_i$ and $E'_j$ are events of the form $(a_{i_1} < b_{j_1}) \wedge (a_{i_2} < b_{j_2}) \wedge \cdots \wedge (a_{i_t} < b_{j_t})$.

*Conjecture.* Assuming all linear extensions of $P$ are equally likely, the events $E$ and $E'$ are mutually favorable, i.e.,

$$\Pr(E|P) \Pr(E'|P) \leqq \Pr(EE'|P).$$

In this paper, we shall prove several results related to this conjecture, which in particular imply the conjecture for the case when both $A$ and $B$ are linearly ordered under $P$ (see Corollary 2 to Theorem 1). The general conjecture, however, remains unresolved.

**2. A monotonicity theorem.** In this section, we shall prove a theorem which implies an important special case of the conjecture, namely, the case when $A$ and $B$ are each linearly ordered under $P$. In fact, in this case the conjecture is true even if $P$ includes relations of both of the types $a_i < b_j$ and $b_k < a_l$.

Let $A = \{a_1 < a_2 < \cdots < a_m\}$ and $B = \{b_1 < b_2 < \cdots < b_n\}$ be linear orders. Let $\Lambda$ denote the set of all linear extensions of $P = A \cup B$. A *cross-relation* between $A$ and $B$ is a set $Z \subseteq (A \times B) \cup (B \times A)$, interpreted as a set of comparisons $a_i < b_j$ and $b_k < a_l$. For a cross-relation $Z$, we define $\hat{Z} = \{\lambda \in \Lambda : \lambda(x) < \lambda(y) \text{ for all } (x, y) \in Z\}$.

It will be convenient to represent each $\lambda \in \hat{Z}$ as a lattice path $\bar{\lambda}$ in $\mathbb{Z}^2$ starting from the origin and terminating at the point $(n, m)$ (see Fig. 2). The interpretation is as follows: As we step along $\bar{\lambda}$ starting from (0, 0), if the $k$th step increases the $A$ (or $B$)

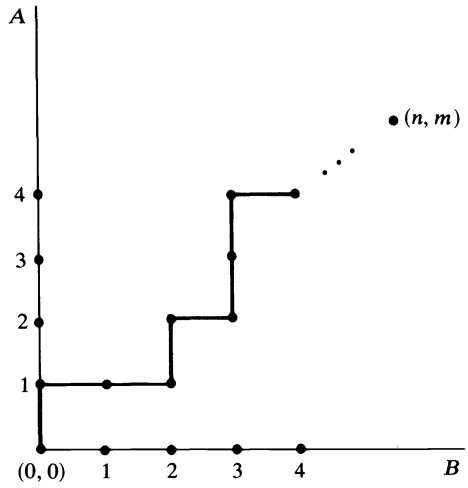



FIG. 1

coordinate from $i-1$ to $i$ then $\lambda$ maps $a_i$ (or $b_i$, respectively) to $k$. Thus, in Fig. 1, $\lambda(a_1)=1, \lambda(b_1)=2, \lambda(b_2)=3, \lambda(a_2)=4$, etc.

Let us consider the geometrical implications of a constraint of the form $\lambda(a_i)<\lambda(b_j)$. By definition, as we go along $\bar{\lambda}$ from (0, 0) to $(n, m)$, $\bar{\lambda}$ must achieve an $A$-value of $i$ before it achieves a $B$-value of $j$. But this means exactly that $\bar{\lambda}$ must not pass through the (closed) vertical line segment joining $(j, i)$ to $(j, 0)$. In general, a set $X \subseteq A \times B$ represents a set of vertical "barriers" of this type which, for any $\lambda \in \hat{X}$, the corresponding lattice path $\lambda$ is prohibited from crossing (Fig. 2). Of course, a set $Y \subseteq B \times A$ corresponds to a set of horizontal barriers in a similar way, with $(b_j, a_i) \in Y$ being represented by the line segment joining $(0, i)$ to $(j, i)$. We will also refer to such vertical and horizontal barriers as $x$-barriers and $y$-barriers. For a cross-relation $Z \subseteq (A \times B) \cup (B \times A)$, we define $Z_X = Z \cap (A \times B)$ and $Z_Y = Z \cap (B \times A)$. Thus $Z_X$ and $Z_Y$ are the vertical and the horizontal barriers determined by $Z$, respectively.

Let $Z$ and $W$ be two cross-relations between $A$ and $B$. We say $Z$ is more *$A$-selective* than $W$ if both $W_X \subseteq Z_X$ and $Z_Y \subseteq W_Y$. (For example, a set of $x$-barriers is

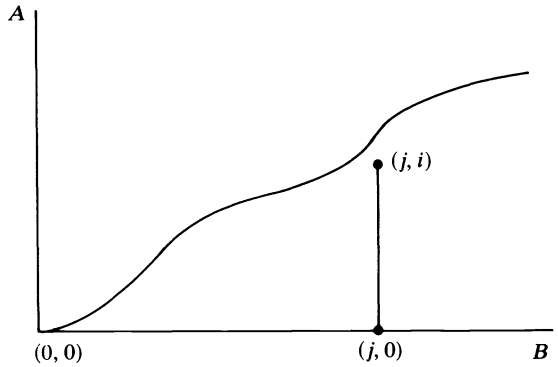



FIG. 2. *A vertical barrier corresponding to the condition* $\lambda(a_i)<\lambda(b_i)$.

always more $A$-selective than a set of $y$-barriers.) Intuitively, one would think that in this case linear extensions of $Z$ should have a greater probability for ranking $A$'s elements below $B$'s. Let $Z'$ and $W'$ be another pair of cross-relations with $Z'$ being more $A$-selective than $W'$. The basic result we prove is the following

THEOREM 1. $|\hat{Z} \cap \hat{Z}'| \cdot |\hat{W} \cap \hat{W}'| \geqq |\hat{Z}' \cap \hat{W}| \cdot |\hat{Z} \cap \hat{W}'|$.

COROLLARY 1. $\Pr(Z'|Z)/\Pr(W'|Z) \geqq \Pr(Z'|W)/\Pr(W'|W)$ *when the denominators are not zero.*

Corollary 1 follows immediately from Theorem 1. It asserts that the ratio $\Pr(Z')/\Pr(W')$ is larger when conditioned on $Z$ than when conditioned on $W$.

COROLLARY 2. $\Pr(V|Z) \geqq \Pr(V|W)$ *for any* $V$ *with* $V_Y \subseteq Z_Y$. *In particular,* $\Pr(X|Z) \geqq \Pr(X|W)$ *for any* $X \subseteq A \times B$.

This follows from Corollary 1 by letting $Z' = V$, and choosing $W'$ so that $W'_X = \varnothing$ and $W'_Y = V_Y$.

*Proof of Theorem* 1. We will construct a $1-1$ mapping of $(\hat{Z}' \cap \hat{W}) \times (\hat{Z} \cap \hat{W}')$ into $(\hat{Z} \cap \hat{Z}') \times (\hat{W} \cap \hat{W}')$. Suppose $\lambda \in \hat{Z}' \cap \hat{W}$ and $\lambda' \in \hat{Z} \cap \hat{W}'$. Let $\bar{\lambda}, \bar{\lambda}'$ be the corresponding lattice paths, and let $\{s_1, s_2, \cdots, s_r\}$ be the set of lattice points common to $\bar{\lambda}$ and $\bar{\lambda}'$.

We assume that the $s_i$ are labeled so that $s_1 = (0, 0)$, $s_r = (n, m)$, and as we move along $\bar{\lambda}$ from $s_1$ to $s_r$, we reach $s_i$ before $s_{i+1}$. Consider the pair of path segments $\bar{\lambda}(s_i, s_{i+1})$ (defined to be the portion of $\bar{\lambda}$ between (and including) $s_i$ and $s_{i+1}$) and $\bar{\lambda}'(s_i, s_{i+1})$. We will call the closed region bounded by these two segments an *olive*, provided that the region is nondegenerate (i.e., $\bar{\lambda}(s_i, s_{i+1})$ and $\bar{\lambda}'(s_i, s_{i+1})$ do not coincide). Let $O_1, O_2, \cdots, O_t$ be the set of olives formed by $\bar{\lambda}$ and $\bar{\lambda}'$. The upper path segment bounding $O_k$ we denote by $O_k^+$; the lower we denote by $O_k^-$. Note that, given $\bar{\lambda} \cup \bar{\lambda}'$, the path $\bar{\lambda}$ can be determined by specifying which $O_i$ contribute $O_i^+$ to $\bar{\lambda}$ and consequently, which $O_j$ contribute $O_j^-$ to $\bar{\lambda}$.

We want to show that for each $\bar{\lambda} \in \hat{Z}' \cap \hat{W}$ with $\bar{\lambda}' \in \hat{Z} \cap \hat{W}'$, we can associate a unique $\bar{\mu} \in \hat{Z} \cap \hat{Z}'$ with $\bar{\mu}' \in \hat{W} \cap \hat{W}'$. In fact, $\bar{\mu}$ and $\bar{\mu}'$ will be constructed from the path segments of $\bar{\lambda}$ and $\bar{\lambda}'$ so that $\bar{\mu} \cup \bar{\mu}' = \bar{\lambda} \cup \bar{\lambda}'$. The rule for obtaining $\bar{\mu}$ (and consequently $\bar{\mu}'$) is as follows:

> Let $\bar{\mu}$ be the same as $\bar{\lambda}$ except that whenever an olive $O_k$ is intersected by a barrier of $Z$ or $W$, we let $O_k^+ \in \bar{\mu}$.

In the example illustrated in Fig. 3, $O_2$ is penetrated (from below) by an $x$-barrier in $Z$–$W$, and $O_4$ is penetrated (from the left) by a $y$-barrier in $W$–$Z$. Note that $\bar{\lambda}$ always contains the lower boundaries $O_k^-$ of the penetrated olives $O_k$. To obtain $\bar{\mu}$, we substitute $O_2^+, O_4^+$ for $O_2^-, O_4^-$ in the path $\bar{\lambda}$.

To show that $\bar{\mu} \in \hat{Z} \cap \hat{Z}'$ and that the complementary path $\bar{\mu}' \in \hat{W} \cap \hat{W}'$ we need only verify that $\bar{\mu}$ and $\bar{\mu}'$ clear their respective sets of barriers in $Z \cup Z'$ and $W \cup W'$.

Suppose $O_k$ is penetrated (from below) by an $x$-barrier in $Z$–$W$, such as $O_2$ in Fig. 3. Then $\bar{\lambda}$ contains $O_k^-$ and $\bar{\lambda}'$ contains $O_k^+$. We want to argue that $O_k^+$ must clear $Z$ and $Z'$, while $O_k^-$ must clear $W$ and $W'$. First of all, if $O_k^+$ clears $W'$ then it clears $W'_Y$ and hence $Z'_Y$. Secondly, $O_k^+$ clears $Z'_X$ since $O_k^-$ clears $Z'$. It follows that $O_k^+$ clears both $Z$ and $Z'$ as desired. The fact that $O_k^-$ clears $W$ and $W'$ can be shown in the same way.

Similarly, if $O_k$ is penetrated by a $y$-barrier in $W$–$Z$, such as $O_4$ in Fig. 3, then assigning $O_k^+$ to $\bar{\mu}$ and $O_k^-$ to $\bar{\mu}'$, will enable $\bar{\mu}, \bar{\mu}'$, to clear their respective barriers.

The mapping $(\bar{\lambda}, \bar{\lambda}') \to (\bar{\mu}, \bar{\mu}')$ is $1-1$, since the path $\bar{\lambda}$ can be reconstructed from $\bar{\mu}$ by substituting $O_k^-$ for $O_k^+$ in those olives $O_k$ penetrated by a barrier of $Z$ or $W$. This completes the proof of Theorem 1. □

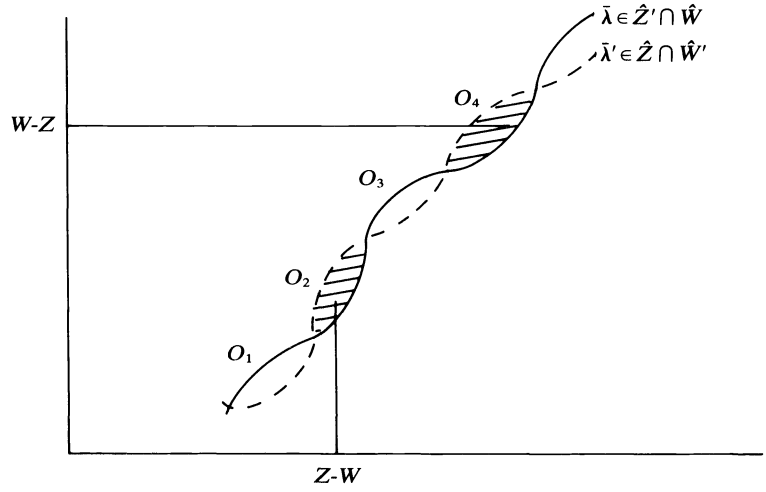



FIG. 3. *Olives which are penetrated by an x-barrier in Z–W and a y-barrier in W–Z.*

**3. Extension to disjunctions of partial orders.** In this section, we will consider pairs of cross relations $(Z, W)$ on $A = \{a_1 < a_2 < \cdots < a_m\}$ and $B = \{b_1 < b_2 < \cdots < b_n\}$, when $Z$ consists of just $x$-barriers and $W$ consists of just $y$-barriers. However, we now incorporate the concept of a disjunction of a set of cross-relations. For a disjunction $\mathscr{X} = \bigcup_i Z_i$, where $Z_i \subseteq (A \times B) \cup (B \times A)$, we let $\hat{\mathscr{X}}$ denote $\bigcap_i \hat{Z}_i$. Suppose $\mathscr{X} = \bigcup_i X_i$ and $\mathscr{Y} = \bigcup_j Y_j$ where $X_i \subseteq A \times B$ and $Y_j \subseteq B \times A$, with $\mathscr{X}' = \bigcup_i X_i'$ and $\mathscr{Y}' = \bigcup_j Y_j'$ defined similarly. The analogue[2] of Theorem 1 is the following

THEOREM 2. $|\hat{\mathscr{X}} \cap \hat{\mathscr{X}}'|\,|\hat{\mathscr{Y}} \cap \hat{\mathscr{Y}}'| \geqq |\hat{\mathscr{X}} \cap \hat{\mathscr{Y}}|\,|\hat{\mathscr{X}}' \cap \hat{\mathscr{Y}}'|$.

As in the case of Theorem 1, here we can also derive as corollaries that $\Pr(\hat{\mathscr{X}}|\hat{\mathscr{X}}')/\Pr(\hat{\mathscr{Y}}|\hat{\mathscr{X}}') \geqq \Pr(\hat{\mathscr{X}}|\hat{\mathscr{Y}}')/\Pr(\hat{\mathscr{Y}}|\hat{\mathscr{Y}}')$, that is, the ratio $\Pr(\hat{\mathscr{X}}/\Pr\hat{\mathscr{Y}})$ is larger when conditioned on $\hat{\mathscr{X}}'$ than when conditioned on $\hat{\mathscr{Y}}$. For the special case that $\mathscr{Y} = \mathscr{Y}' = \varnothing$, we obtain

$$\Pr(\hat{\mathscr{X}}|\hat{\mathscr{X}}') \geqq \Pr(\hat{\mathscr{X}}). \tag{1}$$

*Proof of Theorem* 2. As in the proof of Theorem 1, we will show that for each $\bar{\lambda} \in \hat{\mathscr{X}} \cap \hat{\mathscr{Y}}$ with $\bar{\lambda}' \in \hat{\mathscr{X}}' \cap \hat{\mathscr{Y}}'$, we can associate a unique $\bar{\mu} \in \hat{\mathscr{X}} \cap \hat{\mathscr{X}}'$ with $\bar{\mu} \in \hat{\mathscr{Y}} \cap \hat{\mathscr{Y}}'$. Furthermore, $\bar{\mu}$ and $\bar{\mu}'$, will be constructed from $\bar{\lambda}$ and $\bar{\lambda}'$ by interchanging certain path segments. We may assume without loss of generality that no $X_i$, $X_i'$, $Y_j$, or $Y_j'$ have a barrier which penetrates both $\bar{\lambda}$ and $\bar{\lambda}'$.

Let $O_1, O_2, \cdots, O_t$ be the set of olives formed by $\bar{\lambda}$ and $\bar{\lambda}'$. Thus $\bar{\lambda}$ corresponds to a subset $P \subseteq \{1, 2, \cdots, t\} = T$ such that $\bar{\lambda}$ contains $O_k^+$ iff $k \in P$, and with this association $\bar{\lambda}'$ corresponds to the subset $Q = T - P = P^c$. For a given olive $O_k$, there may be various barriers which intersect it. For each $X_i$, let $G_i$ denote the set $\{k \in T$: a barrier from $X_i$ intersects $O_k\}$. Similarly, define $G_i'$ for $X_i'$, $H_i$ for $Y_i$, and $H_i'$ for $Y_i'$. Observe that

$\bar{\lambda} \in \hat{\mathscr{X}}$ iff $\bar{\lambda} \in \hat{X}_i$ for some $i$,

iff $P \supseteq G_i$ for some $i$,

iff $P \in [\mathscr{G}]_U \equiv$ upper ideal in $2^T$ generated by $\mathscr{G} = \{G_1, G_2, \cdots\}$,

where the meaning of the last statement is as follows.

[2] We could, of course, write this as $|\hat{\mathscr{X}} \cap \hat{\mathscr{X}}'|\,|\hat{\mathscr{Y}} \cap \hat{\mathscr{Y}}'| \geqq |\hat{\mathscr{X}}' \cap \hat{\mathscr{Y}}|\,|\hat{\mathscr{X}} \cap \hat{\mathscr{Y}}'|$ to make it resemble Theorem 1 more.

DEFINITION. For a finite set $T$, let $2^T$ denote the collection of all subsets of $T$ partially ordered by set inclusion (i.e., $C \prec D$ iff $C \supseteq D$). An *upper ideal* in $2^T$ is a subset $\mathcal{U} \subseteq 2^T$ such that if $S \in \mathcal{U}$ then any element $S'$ higher in the partial order (i.e., $S \subseteq S'$) must also be in $\mathcal{U}$. Similarly, a *lower ideal* $\mathcal{L} \subseteq 2^T$ has the property that if $S \in \mathcal{L}$ and $S' \subseteq S$, then $S' \in \mathcal{L}$.

As above, we have

$$\bar{\lambda} \in \hat{\mathcal{Y}} \text{ iff } \bar{\lambda} \in \hat{Y}_j \text{ for some } j,$$

$$\text{iff } P \subseteq H_j^c \text{ for some } j,$$

$$\text{iff } P \in [\mathcal{H}^c]_L \equiv \text{lower ideal in } 2^T \text{ generated by } \mathcal{H}^c = \{H_1^c, H_2^c, \cdots\}.$$

Now, what we are trying to show is that for each $\bar{\lambda} \in \hat{\mathcal{X}} \cap \hat{\mathcal{Y}}$ with $\bar{\lambda}' \in \hat{\mathcal{X}}' \cap \hat{\mathcal{Y}}'$, we can associate a unique $\bar{\mu} \in \hat{\mathcal{X}} \cap \hat{\mathcal{X}}'$ with $\bar{\mu}' \in \hat{\mathcal{Y}} \cap \hat{\mathcal{Y}}'$. Translating this into the language of ideals, we want:

For each $P \in [\mathcal{G}]_U \cap [\mathcal{H}^c]_L$ with $P^c \in [\mathcal{G}']_U \cap [\mathcal{H}'^c]_L$, there can be associated a unique $Q \in [\mathcal{G}]_U \cap [\mathcal{G}']_U$ with $Q^c \in [\mathcal{H}^c]_L \cap [\mathcal{H}'^c]_L$.

We claim that, in fact, we will be able to find such a mapping for *arbitrary* upper ideals $\mathcal{U}, \mathcal{U}'$ and lower ideals $\mathcal{L}, \mathcal{L}'$ in $2^T$. In other words, there is a $1-1$ mapping $(P, P^c) \to (Q, Q^c)$ such that if $P \in \mathcal{U} \cap \mathcal{L}$ and $P^c \in \mathcal{U}' \cap \mathcal{L}'$ then $Q \in \mathcal{U} \cap \mathcal{U}'$ and $Q^c \in \mathcal{L} \cap \mathcal{L}'$. Further, we will restrict the mapping so that

$$P \subseteq Q. \tag{2}$$

If (2) holds then

$$P \in \mathcal{U} \Rightarrow Q \in \mathcal{U} \qquad \text{since } \mathcal{U} \text{ is an upper ideal,}$$

$$P^c \in \mathcal{L}' \Rightarrow Q^c \in \mathcal{L}' \qquad \text{since } \mathcal{L}' \text{ is a lower ideal.}$$

Thus, we want

$$\begin{matrix} P \in \mathcal{U} \cap \mathcal{L} \\ P^c \in \mathcal{U}' \cap \mathcal{L}' \end{matrix} \Rightarrow \begin{matrix} Q \in \mathcal{U}' \\ Q^c \in \mathcal{L} \quad \text{with } P \subseteq Q. \end{matrix}$$

We claim even further that we can find the required mapping for the more general domain

$$\begin{matrix} P \in \mathcal{L} \\ P^c \in \mathcal{U}' \end{matrix} \Rightarrow \begin{matrix} Q \in \mathcal{U}' \\ Q^c \in \mathcal{L} \quad \text{with } P \subseteq Q. \end{matrix}$$

But notice that if $\mathcal{U}'$ is an upper ideal then $\mathcal{U}'^c$ is a lower ideal. Thus, the condition

$$\begin{matrix} P \in \mathcal{L} \\ P^c \in \mathcal{U}' \end{matrix} \Rightarrow \begin{matrix} Q \in \mathcal{U}' \\ Q^c \in \mathcal{L} \quad \text{with } P \subseteq Q \end{matrix}$$

becomes

$$P \in \mathcal{L} \cap \mathcal{U}'^c \equiv \mathcal{W} \Rightarrow Q^c \in \mathcal{W} \quad \text{with } P \subseteq Q,$$

where $\mathcal{W}$, being the intersection of two lower ideals, is also a lower ideal. Of course,

$$P \subseteq Q \quad \text{iff} \quad P \cap Q^c = \varnothing.$$

Thus, the theorem will be proved if we show the following result, which is actually of independent interest:

For an *arbitrary* lower ideal $\mathcal{W}$ in $2^T$, there is always a permutation $\pi : \mathcal{W} \to \mathcal{W}$ such that for all $w \in \mathcal{W}$, $w \cap \pi(w) = \varnothing$.

For each $x \in \mathcal{W}$, let $d(x)$ denote the set $\{w \in \mathcal{W} : x \cap w = \varnothing\}$. By Hall's theorem [1], it is enough to show that

$$\left| \bigcup_{x \in \mathcal{S}} d(x) \right| \geqq |\mathcal{S}|$$

for all $\mathcal{S} \subseteq \mathcal{W}$. In fact, for $\mathcal{S} \subseteq \mathcal{W}$, let $d_{\mathcal{S}}(x)$ denote $d(x) \cap [\mathcal{S}]_L$. What we will actually show is the stronger assertion

$$\left| \bigcup_{x \in \mathcal{S}} d_{\mathcal{S}}(x) \right| \geqq |\mathcal{S}| \tag{3}$$

for any $\mathcal{S} \subseteq 2^T$. So, suppose $\mathcal{S} = \{S_1, \cdots, S_k\}$ with $S_i \subseteq T$. Thus,

$$y \in \bigcup_{x \in \mathcal{S}} d_{\mathcal{S}}(x) \text{ iff } y \in [\mathcal{S}]_L \text{ and } y \cap x = \varnothing \text{ for some } x \in \mathcal{S},$$

$$\text{iff } y \subseteq S_i \text{ for some } i \text{ and } y \cap S_j = \varnothing \text{ for some } j,$$

$$\text{iff } y \subseteq S_i - S_j \text{ for some } i, j.$$

Therefore, if we can in fact show that there are always at least $k$ different sets of the form $S_i - S_j$ then (3) will follow. However, this is exactly the result of Marica and Schönheim [3]. Hence (3) holds and the theorem follows. □

Theorem 2 can be generalized slightly by allowing the partial order $(P, <)$ underlying $\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{X}}', \hat{\mathcal{Y}}'$ to be more than just $A \cup B$; i.e., $P$ may itself include relations of the form $a_i < b_j$ and $b_k < a_l$. In this case, all such relations can also be interpreted as barriers which cannot be crossed by a linear extension $\bar{\sigma}$ of $P$. Since both paths $\bar{\lambda}$ and $\bar{\lambda}'$ avoid all these barriers, then so will any path $\bar{\mu}, \bar{\mu}'$ constructed from their path segments.

**4. Concluding remarks.** We should point out that if we weaken the hypotheses on the structure of $(P, <)$ even slightly, then (1) can fail. To see this, consider the following partial order $(P, <)$ on the set $\{a_1, a_2, b_1, b_2, c\}$ as shown in Fig. 4.

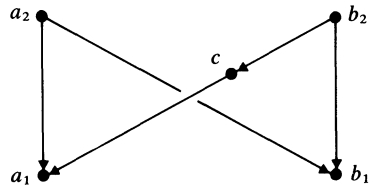



FIG. 4. *An example violating* (1).

Choose $X = X_1 = \{(1, 1)\}$, $X' = X'_1 = \{(2, 2)\}$, and all other $X_i, X'_i, Y_j, Y'_j$ to be $\varnothing$. An easy enumeration yields

$$|\Lambda| = 8, \quad |\hat{X}| = 3 = |\hat{X}'|, \quad |\hat{X} \cap \hat{X}'| = 1.$$

Thus,

$$\Pr(\hat{X}|\hat{X}') = \tfrac{1}{3} < \tfrac{3}{8} = \Pr(\hat{X})$$

which violates (1).

We should also note (as pointed out by D. Kleitman and J. Shearer) that the conjecture is not true if we allow $P$ to have even one relation of the form $a_i < b_j$ as the following example shows.

Let

$$A = \{a_1, a_2\}, \quad B = \{b_1, b_2\}, \quad P = \{a_1 < b_1\},$$
$$E = \{a_1 < b_2\} \quad \text{and} \quad E' = \{a_2 < b_1\}.$$

A simple calculation shows that

$$\Pr(E|P)\Pr(E'|P) = \tfrac{2}{3} \cdot \tfrac{2}{3} = \tfrac{4}{9} > \tfrac{5}{12} = \Pr(EE'|P).$$

*Note added in proof.* V. Chvátal has pointed out that the fact we use concerning mappings of lower ideals seems to be due to Erdös, Herzog and Schönheim in *An extremal problem on the set of noncoprime divisors of a number*, Israel J. Math., (1970), pp. 408–412. Very recently, L. A. Shepp has managed to settle the conjecture in the affirmative by an ingenious application of the FKG inequalities.

## REFERENCES

[1] P. HALL, *On representations of subsets*, J. London Math. Soc. 10 (1935), pp. 26–30.

[2] D. E. KNUTH, *The Art of Computer Programming*, Vol. 3, Sorting and Searching, Addison-Wesley, Reading, MA., 1973.

[3] J. MARICA AND J. SCHÖNHEIM, *Differences of sets and a problem of Graham*, Canad. Math. Bull., 12 (1969), pp. 635–637.

[4] A. C. YAO AND F. F. YAO, *On the average-case complexity of selecting the k-th best*, Proc. 19th Annual IEEE Symp. on Foundations of Computer Science, Ann Arbor, Michigan 1978, pp. 280–289.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0002 $01.00/0

# A CHARACTERIZATION OF WEIGHTED ARITHMETIC MEANS*

J. ACZÉL† AND C. WAGNER‡

**Abstract.** We prove, among other things, that the set of weighted arithmetic means is identical with the set of functions $f: R^n \to R$ satisfying

(i) $\min\{x_j\} \leqq f(x_1, x_2, \cdots, x_n) \leqq \max\{x_j\}$

and

(ii) for $k = 2, 3$: $\sum_{i=1}^{k} x_{ij} = s$ $(j = 1, 2, \cdots, n) \Rightarrow \sum_{i=1}^{k} f(x_{i1}, x_{i2}, \cdots, x_{in}) = s$.

We call a function $f: R^n \to R$ an *averaging function* if

$$\min\{x_j\} \leqq f(x_1, x_2, \cdots x_n) \leqq \max\{x_j\}, \tag{1}$$

and a *weighted arithmetic mean* if $f(x_1, x_2, \cdots, x_n) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$, where $0 \leqq w_j \leqq 1$ and $w_1 + w_2 + \cdots + w_n = 1$. It is easy to check that among the familiar averaging functions (weighted arithmetic, geometric, and harmonic means, weighted medians) weighted arithmetic means uniquely enjoy, for all $k \geqq 1$, what we shall call the *k-allocation property*:

For all $s \in R$, if $(x_{ij})$ is a $k \times n$ matrix with $x_{1j} + x_{2j} + \cdots + x_{kj} = s$ for $1 \leqq j \leqq n$, then $f(x_{11}, x_{12}, \cdots, x_{1n}) + f(x_{21}, x_{22}, \cdots, x_{2n}) + \cdots + f(x_{k1}, x_{k2}, \cdots, x_{kn}) = s$.

We prove in this note that the $k$-allocation property assumed only for $k = 2$ and $k = 3$, characterizes weighted arithmetic means in the set of all averaging functions. In fact, we obtain the following more general result:

THEOREM. *The function $f: R^n \to R$ satisfies the k-allocation property for $k = 2$ and $k = 3$ and is continuous at a point or bounded from one side on an (n-dimensional) interval or just on a set of positive measure if and only if there exist real numbers $w_1, w_2, \cdots, w_n$ with $w_1 + w_2 + \cdots + w_n = 1$ such that $f(x_1, x_2, \cdots, x_n) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$.*

*Proof.* To postulate the $k$-allocation property for $k = 2, 3$ is equivalent to assuming that, for all $s \in R$,

$$f(x_1, x_2, \cdots, x_n) + f(s - x_1, s - x_2, \cdots, s - x_n) = s \tag{2}$$

and

$$f(x_1, x_2, \cdots, x_n) + f(y_1, y_2, \cdots, y_n) + f(s - x_1 - y_1, s - x_2 - y_2, \cdots, s - x_n - y_n) = s. \tag{3}$$

Setting $s = x_1$ in (2), and writing

$$-f(0, u_2, \cdots, u_n) = g(u_2, \cdots, u_n), \tag{4}$$

we have

$$f(x_1, x_2, \cdots x_n) = x_1 + g(x_1 - x_2, \cdots, x_1 - x_n). \tag{5}$$

Setting $s = x_1 + y_1$ in (3), and writing $u_j = x_1 - x_j$ and $v_j = y_1 - y_j$ $(2 \leqq j \leqq n)$, it

* Received by the editors August 8, 1979.

† Universität Graz and University of Waterloo, Ontario, Canada. The work of this author was supported in part by a leave Fellowship of the Social Science and Humanities Research Council of Canada under Grant 451-790424.

‡ Center for Advanced Study in the Behavioral Sciences and University of Tennessee, Knoxville, Tennessee 37916. The work of this author was supported in part by the National Science Foundation under Grant BNS 76-22943 A 02, the Andrew W. Mellon Foundation, and the University of Tennessee.

follows from (4) and (5) that

$$g(u_2, \cdots, u_n) + g(v_1, \cdots, v_n) - g(u_2 + v_2, \cdots, u_n + v_h) = 0,$$

i.e.,

$$g(u_2 + v_2, \cdots, u_n + v_n) = g(u_2, \cdots, u_n) + g(v_2, \cdots, v_n). \tag{6}$$

By [2] and [1, pp. 215–16 and p. 32], the general solution of (6), continuous at a point or bounded from one side on an interval or on a set of positive measure, is

$$g(u_2, \cdots, u_n) = a_2 u_2 + \cdots + a_n u_n. \tag{7}$$

Hence the general solution of (2) and (3), under these same weak regularity conditions, is, by (5),

$$\begin{aligned} f(x_1, x_2, \cdots, x_n) &= (1 + a_2 + \cdots + a_n)x_1 - a_2 x_2 - \cdots - a_n x_n \\ &= w_1 x_1 + w_2 x_2 + \cdots + w_n x_n, \end{aligned} \tag{8}$$

with $w_1 + w_2 + \cdots + w_n = 1$, as asserted. Note that, in (8), one or more of the numbers $w_j$ may be negative.

Now if $f$ is assumed to be an averaging function, the aforementioned boundedness conditions are clearly satisfied, and setting $x_j = 1$ and $x_k = 0$ for $k \neq j$ yields $0 \leqq w_j \leqq 1$. Thus we have as a corollary to the above theorem the following characterization of weighted arithmetic means:

COROLLARY. *Let $f: R^n \to R$. Then $f$ is a weighted arithmetic mean if and only if $f$ is an averaging function satisfying the k-allocation property for $k = 2$ and $k = 3$.*

*Remark* 1. In the statement of the above corollary the averaging function condition (1) may be replaced by a considerably weaker supposition. It is clearly sufficient, for example, that there exist some $c > 0$ (no matter how small) such that $0 \leqq x_j \leqq c$ $(j = 1, 2, \cdots, n)$ implies $0 \leqq f(x_1, x_2, \cdots, x_n)$.

*Remark* 2. The foregoing theorems arose in connection with a study of arithmetic averaging as a method of amalgamating a set of individual opinions as to the most appropriate values of some sequence of decision variables. (See [3].)

Suppose that there are $k$ decision variables and $n$ individuals and we denote by $x_{ij}$ the opinion of individual $j$ as to the most appropriate value of variable $i$. In many of these problems (such as the allocation of a fixed sum of money among $k$ competing projects) the column sums of the matrix $(x_{ij})$ are required to have a common value $s$. If a group adopts as the consensual value of variable $i$ the weighted arithmetic average $\bar{x}_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in}$ the consensual values have the highly desirable property $\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k = s$. The above results assert that weighted arithmetic means are the only averaging functions with this property.

## REFERENCES

[1] J. ACZÉL, *Lectures on Functional Equations and Their Applications*, Academic Press, New York, 1966.

[2] A. OSTROWSKI, *Mathematische Miszellen XIV. Über die Funktionalgleichung der Exponentialfunktion und verwandte Funktionalgleichungen*, Jber. Deutsch. Math. Verein, 38 (1929), pp. 54–62.

[3] C. WAGNER, *Consensus through respect: a model of rational group decision-making*, Philos. Studies, 34 (1978), pp. 335–349.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0003 $01.00/0

# A CANONICAL REPRESENTATION OF SIMPLE PLANT LOCATION PROBLEMS AND ITS APPLICATIONS*

GERARD CORNUEJOLS†, GEORGE L. NEMHAUSER‡ AND LAURENCE A. WOLSEY§

**Abstract.** We consider a location problem whose mathematical formulation is $\max_S\{v(S)\colon S\subseteq N, |S|=K\}$ where $v(S)=v(\emptyset)+\sum_{i\in I}\max_{j\in S}c_{ij}$, $C=(c_{ij})$ is a real matrix with row index set $I=\{1,\cdots,m\}$ and column index set $N=\{1,\cdots,n\}$ and $K\leqq n$ is a positive integer. A set function of the form $v(S)$ is called a simple location function. We give a constructive proof that a set function $w$ is a simple location function if and only if it can be represented in the canonical form $w(S)=r_\emptyset+\sum_{T\cap S\neq\emptyset}r_T$ with $r_T\geqq 0$ for all $\emptyset\subset T\subset N$. The proof is also a polynomial algorithm for reducing the matrix $C$ to the canonical form. We give two applications of this canonical representation. The first is that a large class of algorithms for the location problem need to enumerate all but $l$ of the feasible solutions in order to find one of the $l$th best solutions. The second application is the derivation of new integer programming formulations of the location problem. Some of these formulations seem to be easier to solve than the standard one. We conclude the paper with two alternate representations of set functions and, in each case, characterize the nondecreasing, submodular and simple location functions.

**1. Introduction.** In this paper we continue the study of uncapacitated location problems using a canonical representation of their objective functions. Related work is contained in [1], [2], [3], [7], [8], [9].

Let $C=(c_{ij})$ be a real matrix with column index set $N=\{1,\cdots,n\}$ and row index set $I=\{1,\cdots,m\}$, For each $S\subseteq N$ define

$$v(S)=v(\emptyset)+\sum_{i\in I}\max_{j\in S}c_{ij}, \tag{1.1}$$

where $v(\emptyset)$ is any real number. Set functions of the form (1.1) will be called *simple location functions.* It is easily shown that simple location functions are

(a) *nondecreasing:* $v(S)\geqq v(T)$, $\forall\, S\subseteq T\subseteq N$ whenever $v(\emptyset)\geqq v(\{i\})$ for all $i\in N$ and

(b) *submodular:* $v(S)+v(T)\geqq v(S\cup T)+v(s\cap T)$, $\forall S, T\subseteq N$ whenever $v(\emptyset)\leqq v(\{i\})+v(\{j\})-v(\{i,j\})$ for all $i, j\in N$.

The well-known simple *K-plant location problem* is

$$\max_{S\subseteq N}\{v(S)\colon |S|=K\}, \tag{$L_K$}$$

where $N$ is a set of potential facility locations, $I$ is a set of clients, $c_{ij}$ is the profit obtained by serving client $i$ from facility $j$, $K$ is the number of facilities to be chosen, and $v(S)$, given by (1.1), is the net profit derived from the subset of facilities $S$. Problem $(L_K)$ can have as its input any positive integers $n\geqq K$ and $m$ and any $m\times n$ real matrix $C$.

A general set function $w$ can be expressed in the canonical form

$$w(S)=r_\emptyset+\sum_{T\cap S\neq\emptyset}r_T \quad \forall S\subseteq N. \tag{1.2}$$

* Received by the editors August 28, 1979 and in revised form January 2, 1980.

† Carnegie-Mellon University, Pittsburgh, Pennsylvania. The work of this author was supported by the National Science Foundation under grant ENG-7902506 to Carnegie-Mellon University.

‡ Cornell University, Ithaca, New York. The work of this author was supported by the National Science Foundation under grant ENG-7500568 to Cornell University.

§ Center for Operations Research and Econometrics, University of Louvain, Louvain-La-Neuve, Belgium. The work of this author was supported, in part, by a Senior Visiting Research Fellowship from the Science Research Council while visiting the London School of Economics.

The major part of this paper is concerned with the representation of simple location functions in this canonical form. In §2 we give a constructive proof of the fact that $w$ is a simple location function if and only if $r_T \geqq 0$, $\emptyset \subset T \subset N$. The proof is also an algorithm for reducing the matrix $C$ of (1.1) to a canonical form.

Sections 3–5 present applications of the canonical representation. Let $z$ be a nondecreasing set function that satisfies: if for any $S \subset N$ and $j \notin S$, $z(S \cup \{j\}) = z(S)$, then $z(T \cup \{j\}) = z(T)\ \forall T \supset S$. In § 3 we give a constructive proof of a result of Kreps [6], which states that a set function $z$ that satisfies the above condition is "order equivalent" to some simple location function $v$ in the sense that $v(S) \geqq v(T)$ if and only if $z(S) \geqq z(T)$, $\forall\ S, T \subseteq N$.

In § 4 we use this result on order equivalence to strengthen an earlier result on problem $(L_K)$. We show that algorithms for $(L_K)$ that only have access to the values $v(S)$ need to enumerate all but $l$ of the feasible solutions in order to guarantee finding one of the $l$th best solutions.

In § 5 we give new integer programming formulations of $(L_K)$. The linear programming relaxations of some of these integer programs may, because of their dimensions, have computational advantages.

In § 6 we introduce two other representations of general set functions and establish their relationships to (1.2). In particular, we characterize the nondecreasing submodular functions and the simple location functions for each representation.

**2. A canonical representation of simple location functions.** A simple location function $v$ is defined by (1.1) from a matrix $C$ and $v(\emptyset)$. However, matrix $C$ cannot be determined from $v$. In this section we show constructively that a matrix of a very simple form, which we call the *canonical matrix*, can be uniquely determined from $C$. We say that a row of a matrix is in *canonical form* if

(a) it has a nonzero element,
(b) all of the nonzero elements are equal, and
(c) if it has a negative element then it has no zero elements.

A matrix is in canonical form if all of its rows are, and if

(d) for all $T \subseteq N$ there is at most one row with nonzero elements for the subset of columns $T$ and zero elements for $N - T$.

Note that each row of a canonical matrix $R$ can be indexed by the set $\emptyset \subset T \subseteq N$ of nonzero entries having value $r_{Tj} = r_T$, $j \in T$ and $r_{Tj} = 0$, $j \in N - T$. Thus for all $S \subseteq N$,

$$v(S) = r_\emptyset + \sum_{\emptyset \subset T \subseteq N} \max_{j \in S} r_{Tj} = r_\emptyset + \sum_{T \cap S \neq \emptyset} r_T,$$

where $r_\emptyset = v(\emptyset)$.

The construction of a canonical matrix is based upon a simple procedure that is applied iteratively to each nonzero row that is not in canonical form. An application of the procedure replaces a row by two rows, at least one of which is in canonical form. Below we describe the procedure and prove its validity in a slightly more general setting than is needed now. We can assume that zero rows are deleted before the procedure is applied.

*Canonical Reduction Procedure.* Suppose row $i$ of $C$ is not in canonical form and, for simplicity of notation, $c_{ij} \geqq c_{i,j+1}$, $j = 1, \cdots, n-1$. Let $c_{ip}$ be the nonzero element of largest index. Replace row $i$ by the two rows $i_1$ and $i_2$ with

$$c_{i_1 j} = \begin{cases} c_{ip}, & j = 1, \cdots, p, \\ 0, & j > p, \end{cases}$$

$$c_{i_2 j} = c_{ij} - c_{i_1 j}, \qquad j = 1, \cdots, n.$$

*Examples.* (2 0 −1) is replaced by (−1 −1 −1 ) and (3 1 0); (2 1 0) is replaced by (1 1 0) and (1 0 0).

Row $i_1$ is in canonical form. In row $i_2$ the index of the smallest nonzero element has been reduced by at least one. Rows $i_k$, $k=1, 2$, have the property that $c_{i_kj} \geqq c_{i_kj+1}$, $j=1, \cdots, n-1$.

PROPOSITION 2.1. *Let*

$$f(y)=\max\left\{\sum_{j\in N} c_{ij}x_{ij}\colon \sum_{j\in N} x_{ij}=1,\ 0\leqq x_{ij}\leqq y_j,\ j\in N\right\}$$

*and*

$$f'(y)=\max\left\{\sum_{k=1}^{2}\sum_{j\in N} c_{i_kj}x_{i_kj}\colon \sum_{j\in N} x_{i_kj}=1,\ 0\leqq x_{i_k}\leqq y_j,\ k=1,2,\ j\in N\right\},$$

*where* $y_j \geqq 0$, $j\in N$ *and* $\sum_{j\in N} y_j \geqq 1$. *Then* $f(y)=f'(y)$.

*Proof.* The fact that sequences $\{c_{ij}\}_{j\in N}$ and $\{c_{i_kj}\}_{j\in N}$, $k=1, 2$, are nonincreasing implies $x_{ij}=x_{i_1j}=x_{i_2j}$ $\forall j\in N$ for any $y$ that satisfies $\sum_{j\in N} y_j \geqq 1$. The result then follows from $c_{ij}=c_{i_1j}+c_{i_2j}$ $\forall j\in N$. □

THEOREM 2.1. *Given an* $m\times n$ *matrix* $C$, *there exists uniquely (up to row permutations) an* $m'\times n$ *matrix* $R$ *with rows indexed by* $\varnothing \subset T\subset N$ *such that*

(a) $$r_{Tj}=\begin{cases} r_T & \text{if } j\in T,\\ 0 & \text{if } j\notin T,\end{cases}$$

(b) $$r_T \geqq 0, \qquad \varnothing\subset T\subset N,$$

(c) $$m'\leqq \min\,(mn-m+1,\ 2^n-1),$$

(d) $$v(S)=r_\varnothing+\sum_{i\in I}\max_{j\in S} c_{ij}=r_\varnothing+\sum_{\varnothing\subset T\subseteq N}\max_{j\in S} r_{Tj}=r_\varnothing+\sum_{T\cap S\neq\varnothing} r_T \quad \forall S\subseteq N.$$

*Proof.* Applying the canonical reduction procedure at most $n$ times to each of the $m$ rows of $C$ we obtain a unique matrix $R'$ with each row satisfying (a). Also note that the procedure only produces a row with negative elements when $p=n$ and in this case $T=N$. Therefore (b) holds. Finally recombining rows of $R'$ which have nonzero entries in identical sets of columns, in particular the $m$ possible cases when $T=N$, gives a unique matrix $R$ satisfying (c). Property (d) follows from the repeated application of Proposition 2.1 with $y=y^S$, the characteristic vector of $S$. □

Theorem 2.1 implies

COROLLARY 2.1. $z(S)=r_\varnothing+\sum_{T\cap S\neq\varnothing} r_T$ *is a simple location function if and only if* $r_T\geqq 0$ $\forall\varnothing\subset T\subset N$.

### 3. The order structure of simple location functions.

DEFINITION. Two set functions $w_1$ and $w_2$ that satisfy $w_1(S)\geqq w_1(T)$ if and only if $w_2(S)\geqq w_2(T)$ $\forall S, T\subseteq N$ are *order equivalent.*

Note that this definition implies that $w_1(S)>w_1(T)$ when $w_2(S)>w_2(T)$ and $w_1(S)=w_1(T)$ when $w_2(S)=w_2(T)$.

Let $z$ be a nondecreasing set function that satisfies the following condition:

(3.1) if $z(S\cup\{j\})=z(S)$ for some $S\subset N$ and $j\notin S$, then $z(T\cup\{j\})=z(T)$ for all $T\supset S$.

Strictly increasing functions and nondecreasing submodular functions are classes of set

functions that satisfy (3.1).

In a very different context and by a quite different technique, the following theorem has been proved by Kreps [6].

THEOREM 3.1. *If $z$ is a nondecreasing set function and satisfies condition* (3.1), *then $z$ is order equivalent to a simple location function.*

*Proof.*[1] Given a $z$ that is nondecreasing and satisfies (3.1), we will construct $\{r_T\}$ with $r_T \geqq 0, \forall T \subseteq N$ so that $v$ defined by $v(S) = r_\varnothing + \sum_{T \cap S \neq \varnothing} r_T$ is order equivalent to $z$.

Define the *closure* of $S$ to be $\bar{S} = S \cup \{j \notin S : z(S \cup \{j\}) = z(S)\}$. $S$ is said to be *closed* if $\bar{S} = S$; sets that are not closed are called *open*. Suppose that $S$ is open and $j \in \bar{S} - S$. Then if $T \supset S$ and $j \notin T$, (3.1) implies $j \in \bar{T} - T$ so that $T$ is open. Now, if $j \in \bar{S} - S$ and $v$ exists, we have

$$v(S \cup \{j\}) - v(S) = \sum_{T \subseteq N - S - \{j\}} r_{T \cup \{j\}} = \sum_{T \supseteq S} r_{(N-T) \cup \{j\}} = 0.$$

Thus from nonnegativity we obtain $r_{(N-T)\cup\{j\}} = 0 \; \forall T \supseteq S$. Hence $r_{N-S} = 0$ and $r_{N-T} = 0$ for precisely those sets $T \supseteq S$, $j \notin T$, which (3.1) implies are open. It remains to compute the values of $r_{N-S}$ for all closed sets $S$. This is done by the following procedure.

*Initialization*: We begin by ordering the closed sets $N = S^0, S^1, \cdots, S^t$ so that $z(S^0) > z(S^1) \geqq \cdots \geqq z(S^{t-1}) \geqq z(S^t)$. Then we compute the $\{r_{N-S^i}\}$ recursively in the order $i = 0, 1, \cdots, t$. Define $r_\varnothing = r_{N-S^0} = 0$ and $v(N) = 2^t - 1$. Set $k = 1$.

*Step $k$*: Suppose $r_{N-S^i}$, $i = 0, \cdots, k-1$ have been fixed with the properties

(i) $r_{N-S^i} \geqq 0, \qquad i = 0, \cdots, k-1,$

(ii) $\sum_{i=0}^{k-1} r_{N-S^i} = 2^{k-1} - 1$ and

(iii) $v(S^i) = v(N) - \sum_{T \supseteq S^i} r_{N-T}, \qquad i = 0, \cdots, k-1$ satisfy the order equivalence property.

(Note that $v(S^i)$ is well defined in (iii). To see this consider any $T \supseteq S^i$. Since $S^i$ is closed, either $z(T) > z(S^i)$ or $T = S^i$. Consequently, either $T$ is not closed or $T = S^l$ for some $l \leqq i$. In all cases the value of $r_{N-T}$ has been fixed.) Suppose that we require $v(S^{k-1}) > v(S^k) = \cdots = v(S^{k+j})$, where $j$ is defined so that $v(S^{k+j}) > v(S^{k+j+1})$ or $k + j = t$.

To simplify notation define $\bar{r}_i = r_{N-S^{k+i}}$, $i = 0, \cdots, j$. Choose $\{\bar{r}_i\}_{i=0}^j$ to satisfy $\sum_{i=0}^j \bar{r}_i = 2^{k+j} - 2^{k-1}$ and the $j$ equations

$$\bar{r}_0 - \bar{r}_i = \sum_{T \subset N - S^{k+i}} r_T - \sum_{T \subset N - S^k} r_T, \qquad i = 1, \cdots, j. \tag{3.2}$$

The coefficient matrix of these $j+1$ equations is nonsingular so there is a unique solution. We will show that the solution yields:

(a) $\sum_{i=0}^{k+j} r_{N-S^i} = 2^{k+j} - 1;$

(b) $v(S^k) = \cdots = v(S^{k+j});$

(c) $v(S^k) < v(S^{k-1});$

(d) $r_{N-S^{k+i}} \geqq 0, \; i = 0, \cdots, j.$

[1] The proof is essentially independent of § 2; it only uses the trivial "if" part of Corollary 2.1.

*Proof of* (a).

$$\sum_{i=0}^{k+j} r_{N-S^i} = \sum_{i=0}^{k-1} r_{N-S^i} + \sum_{i=0}^{j} \bar{r}_i = 2^{k-1} - 1 + 2^{k+j} - 2^{k-1} = 2^{k+j} - 1.$$

*Proof of* (b). Equation (3.2) implies

$$v(N) - v(S^k) = \sum_{T \subseteq N - S^k} r_T = \sum_{T \subseteq N - S^{k+i}} r_T = v(N) - v(S^{k+i}), \qquad i = 1, \cdots, j.$$

*Proof of* (c).

$$\sum_{i=0}^{j} \bar{r}_i = 2^{k-1}(2^{j+1} - 1) = 2^{k-1} \sum_{i=0}^{j} 2^i \geqq 2^{k-1}(j+1),$$

which implies that there exists $t$ such that $\bar{r}_t \geqq 2^{k-1}$. Thus

$$v(S^k) = v(S^{k+t}) = v(N) - \sum_{T \subseteq N - S^{k+t}} r_T \leqq v(N) - 2^{k-1} < v(N) - \sum_{i=0}^{k-1} r_{N-S^i} \leqq v(S^{k-1}).$$

*Proof of* (d). $\sum_{i=0}^{k-1} r_{N-S^i} = 2^{k-1} - 1$ and (3.1) imply $\bar{r}_p - \bar{r}_i \leqq 2^{k-1} - 1$, $i, p = 0, \cdots, j$, $i \neq p$. Furthermore, as shown in (c), there exists $t$ such that $\bar{r}_t \geqq 2^{k-1}$. Thus $\bar{r}_i \geqq 2^{k-1} - (2^{k-1} - 1) \geqq 1$ for all $i = 0, \cdots, j$.

The recursive procedure is continued with $k$ replaced by $k + j + 1$. □

**4. Black box algorithms for the simple *K*-plant location problem.** Here we use Theorem 3.1 to obtain a result about a class of algorithms for the optimization problem $(L_K)$. We observe first that any algorithm (approximate or exact) in which all decisions are based only on the relative values of pairs of subsets will behave identically whether applied to the problem

$$(Q_K) \qquad \max_{S \subseteq N} \{z(S): |S| = K\},$$

where $z$ is a nondecreasing set function that satisfies the condition (3.1), or to the location problem $(L_K)$ with $v$ order equivalent to $z$. We call these *ordinal algorithms*. In particular, certain heuristics for $(L_K)$ and $(Q_K)$ such as "greedy" and "interchange" [1] are ordinal algorithms. We define a *black-box algorithm* (see [8]) to be one in which all decisions are based only on function values. Evidently ordinal algorithms are a subset of black-box algorithms.

THEOREM 4.1. *Every black-box algorithm for* $(L_K)$ *that guarantees to find one of the* $l$*th-best solutions, where* $l$ *is any positive integer, needs to enumerate all but* $l$ *of the feasible solutions.*

*Proof.* Let $z(S) = K + 1$ for all the sets $S \subseteq N$ of cardinality larger than $K$ and for $l$ of the sets of cardinality $K$; $z(S) = |S|$ for all the other sets $S \subseteq N$. Obviously $z$ is nondecreasing and satisfies condition (3.1). Therefore, by Theorem 3.1, it is order equivalent to a location function $v$. Since $v(S_1) = v(S_2)$ if $|S_1| = |S_2| \neq K$, no information can be gained from the values of $v(S)$ if $|S| \neq K$. On sets of size $K$ a black-box algorithm will have the same behavior for $v$ and $z$. Consider $z$. To find one of the $l$ sets of size $K$ with $z(S) = K + 1$ one needs to enumerate all but $l$ of the sets of cardinality $K$ since, if $(l+1)$ sets are omitted and it turns out that the value of all the evaluated sets is $K$, then it is impossible to determine which of the $(l+1)$ remaining sets have value $K + 1$. □

**5. Integer programming formulations of the simple *K*-plant location problem and related linear programs.** The most familiar and widely cited integer linear programming

formulation of the simple $K$-plant location problem is

$$\max \sum_{i \in I} \sum_{j \in N} c_{ij} x_{ij}, \tag{5.1}$$

$$\sum_{j \in N} x_{ij} = 1 \quad \forall i \in I, \tag{5.2}$$

$$x_{ij} - y_j \leqq 0 \quad \forall i \in I, j \in N, \tag{5.3}$$

$$\sum_{j \in N} y_j = K, \tag{5.4}$$

$$x_{ij} \geqq 0 \quad \forall i \in I, j \in N, \tag{5.5}$$

$$y_j \in \{0, 1\} \quad \forall j \in N, \tag{5.6}$$

where $y_j = 1$ means that facility $j$ is chosen. The $y_j$'s are the strategic variables since given an $S \subseteq N$ and its characteristic vector $y^S$ ($y_j^S = 1$, $j \in S$ and $y_j^S = 0$, otherwise) an optimal set of $x_{ij}$'s is given for all $i \in I$ by

$$x_{ij}^S = \begin{cases} 1 & \text{for some } j \in S \text{ such that } c_{ij} = \max_{k \in S} c_{ik}, \\ 0 & \text{otherwise.} \end{cases}$$

In this formulation, and in the ones to follow, upper bounds on the $y_j$'s are unnecessary and we obtain linear programming relaxations by replacing (5.6) by

$$y_j \geqq 0 \quad \forall j \in N. \tag{5.7}$$

This integer program and its linear programming relaxation have $mn + n$ variables and $mn + m + 1$ constraints. To solve this linear program or its dual with a general simplex routine requires basis matrices of size larger than $mn \times mn$.[2]

The objective of this section is to give several new integer programming formulations and linear programming relaxations. Some of these relaxations may have computational advantages. In particular, one of the linear programs has a dual with only $n$ constraints other than upper bounds and nonnegativity, and another has only $m + n$ variables. Each of the linear programs has the same optimal value as the linear program (5.1)–(5.5), (5.7).[3]

Consider the $K$-plant location problem in the canonical from given by matrix $R$. $R$ has $n$ columns and no more than $m(n-1)+1$ rows. All nonzero elements in row $T$ equal $r_T$ and $r_T > 0$ except for $T = N$. Let $F = \{T \subset N : r_T > 0\}$. The problem can be stated as

$$\max \sum_{T \in F} \sum_{j \in T} r_T x_{Tj}, \tag{5.8}$$

$$\sum_{j \in N} x_{Tj} = 1 \quad \forall T \in F, \tag{5.9}$$

[2] Schrage [10] treats (5.3) as variable upper bound constraints and therefore can work with "basis" matrices of size $(m+1) \times (m+1)$. This specialized approach is not available in general simplex routines. See Guignard and Spielberg [4] for the results of applying various decomposition schemes to [(5.1) – (5.6)].

[3] Size is not, of course, the only consideration. Replacing (5.3) by the $n$ constraints $\sum_{i \in I} x_{ij} \leqq m y_j \forall j \in N$ yields another equivalent integer programming formulation. Although the linear programming relaxation of this integer program is relatively easy to solve, the bounds it produces are weak. Consequently, this linear program is not as useful as [(5.1)–(5.5), (5.7)] in a branch-and-bound algorithm. A more general discussion of equivalent integer programs is given by Williams [11].

$$x_{Tj} - y_j \leqq 0 \quad \forall T \in F, j \in N, \tag{5.10}$$

$$\sum_{j \in N} y_j = K, \tag{5.11}$$

$$x_{Tj} \geqq 0 \quad \forall T \in F, j \in N, \tag{5.12}$$

$$y_j \in \{0, 1\} \quad \forall j \in N, \tag{5.13}$$

where we have dropped the constant $r_N$ from the objective function.

From § 2 and, in particular, Proposition 2.1, it follows that the integer programs (5.1)–(5.6) and (5.8)–(5.13) are equivalent and their respective linear programming relaxations are also equivalent.

THEOREM 5.1. *The integer program* (5.1)–(5.6) *and its linear programming relaxation are equivalent, respectively, to*

$$\max \sum_{T \in F} r_T x_T, \tag{5.14}$$

$$x_T - \sum_{j \in T} y_j \leqq 0 \qquad \forall T \in F, \tag{5.15}$$

$$\sum_{j \in N} y_j = K, \tag{5.16}$$

$$0 \leqq x_T \leqq 1 \quad \forall T \in F, \tag{5.17}$$

$$y_j \in \{0, 1\} \quad \forall j \in N, \tag{5.18}$$

*and its linear programming relaxation.*

*Proof.* Let $y$ be a nonnegative $n$-vector and $g(y) = \max_{x_{Tj}} \sum_{T \in F} \sum_{j \in T} r_T x_{Tj}$, subject to (5.9), (5.10) and (5.12). Then $g(y) = \sum_{T \in F} r_T \min (\sum_{j \in T} y_j, 1)$. Similarly, define $h(y) = \max_{x_T} \sum_{T \in F} r_T x_T$, subject to (5.15) and (5.17). Then $h(y) = \sum_{T \in F} r_T \min (\sum_{j \in T} y_j, 1) = g(y)$.

The theorem follows since the constraints on $y$ are the same for the two integer programs, and also for their linear programming relaxations. □

Two more equivalent integer programs are obtained from (5.14)–(5.18) by the following changes in variables: (i) $\pi_T = \sum_{j \in T} y_j - x_T$, and (ii) $\eta_T = r_T x_T$.

If follows from the proof of Theorem 5.1 that the linear programming relaxations of these integer programs are also equivalent to (5.1)–(5.5), (5.7). We will now show that the problems obtained from these transformations have some attractive computational features.

First consider (i). When (i) is substituted into (5.14), (5.15) and (5.17) the constraints $\pi_T \leqq \sum_{j \in T} y_j$ obtained from $x_T \geqq 0$ can be eliminated since they are automatically satisfied by any optimal solution. Now replacing (5.18) by (5.7) and taking the dual we obtain

$$\min \sum_{T \in F} u_T + Kw, \tag{5.19}$$

$$\sum_{T \supseteq \{j\}} u_T + w \geqq \sum_{T \supseteq \{j\}} r_T \quad \forall j \in N, \tag{5.20}$$

$$0 \leqq u_T \leqq r_T \quad \forall T \in F. \tag{5.21}$$

This linear program may have computational advantages because it only has $n$ constraints other than upper bands and at most $m(n-1)+1$ variables.

Now consider (ii): Substituting (ii) into (5.15) yields

$$\eta_T - r_T \sum_{j \in T} y_j \leqq 0 \quad \forall T \in F, \tag{5.22}$$

and from $x_T \leqq 1$ we obtain

$$\eta_T \leqq r_T \quad \forall T \in F. \tag{5.23}$$

In many instances some of the constraints of (5.22) and (5.23) can be aggregated. Suppose that the canonical reduction procedure of § 2 was applied to row $i$ of matrix C. Row $i$ is then replaced by $p \leqq n-1$ rows, indexed by $S_t$, $t = 1, \cdots, p$. The $t$th new row has value $r_{S_t} > 0$ for $j \in S_t$ and value 0, otherwise. As above, a row with all nonzero elements can be suppressed. The critical observation is that we can assume $\varnothing \subset S_p \subset S_{p-1} \subset \cdots \subset S_1 \subset N$.

We are interested in the case $p \geqq 2$. Let $\eta^i = \sum_{t=1}^{p} \eta_{S_t}$ and consider the constraints

$$\eta^i \leqq \sum_{t=1}^{k} r_{S_t} + \sum_{t=k+1}^{p} r_{S_T}\Big(\sum_{j \in S_t} y_j\Big), \qquad k = 0, \cdots, p. \tag{5.24}$$

The $k$th constraint in (5.24) is the sum of (5.23) for the sets $\{S_t\}_{t=1}^{k}$ and (5.22) for the sets $\{S_t\}_{t=k+1}^{p}$. Thus replacing the $2p$ constraints (5.22) and (5.23) and $p$ variables $\{n_{S_t}\}_{t=1}^{p}$ obtained from row $i$ of matrix $C$ by the system (5.24) is certainly a relaxation.

To show that the replacement yields an equivalent program consider any $y \geqq 0$ with $\sum_{j \in N} y_j \geqq 1$. Now let $\bar{k}$, $0 \leqq \bar{k} \leqq p$, be such that $\sum_{j \in S_{\bar{k}+1}} y_j \leqq 1$ and $\sum_{j \in S_{\bar{k}}} y_j \geqq 1$, where $S_0 = N$ and $S_{p+1} = \varnothing$. Note that the constraint of (5.24) with $k = \bar{k}$ gives the maximum value of $\eta^i$. Furthermore, in (5.22) and (5.23) the maximum values of the $\{\eta_{S_t}\}$ are given by

$$\eta_{S_t} = r_{S_t} \sum_{j \in S_t} y_j \quad \forall t \geqq \bar{k} + 1$$

and

$$\eta_{S_t} = r_{S_t} \quad \forall t \leqq \bar{k}.$$

Hence

$$\sum_{t=1}^{p} n_{S_t} = \sum_{t=1}^{\bar{k}} r_{S_t} + \sum_{t=\bar{k}+1}^{p} r_{S_t}\Big(\sum_{j \in S_t} y_j\Big).$$

Finally, suppose that $c_{ik} = \sum_{t=1}^{k} r_{S_t}$, $k = 1, \cdots, p$, and define $c_{i0} = 0$ and $a^+ = \max(0, a)$. Then (5.24) yields

$$n^i \leqq c_{ik} + \sum_{j \in N} (c_{ij} - c_{ik})^+ y_j, \qquad k = 0, \cdots, p. \tag{5.25}$$

Thus, by applying (5.25) to each of the $m$ rows of matrix $C$, we obtain the integer programming formulation

$$\max \sum_{i=1}^{m} \eta^i, \tag{5.26}$$

$$\eta^i \leqq c_{ik} + \sum_{j \in N} (c_{ij} - c_{ik})^+ y_j, \qquad k = 0, \cdots, n, \quad i = 1, \cdots, m \tag{5.27}$$

$$\sum_{j \in N} y_j = K, \tag{5.28}$$

$$y_j \in \{0, 1\} \quad \forall j \in N, \tag{5.29}$$

where $c_{i0} = 0$, $i = 1, \cdots, m$.

THEOREM 5.2. *The integer program* (5.26)–(5.29) *and its linear programming relaxation are equivalent to the integer program* (5.1)–(5.6) *and to its linear programming relaxation, respectively.*

The formulation (5.26)–(5.29) and Theorem 5.2 have been obtained by a very different approach in [9]. The linear programming relaxation of (5.26)–(5.29) may be attractive computationally since it has only $m+n$ variables.

We are doing a computational study of the performance of the simplex method on some of these linear programs. Preliminary results on $10\times 10$ and $30\times 30$ problems from the literature, each solved for four values of $K$, indicate the following:

1. In all cases dual formulations are much easier to solve than the corresponding primal formulations. In several cases the primal required more than 10 times as many pivots as the dual.

2. In all cases the formulation (5.19)–(5.21) dominated the duals of the linear programming relaxations of (5.1)–(5.6) and (5.26)–(5.29); in particular (5.19)–(5.21) required fewer than half of the number of pivots required for the other duals for each problem. Most remarkable is that, although the canonical reduction of the $30\times 30$ matrix yielded over five hundred rows, the largest number of pivots required was 18, which occurred for $K=3$.

**6. Representations of set functions.** The canonical representation of simple location functions by (1.2) is natural because of the direct connection between matrix $C$ and the $\{r_T\}$. Other representations may also provide insight. In this section we consider two other representations and establish their relationships to (1.2). In particular, we characterize the nondecreasing submodular directions and simple location functions for each representation.

Let $w$ be a set function defined on all of the subsets of $S$. Consider the representations

(a) $$w(S)=r_\varnothing+\sum_{T\cap S\neq\varnothing} r_T \quad \forall S\subseteq N,$$

(b) $$w(S)=\sum_{T\subseteq S} c_T \quad \forall S\subseteq N,$$

(c) $$\rho_\varnothing(S)=w(S) \quad \forall S\subseteq N \quad \text{and}$$
$$\rho_{J\cup\{j\}}(S)=\rho_J(S\cup\{j\})-\rho_J(S) \quad \forall S\subset N,\quad J\subset N-S,\quad j\in N-S-J.$$

The $\{c_T\}$ in representation (b) can be thought of as coefficients of a Boolean polynomial (see [7]), and the $\{\rho_J(S)\}$ may be interpreted as "set function derivatives".

THEOREM 6.1. *The expressions for the coefficients* $\{c_T\}_{T\subseteq N}$, $\{r_T\}_{T\subseteq N}$ *and* $\{\rho_J(S)\}_{S\subseteq N, J\subseteq N-S}$ *in terms of* $w$ *and each other are summarized in Table* 1. *All of these coefficients are uniquely defined by* (a), (b) *and* (c).

*Proof.* (See Table 1) Expressions (1) and (2) are definitions. We prove (10) by induction on $|J|$. For $J=\varnothing$, (10) gives $\rho_\varnothing(S)=w(S)$, which is a definition. Assume (10) is true for all $S$ and $J$ with $|J|\leqq p$. Thus by the induction hypothesis

$$\rho_{J\cup\{j\}}(S)=\sum_{T\subseteq J}(-1)^{|J-T|}w(S\cup\{j\}\cup T)-\sum_{T\subseteq J}(-1)^{|J-T|}w(S\cup T)$$

$$=\sum_{T\cup\{j\}\subseteq J\cup\{j\}}(-1)^{|J\cup\{j\}-T\cup\{j\}|}w(S\cup T\cup\{j\})+\sum_{T\subseteq J}(-1)^{|J\cup\{j\}-T|}w(S\cup T)$$

$$=\sum_{T\subseteq J\cup\{j\}}(-1)^{|J\cup\{j\}-T|}w(S\cup T).$$

TABLE 1

$\gamma = w(N) + w(\emptyset)$, $R = S \cup J$, $S \cap J = \emptyset$ and $\delta_\emptyset(T) = \begin{cases} 1 & \text{if } T = \emptyset \\ 0 & \text{otherwise.} \end{cases}$

| | $w$ | $c$ | $r$ | $\rho$ |
|---|---|---|---|---|
| $w(R)$ | ——— | (1) $\sum_{T \subseteq R} c_T$ | (2) $r_\emptyset + \sum_{T \cap R \neq \emptyset} r_T$ | (3) $\sum_{T \subseteq J} \rho_T(S)$ |
| $c_R$ | (4) $\sum_{T \subseteq R} (-1)^{\lvert R-T \rvert} w(T)$ | ——— | (5) $(-1)^{\lvert R \rvert+1} \sum_{T \supseteq R} r_T + \gamma\delta_\emptyset(R)$ | (6) $\sum_{T \subseteq S} (-1)^{\lvert S-T \rvert} \rho_J(T)$ |
| $r_R$ | (7) $\sum_{T \subseteq R} (-1)^{\lvert R-T \rvert+1} w(N-T) + \gamma\delta_\emptyset(R)$ | (8) $(-1)^{\lvert R \rvert+1} \sum_{T \supseteq R} c_T + \gamma_T + \gamma\delta_\emptyset(R)$ | ——— | (9) $\sum_{T \subseteq S} (-1)^{\lvert R-T \rvert+1} \rho_J(N-J-T) + \gamma\delta_\emptyset(R)$ |
| $\rho_J(S)$ | (10) $\sum_{T \subseteq J} (-1)^{\lvert J-T \rvert} w(S \cup T)$ | (11) $\sum_{T \subseteq S} c_{J \cup T}$ | (12) $(-1)^{\lvert J \rvert+1} \sum_{J \subseteq T \subseteq N-S} r_T + \gamma\delta_\emptyset(J)$ | ——— |

To derive (3), (4) and (7) we use the Möbius inversion formula (see [5])

$$g(S)=\sum_{T\subseteq S} f(T) \quad \text{if and only if} \quad f(S)=\sum_{T\subseteq S}(-1)^{|S-T|}g(T).$$

Expressions (3) and (4) are obtained by Mobius inversion of (10) and (1) respectively. Expression (7) is obtained by noting that (2) yields $w(R)=\gamma-\sum_{T\subseteq N-R} r_T$, where $\gamma=w(N)+w(\varnothing)$. Hence $\gamma-w(N-R)=\sum_{T\subseteq R} r_T$. Then by Mobius inversion, $r_R=\sum_{T\subseteq R}(-1)^{|R-T|}(\gamma-w(N-T))=\sum_{T\subseteq R}(-1)^{|R-T|+1}w(N-T)+\gamma\delta_{\varnothing}(R)$, where $\delta_{\varnothing}(R)=1$ if $R=\varnothing$, and 0 otherwise.

We now derive (11) and (12) by substitution. From (1) and (10)

$$\rho_J(S)=\sum_{T\subseteq J}(-1)^{|J-T|}w(S\cup T)=\sum_{T\subseteq J}(-1)^{|J-T|}\sum_{P\subseteq S\cup T}c_P$$

$$=\sum_{P\subseteq S\cup J}c_P\sum_{\substack{T\subseteq J\\ T\supseteq P\cap J}}(-1)^{|J-T|}=\sum_{J\subseteq P\subseteq J\cup S}c_P$$

as

$$\sum_{\substack{T\subseteq J\\ T\supseteq P\cap J}}(-1)^{|J-T|}=0 \quad \text{unless } P\supseteq J.$$

From (2) and (10)

$$\rho_J(S)=\sum_{T\subseteq J}(-1)^{|J-T|}[\gamma-\sum_{P\subseteq N-S-T}r_p]$$

$$=\gamma\delta_{\varnothing}(J)+\sum_{P\subseteq N-S}r_P\sum_{\substack{T\subseteq J\\ T\cap P=\varnothing}}(-1)^{|J-T|+1}$$

$$=\gamma\delta_{\varnothing}(J)+(-1)^{|J|+1}\sum_{J\subseteq P\subseteq N-S}r_P$$

as

$$\sum_{\substack{T\subseteq J\\ T\cap P=\varnothing}}(-1)^{|J-T|+1}=0 \quad \text{unless } P\supseteq J.$$

By Möbius inversion of (11) and (12), we obtain (6) and (9) respectively. Setting $S=\varnothing$ and $J=R$ in (6) and (9) we get $c_R=\rho_R(\varnothing)$ and $r_R=(-1)^{|R|+1}\rho_R(N-R)+\gamma\delta_{\varnothing}(R)$. Thus (5) and (8) are special cases of (11) and (12). □

COROLLARY 6.1. *$w$ is nondecreasing if and only if $\forall S\subset N$, $j\in N-S$*

(a) $p_j(S)\geqq 0$

*or*

(b) $\sum_{T\subseteq S} c_{T\cup\{j\}}\geqq 0$

*or*

(c) $\sum_{T\subseteq N-S-\{j\}} r_{T\cup\{j\}}\geqq 0.$

Note that from (b) and (c) we obtain that $c_{\{j\}}\geqq 0$ and $r_{\{j\}}\geqq 0$, $\forall j\in N$ are necessary conditions for $w$ to be nondecreasing.

COROLLARY 6.2. *$w$ is submodular if and only if $\forall S\subset N$, $\{i,j\}\subset N-S$, $i\neq j$*

(a) $\rho_{\{i,j\}}(S)\leqq 0$

*or*

(b) $\sum_{T\subseteq S} c_{T\cup\{i,j\}}\leqq 0$

*or*

$$\text{(c)} \quad \sum_{T \subseteq N-S-\{i,j\}} r_{T \cup \{i,j\}} \geqq 0.$$

Note that from (b) and (c) we obtain that $c_{\{i,j\}} \leqq 0$ and $r_{\{i,j\}} \geqq 0$, $\forall \{i, j\} \subseteq N$, $i \neq j$ are necessary conditions for $w$ to be submodular.

Statements (a) in Corollaries 6.1 and 6.2 are sign conditions on the first and second "derivatives," respectively, of $w$. The next corollary characterizes the simple location functions as these set functions $w$ whose "derivatives" alternate in sign.

COROLLARY 6.3. *$w$ is a simple location function if and only if*

$$\text{(a)} \quad (-1)^{|J|+1} \rho_J(S) \geqq 0 \quad \forall \varnothing \subset J \subseteq N,\ S \subseteq N-J,$$

$$\text{(b)} \quad (-1)^{|T|+1} c_T \geqq 0 \quad \forall \varnothing \subset T \subseteq N,$$

$$\text{(c)} \quad r_T \geqq 0 \quad \forall \varnothing \cup T \subseteq N.$$

*Proof.* Given a simple location function $v$ described by (1.1), we see that adding a constant to a row of matrix $C$ and subtracting the same constant from $v(\varnothing)$ yields a new simple location function $v'$ with $v'(S) = v(S)$, $\forall S \neq \varnothing$. Now since $r_N = \sum_{i \in I} \min_{j \in N} c_{ij}$, we see that $r_N \geqq 0$ for a suitable choice of $v(\varnothing) = r_\varnothing$. Therefore (c) follows from Corollary 2.1.

Now we show (a)$\Leftrightarrow$(c). From (12), $(-1)^{|T|+1} \rho_T(N-T) = r_T \geqq 0$, $\forall \varnothing \subset T \subseteq N$. Conversely if $r_T \geqq 0, \forall \varnothing \subset T \subseteq N$, then (12) implies $(-1)^{|J|+1} \rho_J(S) \geqq 0$ whenever $J \neq \varnothing$. The proof of (a)$\Leftrightarrow$(b) follows similarly from (11). □

**Acknowledgment.** We wish to thank the referees for their helpful comments.

## REFERENCES

[1] G. CORNUEJOLS, M. L. FISHER AND G. L. NEMHAUSER, *Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms*, Management Sci., 23 (1977), pp. 789–810.

[2] G. CORNUEJOLS, G. L. NEMHAUSER AND L. A. WOLSEY, *Worst-case and probabilistic analysis of a location problem*, Operations Res. to appear.

[3] M. L. FISHER, G. L. NEMHAUSER AND L. A. WOLSEY, *An analysis of approximations for maximizing submodular set functions-II*, Math. Programming Stud. 8 (1978), pp. 73–87.

[4] M. GUIGNARD AND K. SPIELBERG, *Algorithms for exploiting the structure of the simple plant location problem*, Annals Discrete Math., 1 (1977), pp. 247–272.

[5] M. HALL, JR., *Combinatorial Theory*, John Wiley and Sons, NY, 1967.

[6] D. M. KREPS, *A Representation Theorem for Preference for Flexibility*, Graduate School of Business, Stanford Univ. Stanford, CA, 1978.

[7] G. L. NEMHAUSER, L. A. WOLSEY AND M. L. FISHER, *An analysis of approximations for maximizing submodular set functions-I*, Mathematical Programming, 14 (1978), pp. 265–294.

[8] G. L. NEMHAUSER AND L. A. WOLSEY, *Best algorithms for approximating the maximum of a submodular set function*, Math. Oper. Res. 3 (1978), pp. 177–188.

[9] G. L. NEMHAUSER AND L. A. WOLSEY, *Maximizing submodular set functions: Formulation and analysis of algorithms*, Tech. Rep. 398, School of Operations Research and Industrial Engineering Cornell University, Ithaca, NY, 1978.

[10] L. SCHRAGE, *Implicit representation of variable upper bounds in linear programming*, Math. Programming Stud., 4 (1975), pp. 118–132.

[11] H. P. WILLIAMS, *Experiments in the formulation of integer programming problems*, Ibid., 2 (1974), pp. 180-197.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0004 $01.00/0

# THE CONDITION OF A FINITE MARKOV CHAIN AND PERTURBATION BOUNDS FOR THE LIMITING PROBABILITIES*

CARL D. MEYER, JR.†

**Abstract.** Let $\mathbf{T}$ denote the transition matrix of an ergodic chain, $\mathscr{C}$, and let $\mathbf{A}=\mathbf{I}-\mathbf{T}$. Let $\mathbf{E}$ be a perturbation matrix such that $\tilde{\mathbf{T}}=\mathbf{T}-\mathbf{E}$ is also the transition matrix of an ergodic chain, $\tilde{\mathscr{C}}$. Let $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ denote the limiting probability (row) vectors for $\mathscr{C}$ and $\tilde{\mathscr{C}}$. The purpose of this paper is to exhibit inequalities bounding the relative error $\|\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}}\|/\|\boldsymbol{\omega}\|$ by a very simple function of $\mathbf{E}$ and $\mathbf{A}$. Furthermore, the inequality will be shown to be the best one which is possible. This bound can be significant in the numerical determination of the limiting probabilities for an ergodic chain.

In addition to presenting a sharp bound for $\|\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}}\|/\|\boldsymbol{\omega}\|$, we will derive an explicit expression for $\tilde{\boldsymbol{\omega}}$, in which $\tilde{\boldsymbol{\omega}}$ is given as a function of $\mathbf{E}$, $\mathbf{A}$, $\boldsymbol{\omega}$ and some other related terms.

**1. Introduction.** Let $\mathbf{T}$ denote the transition matrix of an ergodic chain, $\mathscr{C}$, and let $\mathbf{A}=\mathbf{I}-\mathbf{T}$. (The terminology and notation will be that used in [5] and [6].) Let $\mathbf{E}$ be a perturbation matrix such that $\tilde{\mathbf{T}}=\mathbf{T}-\mathbf{E}$ is also the transition matrix of an ergodic chain, $\tilde{\mathscr{C}}$. Let $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ denote the limiting probability (row) vectors for $\mathscr{C}$ and $\tilde{\mathscr{C}}$. The purpose of this paper is to exhibit inequalities bounding the relative error $\|\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}}\|/\|\boldsymbol{\omega}\|$ by a very simple function of $\mathbf{E}$ and $\mathbf{A}$. Furthermore, the inequality will be shown to be the best one which is possible. This bound can be significant in the numerical determination of the limiting probabilities for an ergodic chain.

In addition to presenting a sharp bound for $\|\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}}\|/\|\boldsymbol{\omega}\|$, an explicit expression for $\tilde{\boldsymbol{\omega}}$ will be derived in which $\tilde{\boldsymbol{\omega}}$ is given as a function of $\mathbf{E}$, $\mathbf{A}$, $\boldsymbol{\omega}$ and some other related terms.

The approach taken in this paper differs from the traditional methods of past authors in that group properties of the matrix $\mathbf{A}$ are used to produce the desired results, whereas previous results have relied upon the so-called "fundamental matrix" given in [5]; (see [9]). Examples will be given which show that the use of the group properties produces results superior to those which can be produced using the traditional theories.

**2. Group properties.** The fundamental fact on which the analysis of this paper is based is the following.

THEOREM 2.1. *If* $\mathbf{A}=\mathbf{I}-\mathbf{T}$ *where* $\mathbf{T}$ *is any row stochastic matrix, then* $\mathbf{A}$ *belongs to a multiplicative matrix.*

A proof of this is given in [2] and [6]. It also follows from well-known results found in [4] and [8].

Since $\mathbf{A}$ belongs to some multiplicative group, $\mathscr{G}$, $\mathbf{A}$ must possess an inverse in $\mathscr{G}$. This matrix is called the *group inverse* of $\mathbf{A}$ and is denoted by $\mathbf{A}^{\#}$. The identity in $\mathscr{G}$ is $\mathbf{P}=\mathbf{A}\mathbf{A}^{\#}$, the projector whose range is $R(\mathbf{A})$ and whose nullspace is $N(\mathbf{A})$.

As is shown in [6] and [2], almost all of the important information concerning an ergodic chain is available in terms of the entries of $\mathbf{A}^{\#}$. In particular, the limiting matrix, $\mathbf{W}$, for a chain with transition matrix $\mathbf{T}$ is given by

$$\mathbf{W}=\lim_{n\to\infty}\frac{\mathbf{I}+\mathbf{T}+\mathbf{T}^2+\cdots+\mathbf{T}^{n-1}}{n}=\mathbf{I}-\mathbf{A}\mathbf{A}^{\#}, \qquad \text{(see [6] or [2])}. \tag{2.1}$$

* Received by the editors June 14, 1979, and in final form January 7, 1980.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina, 27650. This work was supported in part by the National Aeronautics and Space Administration under grant NSG-1532, and by the National Science Foundation under grant MCS76-11989.

As pointed out in [6], the computation of $\mathbf{A}^{\#}$ is not unduly complicated. Indeed, computing $\mathbf{A}^{\#}$ is less of a chore than calculating the fundamental matrix. Further properties of $\mathbf{A}^{\#}$ are presented in [2].

**3. A perturbation formula for $(\mathbf{A}+\boldsymbol{E})^{\#}$.** Suppose $\mathbf{T}$ and $\tilde{\mathbf{T}}$ are transition matrices for ergodic chains $\mathscr{C}$ and $\tilde{\mathscr{C}}$, respectively, where $\tilde{\mathbf{T}}=\mathbf{T}-\mathbf{E}$ so that $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{E}$. In order to analyze $\tilde{\mathscr{C}}$, it suffices to analyze $\tilde{\mathbf{A}}^{\#}$. The purpose of this section is to provide an expression for $(\mathbf{A}+\mathbf{E})^{\#}$ which will hold for all possible values of $\mathbf{E}$. Notice that $\mathbf{E}$ cannot be arbitrary. Since $\tilde{\mathbf{T}}$ must be a stochastic matrix, the elements, $e_{ij}$, of $\mathbf{E}$ are constrained so that $|e_{ij}|\leqq 1$. There are, of course, other additional restrictions.

If $\mathbf{j}=[1, 1, 1, \cdots, 1]^{\mathrm{T}}$, then $\mathbf{Aj}=\mathbf{0}$ and $(\mathbf{A}+\mathbf{E})\mathbf{j}=\mathbf{0}$ so that $\mathbf{Ej}=\mathbf{0}$. If $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ denote the limiting probability (row) vectors for $\mathscr{C}$ and $\tilde{\mathscr{C}}$, respectively, then (2.1) implies that

$$\mathbf{E}(\mathbf{I}-\mathbf{AA}^{\#})=\mathbf{E}(\mathbf{j}\boldsymbol{\omega})=\mathbf{0},$$

so that

$$\mathbf{EAA}^{\#}=\mathbf{E}, \qquad \text{(i.e., Row Sp}(\mathbf{E})\subseteq\text{Row Sp}(\mathbf{A})). \tag{3.1}$$

Since $\mathbf{A}_{n\times n}$ belongs to a matrix group, there exist nonsingular matrices $\mathbf{P}$ and $\mathbf{C}_{(n-1)\times(n-1)}$ such that

$$\mathbf{A}=\mathbf{P}\begin{bmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}\mathbf{P}^{-1}, \qquad \mathbf{A}^{\#}=\mathbf{P}\begin{bmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}\mathbf{P}^{-1}, \quad \text{and} \quad \mathbf{I}-\mathbf{AA}^{\#}=\mathbf{P}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & 1\end{bmatrix}\mathbf{P}^{-1}. \tag{3.2}$$

(These statements are evident, but the reader may wish to consult [2].) Write $\mathbf{E}$ in the form

$$\mathbf{E}=\mathbf{P}\begin{bmatrix}\mathbf{E}_1 & \mathbf{E}_3\\ \mathbf{E}_2 & \mathbf{E}_4\end{bmatrix}\mathbf{P}^{-1}, \tag{3.3}$$

where $\mathbf{E}_1$ is $(n-1)\times(n-1)$. The fact that $\mathbf{EAA}^{\#}=\mathbf{E}$ implies that $\mathbf{E}_3=\mathbf{0}$ and $\mathbf{E}_4=0$, so that

$$\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{E}=\mathbf{P}\begin{bmatrix}\mathbf{C}+\mathbf{E}_1 & \mathbf{0}\\ \mathbf{E}_2 & 0\end{bmatrix}\mathbf{P}^{-1}. \tag{3.4}$$

Since $\tilde{\mathscr{C}}$ is again an ergodic chain, it must be the case that the limiting matrix, $\tilde{\mathbf{W}}$, must be a rank 1 matrix. By virtue of (2.1), it follows that rank $(\mathbf{I}-\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\#})=1$. By using the formula

$$\begin{bmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{Y} & \mathbf{0}\end{bmatrix}^{\#}=\begin{bmatrix}\mathbf{X}^{\#} & \mathbf{0}\\ \mathbf{YX}^{2\#} & \mathbf{0}\end{bmatrix}, \qquad \text{(found in [2] or [7])}, \tag{3.5}$$

it is easy to see from (3.4) that

$$\mathbf{I}-\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\#}=\mathbf{P}\begin{bmatrix}\mathbf{I}-(\mathbf{C}+\mathbf{E}_1)(\mathbf{C}+\mathbf{E}_1)^{\#} & \mathbf{0}\\ -\mathbf{E}_2(\mathbf{C}+\mathbf{E}_1)^{\#} & 1\end{bmatrix}\mathbf{P}^{-1}.$$

The fact that rank $(\mathbf{I}-\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\#})=1$ now implies that $\mathbf{I}-(\mathbf{C}+\mathbf{E}_1)(\mathbf{C}+\mathbf{E}_1)^{\#}=\mathbf{0}$. That is, $\mathbf{C}+\mathbf{E}_1$ is a nonsingular matrix. Since $\mathbf{C}+\mathbf{E}_1=(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})\mathbf{C}$, it follows that $(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})$ is nonsingular, so that

$$\mathbf{I}+\mathbf{EA}^{\#}=\mathbf{P}\begin{bmatrix}\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{E}_2\mathbf{C}^{-1} & 1\end{bmatrix}\mathbf{P}^{-1},$$

is also nonsingular and

$$(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}=\mathbf{P}\left[\begin{array}{c|c}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1} & \mathbf{0}\\ \hline -\mathbf{E}_2\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1} & 1\end{array}\right]\mathbf{P}^{-1}. \tag{3.6}$$

Now write the expression for $(\mathbf{A}+\mathbf{E})^{\#}$. Using (3.4), (3.5), together with the fact that $(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})$ is nonsingular, yields

$$(\mathbf{A}+\mathbf{E})^{\#}=\mathbf{P}\left[\begin{array}{c|c}\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1} & \mathbf{0}\\ \hline \mathbf{E}_2\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1}\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1} & 0\end{array}\right]\mathbf{P}^{-1}. \tag{3.7}$$

From (3.2) and (3.6) it is easy to see that

$$\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}=\mathbf{P}\left[\begin{array}{c|c}\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1}) & \mathbf{0}\\ \hline \mathbf{0} & 0\end{array}\right]\mathbf{P}^{-1},$$

$$(\mathbf{I}-\mathbf{A}\mathbf{A}^{\#})(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}=\mathbf{P}\left[\begin{array}{c|c}\mathbf{0} & \mathbf{0}\\ \hline -\mathbf{E}_2\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1} & 1\end{array}\right]\mathbf{P}^{-1}$$

and

$$(\mathbf{I}-\mathbf{A}\mathbf{A}^{\#})(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}$$
$$=\mathbf{P}\left[\begin{array}{c|c}\mathbf{0} & \mathbf{0}\\ \hline -\mathbf{E}_2\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1}\mathbf{C}^{-1}(\mathbf{I}+\mathbf{E}_1\mathbf{C}^{-1})^{-1} & 0\end{array}\right]\mathbf{P}^{-1}.$$

Thus (3.7) becomes

$$(\mathbf{A}+\mathbf{E})^{\#}=\mathbf{A}^{\#}-(\mathbf{I}-\mathbf{A}\mathbf{A}^{\#})(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}. \tag{3.8}$$

By using the identity $(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}=\mathbf{I}-\mathbf{E}\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}$, together with (2.1), one arrives at the following result.

THEOREM 3.1. *Let $\mathscr{C}$ be an ergodic chain with transition matrix $\mathbf{T}$ and limiting matrix $\mathbf{W}$ and let $\tilde{\mathscr{C}}$ be an ergodic chain with transition matrix $\tilde{\mathbf{T}}=\mathbf{T}-\mathbf{E}$. If $\mathbf{A}=\mathbf{I}-\mathbf{T}$, then $\mathbf{I}+\mathbf{E}\mathbf{A}^{\#}$ is nonsingular and*

$$(\mathbf{A}+\mathbf{E})^{\#}=\mathbf{A}^{\#}-\mathbf{A}^{\#}\mathbf{E}\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}-\mathbf{W}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}.$$

It is clear that this theorem guarantees that for the situation under question,

$$\lim_{\mathbf{E}\to\mathbf{0}}(\mathbf{A}+\mathbf{E})^{\#}=\mathbf{A}^{\#},$$

so that the following corollary is obtained.

COROLLARY 3.1. *For the situation of Theorem* 3.1, *the elements of $\mathbf{A}^{\#}$ depend continuously on the elements of $\mathbf{A}$.*

This result can also be proven using the information in [2] or [3].

Now that an explicit representation for $(\mathbf{A}+\mathbf{E})^{\#}$ is known, one can obtain almost all of the important information regarding $\tilde{\mathscr{C}}$ through the results of [6]. However, the purpose here is to concentrate on the problem of obtaining a perturbation formula and bounds for the limiting probabilities, because it is these quantities which lie at the heart of any analysis of the chain.

**4. A perturbation formula and perturbation bounds for the limiting probabilities.** If $\mathbf{W}$ and $\tilde{\mathbf{W}}$ are the limiting matrices for ergodic chains $\mathscr{C}$ and $\tilde{\mathscr{C}}$, respectively, then using Theorem 3.1 together with (2.1) yields an explicit expression for $\tilde{\mathbf{W}}$. One has the following result.

THEOREM 4.1. *If $\mathscr{C}$ and $\tilde{\mathscr{C}}$ are ergodic chains with transition matrices* $\mathbf{T}$ *and* $\tilde{\mathbf{T}} = \mathbf{T} - \mathbf{E}$ *and limiting matrices* $\mathbf{W}$ and $\tilde{\mathbf{W}}$, *respectively, then*

$$\tilde{\mathbf{W}} = \mathbf{W}(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1} = \mathbf{W} - \mathbf{W}\mathbf{E}\mathbf{A}^{\#}(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1}, \quad \textit{where } \mathbf{A} = \mathbf{I} - \mathbf{T}.$$

In passing, it is pointed out that as a corollary one obtains $\lim_{\mathbf{E}\to\mathbf{0}} \tilde{\mathbf{W}} = \mathbf{W}$, which is of course the well known result stating that the limiting probabilities are continuous functions of the elements of $\mathbf{T}$. By making use of (3.1) and (2.1) another important corollary of Theorem 4.1 is obtained. It is the one which reveals the structure necessary in order for the limiting probabilities to remain invariant under a perturbation.

COROLLARY 4.1. *For ergodic chains $\mathscr{C}$ and $\tilde{\mathscr{C}}$, it is the case that* $\mathbf{W} = \tilde{\mathbf{W}}$ *if and only if* $R(\mathbf{E}) \subseteq R(\mathbf{A})$. (*i.e., the limiting probabilities are unaltered if and only if the columns of* $\mathbf{E}$ *are linear combinations of columns of* $\mathbf{A}$.)

Consider now the problem of bounding the relative error term $\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|/\|\boldsymbol{\omega}\|$, where $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ are the limiting probability vectors for $\mathscr{C}$ and $\tilde{\mathscr{C}}$, respectively. Since every row of $\mathbf{W}$ is equal to $\boldsymbol{\omega}$ and every row of $\tilde{\mathbf{W}}$ is equal to $\tilde{\boldsymbol{\omega}}$, Theorem 4.1 yields,

$$\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} = \tilde{\boldsymbol{\omega}}\mathbf{E}\mathbf{A}^{\#} \tag{4.1}$$

and

$$\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} = \boldsymbol{\omega}\mathbf{E}\mathbf{A}^{\#}(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1}. \tag{4.2}$$

For the vector 1-norm ($\|\mathbf{x}\|_1 = \sum_j |x_j|$), the induced matrix norm is $\|\mathbf{A}\|_1 = \max_{\|\mathbf{x}\|_1 = 1} \|\mathbf{x}\mathbf{A}\|_1 = \max_i \sum_j |a_{ij}|$, because one is dealing with row vectors and left hand multiplication. A trivial observation is that the relative error in $\boldsymbol{\omega}$ for the 1-norm is always bounded by 2. That is,

$$\frac{\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_1}{\|\boldsymbol{\omega}\|_1} = \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_1 < 2,$$

and $\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_1$ can be made to be arbitrarily close to 2 with particular choices of $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$. However, this does not take into account the relative size of $\|\mathbf{E}\|_1$. The expression in (4.1) can provide a more useful bound in the case of the 1-norm. Using (4.1) to bound the relative error in $\boldsymbol{\omega}$ provides an additional desirable feature, namely, that the bound is obtainable without having to impose any additional hypothesis on the magnitude of the elements of $\mathbf{E}$.

The above remarks are summarized in the following.

THEOREM 4.2. *For ergodic chains $\mathscr{C}$ and $\tilde{\mathscr{C}}$ with transition matrices* $\mathbf{T}$ *and* $\tilde{\mathbf{T}} = \mathbf{T} - \mathbf{E}$ *and limiting probability vectors* $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$, *the relative error in* $\boldsymbol{\omega}$ *for the* 1*-norm is*

$$\frac{\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_1}{\|\boldsymbol{\omega}\|_1} = \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_1 \leqq \|\mathbf{E}\mathbf{A}^{\#}\|_1 \leqq \frac{\|\mathbf{E}\|_1}{\|\mathbf{A}\|_1} \kappa_1(\mathscr{C}),$$

*where* $\mathbf{A} = \mathbf{I} - \mathbf{T}$ *and* $\kappa_1(\mathscr{C}) = \|\mathbf{A}\|_1 \|\mathbf{A}^{\#}\|_1$.

The 1-norm may not be the most desirable choice of norms. It seems that the $\infty$-norm is a more natural choice of norm when investigating the sensitivity of the limiting probabilities to perturbations in the transition probabilities.

It is worth completing the statement on norms by noting that for any two probability vectors, $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$, the following relations always hold for an $n$-state chain.

$$\|\boldsymbol{\omega}\|_1 = 1 \quad \text{and} \quad \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_1 \leqq 2.$$

$$\frac{1}{\sqrt{n}} \leqq \|\boldsymbol{\omega}\|_2 \leqq 1 \quad \text{and} \quad \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_2 \leqq \sqrt{2}, \quad \text{so that} \quad \frac{\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_2}{\|\boldsymbol{\omega}\|_2} \leqq \sqrt{2n}.$$

$$\frac{1}{n} \leqq \|\boldsymbol{\omega}\|_\infty \leqq 1 \quad \text{and} \quad \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_\infty \leqq 1, \quad \text{so that} \quad \frac{\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_\infty}{\|\boldsymbol{\omega}\|_\infty} \leqq n.$$

Consider now an arbitrary vector norm and a compatible matrix norm such that $\|\mathbf{I}\| = 1$. Take the norm of both sides of (4.2) to obtain

$$\frac{\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|}{\|\boldsymbol{\omega}\|} < \|\mathbf{E}\mathbf{A}^{\#}\| \, \|(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1}\|.$$

If $\|\mathbf{E}\mathbf{A}^{\#}\| < 1$, then,

$$\|(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1}\| \leqq \frac{1}{1 - \|\mathbf{E}\mathbf{A}^{\#}\|}$$

and the inequality takes a familiar form which is given below.

THEOREM 4.3. *Let $\mathscr{C}$ and $\tilde{\mathscr{C}}$ be ergodic chains with transition matrices $\mathbf{T}$ and $\tilde{\mathbf{T}}$, respectively, where $\tilde{\mathbf{T}} = \mathbf{T} - \mathbf{E}$. Let $\mathbf{d} = \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}$ where $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ are the limiting probability vectors for $\mathscr{C}$ and $\tilde{\mathscr{C}}$, respectively, and let $\mathbf{A} = \mathbf{I} - \mathbf{T}$. If $\|\mathbf{E}\mathbf{A}^{\#}\| < 1$, then*

$$\frac{\|\mathbf{d}\|}{\|\boldsymbol{\omega}\|} \leqq \frac{\|\mathbf{E}\mathbf{A}^{\#}\|}{1 - \|\mathbf{E}\mathbf{A}^{\#}\|}.$$

*If $\|\mathbf{E}\| \, \|\mathbf{A}^{\#}\| \leqq 1$, then*

(4.3)
$$\frac{\|\mathbf{d}\|}{\|\boldsymbol{\omega}\|} \leqq \frac{\dfrac{\|\mathbf{E}\|}{\|\mathbf{A}\|} \kappa(\mathscr{C})}{1 - \dfrac{\|\mathbf{E}\|}{\|\mathbf{A}\|} \kappa(\mathscr{C})},$$

*where $\kappa(\mathscr{C}) = \|\mathbf{A}\| \, \|\mathbf{A}^{\#}\|$. Moreover, there are nontrivial cases where equality is actually attained in each of the above.*

Note that (3.1) guarantees that $\|\mathbf{E}\|/\|\mathbf{A}\| \leqq \|\mathbf{E}\mathbf{A}^{\#}\|$, which is less than 1, by hypothesis.

The term $\|\mathbf{d}\|/\|\boldsymbol{\omega}\|$ is the relative error in $\boldsymbol{\omega}$ while $\|\mathbf{E}\|/\|\mathbf{A}\|$ is the relative error in $\mathbf{A}$. This inequality is of exactly the same form as the familiar inequality obtained when analyzing a perturbed nonsingular linear system of equations. The only difference is the term $\kappa(\mathscr{C})$. The fact that the analysis of any ergodic chain revolves about the limiting probabilities, together with the appearance of $\kappa(\mathscr{C})$ in Theorems 4.2 and 4.3, motivates one to make the following definition.

DEFINITION. Let $\mathscr{C}$ be an ergodic chain whose transition matrix is $\mathbf{T}$, and let $\mathbf{A} = \mathbf{I} - \mathbf{T}$. The *condition of the chain* $\mathscr{C}$ is defined to be the number $\kappa(\mathscr{C}) = \|\mathbf{A}\| \, \|\mathbf{A}^{\#}\|$.

Clearly, if the condition of the chain is relatively small, then the limiting probabilities will be relatively insensitive to small changes in the transition probabilities. If the condition of the chain is relatively large, then the limiting probabilities may or may not be sensitive. Although the bound in (4.3) can sometimes be pessimistic, it is important to point out that there are nontrivial cases where equality is actually attained. Examples are given in following sections.

As a final observation, note that since $\mathbf{A}\mathbf{A}^{\#} = \mathbf{I} - \mathbf{W}$, where $\mathbf{W}_{n\times n}$ is the limiting matrix, one has $\|\mathbf{A}\mathbf{A}^{\#}\|_1 = 2 - 2\min \omega_i$, $\|\mathbf{A}\mathbf{A}^{\#}\|_2 \geqq 1$, and $\|\mathbf{A}\mathbf{A}^{\#}\|_\infty = 1 + (n-2)\max \omega_i$, so that

$$\kappa_1(\mathscr{C}) \geqq 2 - 2\min \omega_i \geqq 2 - \frac{2}{n},$$

$$\kappa_2(\mathscr{C}) \geqq 1,$$

$$\kappa_\infty(\mathscr{C}) \geqq 1 + (n-2)\max \omega_i \geqq 2 - \frac{2}{n}.$$

A special case which is of frequent interest is that in which the perturbation affects only a single state. That is, only the probabilities for leaving (or entering) the $i$th state are perturbed. The question is, "How does this affect $\omega_i$ and perhaps the rest of $\boldsymbol{\omega}$?"

In this case, the $i$th row of $\mathbf{T}$, denoted by $\mathbf{t}_i$, is perturbed so as to produce $\tilde{\mathbf{t}}_i$, the $i$th row of $\tilde{\mathbf{T}}$. If $\mathbf{u}_i = [0, 0, \cdots, 0, 1, 0, \cdots, 0]^{\mathbf{T}}$ is the $i$th unit vector, then $\mathbf{E}$ is the rank 1 matrix $\mathbf{E} = \mathbf{u}_i(\mathbf{t}_i - \tilde{\mathbf{t}}_i)$. Equation (4.1) degenerates to

$$\mathbf{d} = \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} = \omega_i \mathbf{e}_i \mathbf{A}^{\#}(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1},$$

where $\mathbf{e}_i = \mathbf{t}_i - \tilde{\mathbf{t}}_i$. Since $\mathbf{E} = \mathbf{u}_i\mathbf{e}_i$, one can write

$$(\mathbf{I} + \mathbf{E}\mathbf{A}^{\#})^{-1} = (\mathbf{I} + \mathbf{u}_i\mathbf{e}_i\mathbf{A}^{\#})^{-1} = \mathbf{I} - \frac{\mathbf{u}_i\mathbf{e}_i\mathbf{A}^{\#}}{1 + \mathbf{e}_i\mathbf{A}^{\#}\mathbf{u}_i},$$

so that (4.2) reduces to the following.

COROLLARY 4.2. *If $\mathscr{C}$ is an ergodic chain and the transition probabilities for leaving the ith state are perturbed so as to form an ergodic chain $\tilde{\mathscr{C}}$, then*

$$\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} = \omega_i \left[\frac{\mathbf{e}_i\mathbf{A}^{\#}}{1 + \mathbf{e}_i\mathbf{A}^{\#}\mathbf{u}_i}\right], \tag{4.4}$$

*where $\boldsymbol{\omega}$, $\tilde{\boldsymbol{\omega}}$, $\mathbf{e}_i$, $\mathbf{A}^{\#}$, and $\mathbf{u}_i$ are as described earlier.*

*In particular,*

$$\frac{\omega_i - \tilde{\omega}_i}{\omega_i} = \frac{\sigma}{1 + \sigma},$$

*where $\sigma = \mathbf{e}_i\mathbf{A}^{\#}\mathbf{u}_i$ and*

$$\frac{\omega_k - \tilde{\omega}_k}{\omega_k} = \frac{\omega_i}{\omega_k}\left[\frac{\mathbf{e}_i\mathbf{A}^{\#}\mathbf{u}_k}{1 + \mathbf{e}_i\mathbf{A}^{\#}\mathbf{u}_i}\right].$$

*If $\|\mathbf{e}_i\mathbf{A}^{\#}\| < 1$, then*

$$\left|\frac{\omega_i - \tilde{\omega}_i}{\omega_i}\right| \leqq \frac{\|\mathbf{e}_i\mathbf{A}^{\#}\|}{1 - \|\mathbf{e}_i\mathbf{A}^{\#}\|} \quad \text{and} \quad \left|\frac{\omega_k - \tilde{\omega}_k}{\omega_k}\right| \leqq \frac{\omega_i}{\omega_k}\left[\frac{\|\mathbf{e}_i\mathbf{A}^{\#}\|}{1 - \|\mathbf{e}_i\mathbf{A}^{\#}\|}\right].$$

*If $\|\mathbf{e}_i\|\,\|\mathbf{A}^{\#}\| < 1$, then*

$$\left|\frac{\omega_i - \tilde{\omega}_i}{\omega_i}\right| \leqq \frac{\|\mathbf{e}_i\|/\|\mathbf{A}\|\kappa(\mathscr{C})}{1 - \|\mathbf{e}_i\|/\|\mathbf{A}\|\kappa(\mathscr{C})} \quad \text{and} \quad \left|\frac{\omega_k - \tilde{\omega}_k}{\omega_k}\right| \leqq \frac{\omega_i}{\omega_k}\left[\frac{\|\mathbf{e}_i\|/\|\mathbf{A}\|\kappa(\mathscr{C})}{1 - \|\mathbf{e}_i\|/\|\mathbf{A}\|\kappa(\mathscr{C})}\right].$$

**5. Example.** Below, a general example is constructed to show that equality in (4.3) can be attained for the $\infty$-norm. Note that the fact that row vectors, rather than column vectors, are involved means that the $\infty$-matrix norm is given by $\|\mathbf{A}\|_\infty = \max_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{x}\mathbf{A}\|_\infty = \max_j \sum_i |a_{ij}|$.

Consider the regular chain whose transition matrix is the symmetric circulant,

$$\mathbf{T}=\frac{1}{4n}\begin{bmatrix}3&1&3&1&\cdots&3&1\\1&3&1&3&\cdots&1&3\\3&1&3&1&\cdots&3&1\\\vdots&&&&&&\\1&3&1&3&\cdots&1&3\end{bmatrix}_{2n\times 2n}$$

Since $\mathbf{T}$ is symmetric, the limiting probability vector is

$$\boldsymbol{\omega}=\frac{1}{2n}[1,1,\cdots,1].$$

It is easy to check, that $\mathbf{A}^{\#}$ is given by the symmetric circulant,

$$\mathbf{A}^{\#}=\frac{1}{n}\begin{bmatrix}n&-1&0&-1&0&\cdots&-1&0&-1\\-1&n&-1&0&-1&\cdots&0&-1&0\\0&-1&n&-1&0&\cdots&-1&0&-1\\\vdots&&&&&&&&\\-1&0&-1&0&-1&\cdots&-1&0&n\end{bmatrix},$$

by verifying that $\mathbf{A}\mathbf{A}^{\#}\mathbf{A}=\mathbf{A}$, $\mathbf{A}^{\#}\mathbf{A}\mathbf{A}^{\#}=\mathbf{A}^{\#}$, and $\mathbf{A}\mathbf{A}^{\#}=\mathbf{A}^{\#}\mathbf{A}$. (These three conditions suffice to define $\mathbf{A}^{\#}$; see [2] or [6].) Note that $\|\mathbf{A}^{\#}\|_\infty=2$. Let the perturbation be

$$\mathbf{E}=\begin{bmatrix}-\varepsilon&\varepsilon&-\varepsilon&\varepsilon&\cdots&\varepsilon&-\varepsilon&\varepsilon\\0&0&0&0&\cdots&0&0&0\\\vdots&&&&&&&\\0&0&0&0&\cdots&0&0&0\end{bmatrix},$$

where $0<\varepsilon<1/4n$. Then $\tilde{\mathbf{T}}=\mathbf{T}-\mathbf{E}$ is the transition matrix for a regular chain, and $\|\mathbf{E}\|_\infty=\varepsilon$ so that $\|\mathbf{E}\|_\infty\|\mathbf{A}^{\#}\|_\infty=2\varepsilon<1$. From (4.2),

$$\mathbf{d}=\boldsymbol{\omega}-\tilde{\boldsymbol{\omega}}=\boldsymbol{\omega}\mathbf{E}\mathbf{A}^{\#}(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}.$$

It is easy to see that $\mathbf{E}\mathbf{A}^{\#}=2\varepsilon\mathbf{u}_1\mathbf{v}$, where $\mathbf{u}_1=[1,0,0,\cdots,0]^{\mathrm{T}}$ and $\mathbf{v}=[-1,1,-1,1,\cdots,-1,1]$, so that $\boldsymbol{\omega}\mathbf{E}\mathbf{A}^{\#}=2\varepsilon\omega_1\mathbf{v}=(\varepsilon/n)\mathbf{v}$ and

$$(\mathbf{I}+\mathbf{E}\mathbf{A}^{\#})^{-1}=(\mathbf{I}+2\varepsilon\mathbf{u}_1\mathbf{v})^{-1}=\mathbf{I}-\frac{2\varepsilon}{1-2\varepsilon}\mathbf{u}_1\mathbf{v}.$$

Thus,

$$\mathbf{d}=\frac{\varepsilon}{n(1-2\varepsilon)}\mathbf{v},$$

and

$$\frac{\|\mathbf{d}\|_\infty}{\|\boldsymbol{\omega}\|_\infty}=\frac{2\varepsilon}{1-2\varepsilon}=\frac{\|\mathbf{E}\mathbf{A}^{\#}\|_\infty}{1-\|\mathbf{E}\mathbf{A}^{\#}\|_\infty}=\frac{\|\mathbf{E}\|_\infty\|\mathbf{A}^{\#}\|_\infty}{1-\|\mathbf{E}\|_\infty\|\mathbf{A}^{\#}\|_\infty}=\frac{\|\mathbf{E}\|_\infty/\|\mathbf{A}\|_\infty\kappa_\infty(\mathscr{C})}{1-\|\mathbf{E}\|_\infty/\|\mathbf{A}\|_\infty\kappa_\infty(\mathscr{C})}.$$

One can also construct examples so that equality in (4.3) will hold for other norms.

**6. Why not treat this strictly as an eigenvector problem?** In principle, the problem is an eigenvector problem. That is, one is analyzing the normalized left hand eigenvector associated with the eigenvalue $\lambda_{\mathbf{T}}=1$ for a row stochastic matrix $\mathbf{T}$, or equivalently, the normalized left hand eigenvector associated with the eigenvalue $\lambda_{\mathbf{A}}=0$ for $\mathbf{A}=\mathbf{I}-\mathbf{T}$. The known facts concerning the eigenvectors of a perturbed matrix are therefore, to some extent, relevant. However, there are some peculiar aspects which the general eigenvector theory does not capitalize upon. For example, the stochastic nature of the problem sets it apart. The fact that the relevant eigenvalue, $\lambda_{\mathbf{A}}=0$, (as well as its multiplicity) is unaltered by the perturbation is certainly special. The perturbation $\mathbf{E}$ is constrained to be one of a special kind, namely one which preserves the ergodic nature of the chain.

Moreover, the problem at hand is not concerned with the sensitivity of the *entire* eigensystem of a stochastic matrix. Only a very special eigenvalue and eigenvector with peculiar properties are involved. One should therefore not be too surprised to find some sort of special behavior exhibited which is not present in the general theory.

In general, if $\mathbf{x}$ is an eigenvector for $\mathbf{B}$ such that $(\mathbf{B}-\lambda_1\mathbf{I})\mathbf{x}=\mathbf{0}$ and there exists another eigenvalue $\lambda_2$, of $\mathbf{B}$ which is close to $\lambda_1$, then one expects $\mathbf{x}$ to be sensitive to perturbations in $\mathbf{B}$. (See [10]). However, this can produce some wrong impressions when applied to the special case at hand. The following example illustrates how applying this general theory can be somewhat misleading. Let $\mathscr{C}_1$ and $\mathscr{C}_2$ be two ergodic chains whose transition matrices are given by

$$\mathbf{T}_1=\begin{bmatrix} .99995 & .00005 & 0 \\ 0 & .99995 & .00005 \\ .99995 & 0 & .00005 \end{bmatrix} \tag{6.1}$$

and

$$\mathbf{T}_2=\begin{bmatrix} .99995 & .00005 \\ .00005 & .99995 \end{bmatrix}.$$

The eigenvalues for $\mathbf{A}_1=\mathbf{I}-\mathbf{T}_1$ are $\lambda_1=0$, $\lambda_2\approx.000100002$ and $\lambda_3\approx.999949998$, while the limiting probability vector is $\boldsymbol{\omega}_1\approx(.4999875, .4999875, .000025)$. The eigenvalues for $\mathbf{A}_2=\mathbf{I}-\mathbf{T}_2$ are $\mu_1=0$ and $\mu_2=.0001$ and the limiting probability vector is $\boldsymbol{\omega}_2=(.5, .5)$. In each case the matrices have another eigenvalue very close to the eigenvalue 0. The general perturbed eigenvector theory therefore suggests that the eigenvectors associated with the eigenvalue 0 should be sensitive to perturbations in the elements of each of the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$.

However, if one allows the term "sensitive" to mean that small relative errors in the $\mathbf{A}$ matrix can produce large relative errors in the limiting vector $\boldsymbol{\omega}$, then the sensitivity of the limiting probabilities may or may not be greatly influenced by the distance between the eigenvalue 0 and the other eigenvalues of $\mathbf{A}$.

For the two chains, $\mathscr{C}_1$ and $\mathscr{C}_2$, of the above example, one finds that

$$\mathbf{A}_1^{\#}\approx\begin{bmatrix} 5000 & -4999.75 & -.25 \\ -5000.25 & 5000 & .25 \\ 4999.5 & -5000.25 & .75 \end{bmatrix},$$

and

$$\mathbf{A}_2^{\#}=\begin{bmatrix} 5000 & -5000 \\ -5000 & 5000 \end{bmatrix},$$

so that $\kappa_\infty(\mathscr{C}_1) \approx 15{,}000$ while $\kappa_\infty(\mathscr{C}_2) = 1$. Theorem 4.3 guarantees that the chain $\mathscr{C}_2$ is well conditioned while $\mathscr{C}_1$ is more badly conditioned. Indeed, if $\mathscr{C}_1$ is perturbed so as to produce $\tilde{\mathscr{C}}_1$ with $\tilde{\mathbf{T}}_1 = \mathbf{T}_1 - \mathbf{E}$, where

$$\mathbf{E} = \begin{bmatrix} .001 & -.001 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

one finds that

$$\tilde{\boldsymbol{\omega}}_1 \approx (.045452376, .954499897, .000047727),$$

so that a relative error of $10^{-3}$ in $\mathbf{A}_1$ (using the $\infty$-norm) produces a relative error of about .91 in $\boldsymbol{\omega}_1$. In contrast, Theorem 4.3 guarantees that a relative error of $10^{-3}$ in $\mathbf{A}_2$ can produce a relative error of *at most* $1/999 \approx 10^{-3}$ in $\boldsymbol{\omega}_2$.

The conclusion is that one cannot always use the distance between $\lambda = 0$ and the nearest nonzero eigenvalue of $A$ as a measure of how sensitive the limiting probabilities are to perturbations, so that this criterion is not an accurate measure of the condition of an ergodic chain.

**7. Why not treat this strictly as a system of linear equations?** If $\mathbf{T}_{n\times n}$ is the transition matrix of an ergodic chain, it follows that $\mathbf{A} = \mathbf{I} - \mathbf{T}$ has rank $n-1$, and any subset of $n-1$ columns of $\mathbf{A}$ is linearly independent. The problem of finding the limiting probability vector is simply that of solving the system

$$\boldsymbol{\omega}\mathbf{A} = \mathbf{0}, \qquad \sum \omega_i = 1.$$

Clearly, this is equivalent to one $n \times n$ nonsingular system of the form $\boldsymbol{\omega}\mathbf{M} = \mathbf{b}$, where $\mathbf{M}$ is obtained from $\mathbf{A}$ by replacing one column (say the $k$th one) by the column $\mathbf{j} = [1, 1, \cdots, 1]^{\mathbf{T}}$ and $\mathbf{b}$ is the $k$th unit vector.

Since $\mathbf{M}$ is nonsingular and $\mathbf{b}$ is not subject to perturbation, the standard result (which is the analogue of Theorem 4.3) holds. That is, if a perturbation of the transition probabilities causes $\mathbf{M}$ to go to $\tilde{\mathbf{M}} = \mathbf{M} + \mathbf{F}$ where $\|\mathbf{F}\|\,\|\mathbf{M}^{-1}\| < 1$, then

$$\frac{\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|}{\|\boldsymbol{\omega}\|} \leqq \frac{\|\mathbf{F}\|/\|\mathbf{M}\| \text{ cond } (\mathbf{M})}{1 - \|\mathbf{F}\|/\|\mathbf{M}\| \text{ cond } (\mathbf{M})}, \qquad \text{where cond } (\mathbf{M}) = \|\mathbf{M}\|\,\|\mathbf{M}^{-1}\| \tag{7.1}$$

and $\|\mathbf{I}\| = 1$. (See [10]). This suggests that cond $(\mathbf{M})$ might also be used as a measure of the condition of the chain.

However, converting the singular matrix $\mathbf{A}$ into the nonsingular matrix $\mathbf{M}$ can drastically alter the condition of the problem. That is, although $\mathbf{A}$ is singular, it can be well conditioned in the sense that $\|\mathbf{A}\|\,\|\mathbf{A}^{\#}\|$ is small, whereas the modified matrix $\mathbf{M}$ is nonsingular but $\|\mathbf{M}\|\,\|\mathbf{M}^{-1}\|$ can be very large.

For example, consider the chain $\mathscr{C}$ whose transition matrix is

$$\mathbf{T}_{n\times n} = \begin{bmatrix} 1-\varepsilon & \frac{\varepsilon}{n-1} & \frac{\varepsilon}{n-1} & \cdots & \frac{\varepsilon}{n-1} \\ \frac{\varepsilon}{n-1} & 1-\varepsilon & \frac{\varepsilon}{n-1} & \cdots & \frac{\varepsilon}{n-1} \\ \vdots & & & & \\ \frac{\varepsilon}{n-1} & \frac{\varepsilon}{n-1} & \frac{\varepsilon}{n-1} & \cdots & 1-\varepsilon \end{bmatrix}, \qquad 0 < \varepsilon < 1,$$

so that

$$\mathbf{A}_{n\times n} = \frac{\varepsilon}{n-1}\begin{bmatrix} n-1 & -1 & -1 & \cdots & -1 \\ -1 & n-1 & -1 & \cdots & -1 \\ \vdots & & & & \\ -1 & -1 & -1 & \cdots & n-1 \end{bmatrix}.$$

Clearly, $\mathbf{A}$ is positive semidefinite so that

$$\kappa_2(\mathscr{C}) = \|\mathbf{A}\|_2\|\mathbf{A}^{\#}\|_2 = \frac{\max \lambda_i}{\min_{\lambda_i \neq 0} \lambda_i}, \qquad \text{where } \lambda_i \text{ denotes eigenvalue.}$$

It is easy to verify that the eigenvalues of $\mathbf{A}$ are given by

$$\left\{0, \frac{\varepsilon n}{n-1}, \frac{\varepsilon n}{n-1}, \frac{\varepsilon n}{n-1}, \cdots \frac{\varepsilon n}{n-1}\right\},$$

so that $\kappa_2(\mathscr{C}) = 1$, regardless of what value is assigned to $\varepsilon$ and what the size of $n$ is. Now replace any column (say the $k$th one) of $\mathbf{A}$ by $\mathbf{j} = [1, 1, \cdots, 1]^{\mathrm{T}}$, so as to form the matrix $\mathbf{M}$. The matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ then has the form

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = n\left(\frac{\varepsilon}{n-1}\right)^2\begin{bmatrix} n-1 & -1 & -1 & \cdots & 0 & -1 & \cdots & -1 \\ -1 & n-1 & -1 & \cdots & 0 & -1 & \cdots & -1 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & 0 & \cdots & \left(\frac{n-1}{\varepsilon}\right)^2 & 0 & \cdots & 0 \\ -1 & -1 & -1 & \cdots & 0 & n-1 & \cdots & -1 \\ \vdots & & & & \vdots & & & \\ -1 & -1 & -1 & \cdots & 0 & -1 & \cdots & n-1 \end{bmatrix}.$$

It is not difficult to see that the eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ are given by

$$\left\{n, n\left(\frac{\varepsilon}{n-1}\right)^2, \left(\frac{n\varepsilon}{n-1}\right)^2, \cdots, \left(\frac{n\varepsilon}{n-1}\right)^2\right\}$$

so that for large $n$ or small $\varepsilon$,

$$\operatorname{cond}_2(\mathbf{M}) = \frac{\text{max singular value}}{\text{min singular value}} = \frac{n-1}{\varepsilon}.$$

Thus $\operatorname{cond}_2(\mathbf{M})$ can be made arbitrarily large by taking either $\varepsilon$ small or $n$ large. Note also that $\operatorname{cond}_2(\mathbf{M})$ is independent of which column is selected to contain the 1's.

It clearly would be a mistake to use cond $(\mathbf{M})$ as any sort of guide to the sensitivity the limiting probabilities might exhibit to perturbations in the transition probabilities. Aside from the theoretical hazards which the matrix $\mathbf{M}$ can produce, it is obvious that $\mathbf{M}$ could also present numerical difficulties if it were used in any sort of computational scheme.

The bound produced by using $\mathbf{M}$ and (7.1) is almost always inferior to the bound obtained from Theorem 4.3. As an example, consider again the three-state chain $\mathscr{C}$, whose transition matrix is given by (6.1). Suppose this chain is perturbed so that the transition matrix becomes $\tilde{\mathbf{T}} = \mathbf{T} - \mathbf{E}$, where

$$\mathbf{E} = \begin{bmatrix} -2.5\times 10^{-5} & 2.5\times 10^{-5} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$. $\tilde{\mathbf{M}}$ is obtained from $\tilde{\mathbf{A}}$ by replacing some column of $\tilde{\mathbf{A}}$ by $\mathbf{j}$. Assume that in $\mathbf{M}$ as well as in $\tilde{\mathbf{M}}$, the column which is $\mathbf{j}$ is taken to be the second column. Then,

$$\mathbf{F} = \tilde{\mathbf{M}} - \mathbf{M} = \begin{bmatrix} -2.5 \times 10^{-5} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

is the perturbation in $\mathbf{M}$ in (7.1). Using the $\infty$-norm, one finds that $\text{cond}_\infty(\mathbf{M}) \sim 60{,}000$ whereas $\kappa_\infty(\mathscr{C}) \approx 15{,}000$. The bound for the relative error in $\boldsymbol{\omega}$ which (7.1) provides is approximately 1 whereas the bound produced by (4.3) is about .6. In this case, the actual relative error (with the $\infty$-norm) is about .3334.

This example exhibits only a single case where (4.3) is superior to (7.1). However, experience has shown this to be typical. For each value of $n = 3$, 5, 10, 20, and 30, twenty $n$-state ergodic chains were randomly generated. A random perturbation (which satisfied the hypothesis of Theorem 4.3 and (7.1)) was introduced, and the bounds given by (4.3) and (7.1) were computed using the $\infty$-norm. For $n = 3$, (4.3) gave a better bound than (7.1) in 13 out of the 20 trials. For $n = 5$, (4.3) gave a better bound in 18 out of 20 trials. For each of the cases $n = 10$, $n = 20$, and $n = 30$, (4.3) was found to be superior in 20 out of 20 trials. Moreover, for each of the 100 chains generated, $\kappa_\infty(\mathscr{C})$ was never significantly greater than 5 whereas $\text{cond}_\infty(\mathbf{M}_{n\times n})$ was always in the neighborhood of $n^2$.

Since the goal was not to use $\mathbf{M}$ in any sort of computational scheme, but rather to determine the degree to which characteristics of $\mathbf{M}$ (e.g., cond ($\mathbf{M}$)) reflect the relative sensitivity of the limiting probabilities, no attempt was made to scale $\mathbf{M}$. This, of course, could be done, and should be done if $\mathbf{M}$ is specifically given and is to be used in computations. However, when $\mathbf{M}$ is not specifically given, no theoretical advantage as far as producing a general analytical bound on the relative error can be realized.

**Acknowledgments.** The author wishes to thank Professor Gene Golub for his valuable comments and suggestions, as well as for his hospitality during the author's stay at Stanford University.

## REFERENCES

[1] EDWARD T. BROWNE, *Introduction to the Theory of Determinants and Matrices*, University of North Carolina Press, Chapel Hill NC, 1958.

[2] S. L. CAMPBELL AND CARL D. MEYER, *Generalized Inverses of Linear Transformations*, Surveys and Reference Works in Mathematics, Pitman Pub. Co., London, 1979.

[3] ———, *Continuity properties of the Drazin pseudoinverse*, Linear Algebra and Appl., 10 (1975), pp. 77–83.

[4] F. R. GANTMACHER, *Matrix Theory*, Vol. 2., Chelsea Pub. Co., New York, 1960.

[5] J. G. KEMEMY AND J. L. SNELL, *Finite Markov Chains*, D. Van Nostrand Co., New York, 1960.

[6] CARL D. MEYER, *The role of the group generalized inverse in the theory of finite markov chains*, SIAM Rev., 17 (1975), pp. 443–464.

[7] CARL D. MEYER AND N. J. ROSE, *The index and the Drazin inverse of block triangular matrices*, SIAM J. Appl. Math., 33 (1977), pp. 1–7.

[8] L. MIRSKY, *An Introduction to Linear Algebra*, Oxford at the Clarendon Press, London, 1963.

[9] PAUL J. SCHWEITZER, *Perturbation theory and finite Markov chains*, J. Appl. Prob., 5 (1968), pp. 401–413.

[10] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0005 $01.00/0

# NONNEGATIVE CHOLESKY DECOMPOSITION AND ITS APPLICATION TO ASSOCIATION OF RANDOM VARIABLES*

ALLAN R. SAMPSON†

**Abstract.** The concept of a multivariate family of distributions indexed by a covariance scale parameter $\Sigma$ is formally defined and examples given. The multivariate normal is one such family. Sufficient conditions are given so that a positive definite matrix has a nonnegative Cholesky decomposition. These conditions also yield the association of random variables with a covariance scale parameter distribution. These results are related to other matrix results and to Barlow and Proschan's stronger conditions (*Statistical Theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart, Winston, New York, 1975), for the association of the multivariate normal, namely, $\lambda_{ij} \leqq 0$, $i \neq j$ where $\Sigma^{-1} = \Lambda = \{\lambda_{ij}\}$.

**1. Introduction and summary.** The association of random variables having a multivariate normal distribution or some related distributions has been studied among others by Barlow and Proschan (1975), Abdel-Hameed and Sampson (1978), and Ahmed, Leon and Proschan (1978). Most of the approaches involve the specific distributional form of the multivariate normal or some relationships specific to the multivariate normal.

In this paper we give new sufficient conditions for the association of random variables having the multivariate normal, and show more importantly that these conditions are sufficient for the association of a broad class of multivariate random variables. For the random variables of interest, this condition is given in terms of a scale parameter which is an analogue of the covariance matrix. Additionally we connect the previously known association conditions with this new covariance condition. The approach we follow involves deriving a nonnegative version of the Cholesky decomposition which can then be used with standard association theorems to derive our results. Related matrix results are also considered, as well as, applications to a bivariate Cauchy and a bivariate extended gamma distribution.

In § 2, we give the matrix theory results concerning the Cholesky decomposition and discuss the related concept of completely positive matrices. In § 3 we define and give examples of distributions with a covariance scale parameter. We then apply the matrix results in § 4 to obtain sufficient conditions for the association of random variables with a distribution having a covariance scale parameter. Some additional comments are presented in § 5.

**2. Nonnegative Cholesky decomposition and completely positive matrices.** The basic matrix results that are obtained provide sufficient conditions so that a positive definite matrix has a nonnegative Cholesky decomposition. Additionally, we make use of certain results pertaining to completely positive matrices.

Denote by $\tau$ the set of lower triangular matrices with positive diagonal elements. If a matrix $\mathbf{A}$ has all positive (nonnegative) elements, we write $\mathbf{A} > 0 (\geqq 0)$; and if $\mathbf{S}$ is a positive definite matrix, we write $\mathbf{S}$ is p.d. The multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$ is denoted $N(\boldsymbol{\mu}, \Sigma)$.

DEFINITION 2.1. (See Varga (1962, p. 85)). $\Lambda = \{\lambda_{ij}\}$ is a *Stieltjes matrix* if $\Lambda$ is p.d. and $\lambda_{ij} \leqq 0$, for all $i \neq j$.

* Received by the editors October 15, 1979, and in final revised form January 15, 1980.

† Institute for Statistics and Applications, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh PA 15260. The work of this author is sponsored by the Air Force Office of Scientific Research under Contract F49620-79-C-0161.

THEOREM 2.1. (See Varga (1962, p. 85)). *If $\mathbf{\Lambda}$ is a Stieltjes matrix then $\mathbf{\Lambda}^{-1} \geqq 0$.*

In general the converse to Theorem 2.1 is false. Theorem 2.2 and Corollary 2.1 that follow will be used to connect standard association conditions to the new condition.

THEOREM 2.2. *Let $\mathbf{\Sigma}$ be a $p \times p$ symmetric matrix and partitioned as*

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}' & s \end{bmatrix},$$

*where $\mathbf{\Sigma}_{11}$ is $(p-1)\times(p-1)$. If $\mathbf{\Sigma}^{-1}$ is a Stieltjes matrix, then $\mathbf{\Sigma}_{11}^{-1}$ is a Stieltjes matrix.*

*Proof.* Note that

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} (\mathbf{\Sigma}_{11} - s^{-1}\boldsymbol{\sigma}\boldsymbol{\sigma}')^{-1} & -\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma}(s - \boldsymbol{\sigma}'\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma})^{-1} \\ -s^{-1}\boldsymbol{\sigma}'(\mathbf{\Sigma}_{11} - s^{-1}\boldsymbol{\sigma}\boldsymbol{\sigma}')^{-1} & (s - \boldsymbol{\sigma}'\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma})^{-1} \end{bmatrix}.$$

Because $\mathbf{\Sigma}^{-1}$ is a Stieltjes matrix, it follows that $\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma} \geqq 0$, $(s - \boldsymbol{\sigma}'\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma})^{-1} \geqq 0$, and $(\mathbf{\Sigma}_{11} - s^{-1}\boldsymbol{\sigma}\boldsymbol{\sigma}')^{-1}$ is a Stieltjes matrix. By the result of Woodbury (1950) (also see Rao (1973, p. 33)), we can write

$$\mathbf{\Sigma}_{11}^{-1} = (\mathbf{\Sigma}_{11} - s^{-1}\boldsymbol{\sigma}\boldsymbol{\sigma}')^{-1} - (s - \boldsymbol{\sigma}'\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma})^{-1}(\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma})(\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\sigma})';$$

hence, it now follows that $\mathbf{\Sigma}_{11}^{-1}$ is a Stieltjes matrix. □

DEFINITION 2.2. Suppose $\mathbf{\Sigma} = \{\sigma_{ij}\}$ is a $p \times p$ matrix. For $i > j$, $\mathbf{\Sigma}(i, j)$ is defined by

$$\mathbf{\Sigma}(i, j) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1j} & \sigma_{1i} \\ \vdots & & \vdots & \vdots \\ \sigma_{j1} & \cdots & \sigma_{jj} & \sigma_{ji} \\ \sigma_{i1} & \cdots & \sigma_{ij} & \sigma_{ii} \end{bmatrix}.$$

COROLLARY 2.1. *If $\mathbf{\Sigma}^{-1}$ is a Stieltjes matrix then $(\mathbf{\Sigma}(i, j))^{-1}$ is a Stieltjes matrix.*

*Proof.* Note that if $\mathbf{\Sigma}^{-1}$ is a Stieltjes matrix then $(\mathbf{P}\mathbf{\Sigma}\mathbf{P}')^{-1}$ is a Stieltjes matrix for any permutation matrix $\mathbf{P}$. Now there exists $\mathbf{P}$ such that

$$\mathbf{P}\mathbf{\Sigma}\mathbf{P}' = \begin{bmatrix} \mathbf{\Sigma}(i, j) & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix},$$

and the result follows from Theorem 2.2. □

DEFINITION 2.3. (*Cholesky Decomposition*). (See Wilkinson (1965, pp. 229–232)). If $\mathbf{\Sigma}$ is p.d., then there exists a unique $\mathbf{T} \in \tau$, such that $\mathbf{\Sigma} = \mathbf{T}\mathbf{T}'$.

There are a number of techniques for computing $\mathbf{T}$ based upon $\mathbf{\Sigma}$. The usual numerical techniques are recursive in nature (see Wilkinson). For our purposes a direct relationship is required; a suitable version of this is given in Lemma 2.1. Similar versions are used in standard existence proofs for the Cholesky decomposition.

LEMMA 2.1. *Suppose $\mathbf{\Sigma} = \mathbf{T}\mathbf{T}'$, $\mathbf{T} = \{t_{ij}\} \in \tau$. Then for $i > j$,*

$$t_{ij} = -\left(t_{jj} \prod_{k=1}^{j-1} t_{kk}^2\right)^{-1} |\mathbf{\Sigma}(i, j)| \gamma_{j+1,j},$$

*where $\gamma_{j+1,j}$ is the $(j+1, j)$th entry of $(\mathbf{\Sigma}(i, j))^{-1}$.*

*Proof.* The proof follows immediately from the following algebraic identity:

$$\det\left(\begin{bmatrix} t_{11} & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ \cdot & & & 0 \\ \cdot & & & \\ t_{j1} & \cdots & \cdots & t_{jj} \end{bmatrix}\begin{bmatrix} t_{11} & \cdots & t_{j-1,1} & t_{i1} \\ \vdots & & \vdots & \vdots \\ 0 & & t_{j-1,j-1} & t_{i,j-1} \\ 0 & \cdots & 0 & t_{ij} \end{bmatrix}\right) = -\text{cofactor}\,(\mathbf{C}_{j+1,j}),$$

where $\mathbf{C}_{j+1,j}$ is the submatrix of $\mathbf{\Sigma}(i, j)$ where the $(j+1)$st row and $j$th column are deleted. □

The two principal results of direct use to us now follow.

THEOREM 2.3. *Let $\Sigma$ be a positive definite $p \times p$ matrix. If the $(j+1, j)$th entry of $(\Sigma(i, j))^{-1}$ is nonpositive for $i > j$, $1 \leqq j \leqq p-1$, then $\Sigma = \mathbf{TT}'$ where $\mathbf{T} \in \tau$ and $\mathbf{T} \geqq 0$.*

*Proof.* This follows directly from Lemma 2.1. □

COROLLARY 2.2. *If $\Sigma^{-1}$ is a Stieltjes matrix then $\Sigma = \mathbf{TT}'$ where $\mathbf{T} \in \tau$ and $\mathbf{T} \geqq 0$.*

*Proof.* This follows directly from Theorem 2.3, Corollary 2.1 and the definition of Stieltjes matrix. □

Theorem 2.3 and Corollary 2.2 are related to certain results concerning completely positive matrices.

DEFINITION 2.4. (Hall and Newman (1963)). $\Sigma$, a $p \times p$ matrix, is a completely positive matrix if $\Sigma$ is p.d. and there exists $\mathbf{C}$, a $p \times n$ matrix, with $\mathbf{C} \geqq 0$ such that $\Sigma = \mathbf{CC}'$.

THEOREM 2.4. (Diananda (1962)). *Let $\Sigma$ be a p.d. $p \times p$ matrix with $p \leqq 4$. If $\Sigma \geqq 0$, then $\Sigma$ is completely positive.*

Hall and Newman show by counterexample that Theorem 2.4 does not hold for $p \geqq 5$. Gray and Wilson (1979) comment further on the results of Diananda, and Hall and Newman. Hence, Theorem 2.3 can be viewed as providing conditions when $p \geqq 5$ for $\Sigma$ to be completely positive; also for $p = 2$, the condition of Theorem 2.3 reduces to $\sigma_{12} \geqq 0$. However, the conditions of Theorem 2.3 actually provide a stronger result in that the appropriate decomposition of $\Sigma$ can be accomplished with nonnegative triangular matrices. An interesting alternative version of Theorem 2.4 can be obtained for the case $p = 3$.

LEMMA 2.2. *Let $\Sigma$ be a p.d. $3 \times 3$ matrix. If $\Sigma \geqq 0$, then there exist a permutation matrix $\mathbf{P}$ and $\mathbf{T} \in \tau$ with $\mathbf{T} \geqq 0$ such that $\mathbf{P}\Sigma\mathbf{P}' = \mathbf{TT}'$.*

*Proof.* Write $\Sigma = \mathbf{TT}'$, $\mathbf{T} \in \tau$. If $\mathbf{T} \geqq 0$, then the result is immediate. If $\mathbf{T} \not\geqq 0$, then the (3, 2) entry of $\mathbf{T}$ is negative; having the (2, 1) or (3, 1) entries negative would imply $\Sigma \not\geqq 0$. Without loss of generality we write

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & -c & 1 \end{bmatrix},$$

where $a \geqq 0$, $b \geqq 0$, $c > 0$. There exists a permutation matrix $\mathbf{P}$ such that

$$\mathbf{PTT'P'} = \begin{bmatrix} 1+b^2+c^2 & b & ab-c \\ b & 1 & a \\ ab-c & a & 1+a^2 \end{bmatrix}.$$

The Cholesky decomposition $\mathbf{UU}'$ for $\mathbf{PTT'P'}$ is

$$\mathbf{U} = \begin{bmatrix} (1+b^2+c^2)^{1/2} & 0 & 0 \\ \dfrac{b}{(1+b^2+c^2)^{1/2}} & \dfrac{(1+c^2)^{1/2}}{(1+b^2+c^2)^{1/2}} & 0 \\ \dfrac{ab-c}{(1+b^2+c^2)^{1/2}} & \dfrac{(1+b^2+c^2)^{1/2}}{(1+c^2)^{1/2}}\left(a - \dfrac{b(ab-c)}{1+b^2+c^2}\right) & \dfrac{1}{(1+c^2)^{1/2}} \end{bmatrix}.$$

The nonnegativity of $u_{21}, u_{31}$, where $\mathbf{U} = \{u_{ij}\}$, follows from the fact $\mathbf{PTT'P'} \geqq 0$. The sign of $u_{32}$ is determined by $a(1+b^2+c^2) - ab^2 + bc = a + ac^2 + bc \geqq 0$. Hence, $\mathbf{U} \geqq 0$.

**3. Covariance scale parameters.** An important result concerning the multivariate normal distribution with known mean vector is that the covariance matrix uniquely indexes the distribution. Moreover, the covariance matrix being diagonal is equivalent

to the vector entries being independent. In generating new multivariate distributions, it is sometimes desirable to preserve these two properties. For example, elliptically symmetric families (e.g., Kelker (1970)) are uniquely indexed by their covariance matrices, but diagonality does not necessarily imply independence.

One approach to generating multivariate families with these two properties is to employ the covariance matrix or its analogue as a scale parameter in a multivariate distribution whose components were originally independent. Suppose that $\mathbf{X} = (X_1, \cdots, X_p)'$ is a vector with independent entries. Let $\mathbf{Y} = \mathbf{TX}$ where $\mathbf{T} \in \Theta \subseteq \tau$. The family of distributions of $\mathbf{Y}$ is then indexed by $\mathbf{T} \in \tau$. By the uniqueness of the Cholesky decomposition (Definition 2.3), the family of distributions can be equivalently indexed by $\mathbf{\Sigma} = \mathbf{TT}'$, where $\mathbf{\Sigma}$ lies in a parameter space $S$. Denote the family of distributions of $\mathbf{Y}$ by $F(\mathbf{y}; \mathbf{\Sigma})$.

DEFINITION 3.1. The family of distributions $F(\mathbf{y}; \mathbf{\Sigma})$ indexed by $\mathbf{\Sigma} \in S$ is said to be a family of distributions with *covariance scale parameter* $\mathbf{\Sigma}$.

The following lemma provides some essential results for families of this type. The proof is straightforward.

LEMMA 3.1. *For a family of distributions with covariance scale parameter* $\mathbf{\Sigma}$, *the entries of* $\mathbf{Y}$ *are independent if and only if* $\mathbf{\Sigma} = \mathbf{D}$, *where* $\mathbf{D}$ *is a diagonal matrix. Moreover, if* $\operatorname{var} X_1 = \cdots = \operatorname{var} X_p = 1$, *then* $\mathbf{\Sigma}$ *is the covariance matrix of* $\mathbf{Y}$.

This basic approach to generating families of multivariate distributions has been considered in a number of contexts. Steffenson (1922) originally used a method analogous to this to generate multivariate distributions with certain correlational properties. Mardia (1970) reviews this approach and discusses related estimation topics. Triangular transformations were employed by Arnold (1967) for constructing bivariate distributions with certain dependence properties. Just to mention a couple of other contexts in which this general approach has been useful, we note its use in simulation and normality testing (Gnanadesikan (1974, p. 177)), nonparametric power calculations (Jogdeo (1964)), and reliability applications (Krishnaiah (1977)).

Clearly the family of distributions $N(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is p.d., is a family of distributions with a covariance scale parameter $\mathbf{\Sigma}$.

Density functions for two other bivariate covariance scale parameter examples are

$$(3.1)\qquad f(y_1, y_2) = \pi^{-2}\sigma_1\sigma_2(1-\rho^2)^{1/2}(\sigma_1^2 + y_1^2)^{-1} \cdot ((1-\rho^2)\sigma_2^2 + (y_2 - \rho(1-\rho^2)^{-1/2}y_1)^2)^{-1},$$

where $-\infty < y_1 < \infty$, $-\infty < y_2 < \infty$; and

$$(3.2)\qquad g(y_1, y_2) = c(\lambda_1, \lambda_2, \mathbf{\Sigma}) y_1^{\lambda_1^2 - 1}(y_2 - \rho(1-\rho^2)^{-1/2}y_1)^{\lambda_2^2 - 1} \cdot \exp[-\lambda_1\sigma_1^{-1}y_1 - \lambda_2\sigma_2^{-1}(1-\rho^2)^{-1/2}(y_2 - \rho(1-\rho^2)^{-1/2}y_1)],$$

where $y_1 > 0$, $y_2 - \rho(1-\rho^2)^{-1/2}y_1 > 0$, $\lambda_1 > 0$, $\lambda_2 > 0$, and

$$c(\lambda_1, \lambda_2, \mathbf{\Sigma}) = [\Gamma(\lambda_1^2)\Gamma(\lambda_2^2)]^{-1}(\lambda_1/\sigma_1)^{\lambda_1^2}(\lambda_2/(\sigma_2(1-\rho^2)^{1/2}))^{\lambda_2^2}.$$

The density given by (3.1) is a bivariate Cauchy where each marginal has a Cauchy distribution. The density given by (3.2) has a $y_1$-marginal that is gamma and a $y_2$-marginal that is a weighted sum of independent gammas. For $Y_1$, $Y_2$ having the density of (3.2), the parameters $\sigma_1$, $\sigma_2$, and $\rho$ satisfy $\sigma_1^2 = \operatorname{var} Y_1$, $\sigma_2^2 = \operatorname{var} Y_2$ and $\rho = \operatorname{correl}(Y_1, Y_2)$.

**4. Association and covariance scale parameters.** In this section, sufficient conditions are given for the association of random variables with a covariance scale

parameter distribution. However, before providing the main result, we review some positive dependence concepts and note related results.

DEFINITION 4.1. (Barlow and Proschan (1975)). A nonnegative real valued function $f(u_1, \cdots, u_p)$ is said to be *totally positive of order* 2 *in pairs* ($TP_2$ in pairs) if for any pair of arguments $u_a$, $u_b$, $f(u_1, \cdots, u_a, \cdots, u_b, \cdots, u_p)$ viewed as a function of $u_a$, $u_b$ is $TP_2$. (For a definition of $TP_2$ see Karlin (1968)).

DEFINITION 4.2. (Barlow and Proschan (1975)). The random variables $U_1, \cdots, U_p$ are *conditionally increasing in sequence* (CIS) if for $i = 2, \cdots, p$, $P(U_i > u_i | U_{i-1} = u_{i-1}, \cdots, U_1 = u_1)$ is increasing in $u_1, \cdots, u_{i-1}$.

DEFINITION 4.3. (Esary, Proschan, and Walkup (1967)). The random variables $U_1, \cdots, U_p$ are *associated* if $\operatorname{Cov}[f(U_1, \cdots, U_p), g(U_1, \cdots, U_p)] \geqq 0$ for all nondecreasing functions $f$, $g$.

DEFINITION 4.4. (Lehmann (1966)). The random variables $U_1, \cdots, U_p$ are *positively quadrant dependent* (PQD) if $P(U_1 \leqq u_1, \cdots, U_p \leqq u_p) \geqq \prod P(U_i \leqq u_i)$ for all $u_1, \cdots, u_p$.

DEFINITION 4.5. (Ahmed, Langberg, Leon and Proschan (1978)). The random variables $U_1, \cdots, U_p$ are *positively orthant dependent* (POD) if $P(U_1 > u_1, \cdots, U_p > u_p) \geqq \prod P(U_i > u_i)$ for all $u_1, \cdots, u_p$.

LEMMA 4.1. *Let $U_1, \cdots, U_p$ have probability density function $f(u_1, \cdots, u_p)$. Then the following implications hold*: *($f(u_1, \cdots, u_p)$ is $TP_2$ in pairs) implies ($U_1, \cdots, U_p$ are CIS) implies ($U_1, \cdots, U_p$ are associated) implies ($U_1, \cdots, U_p$ are PQD) and also ($U_1, \cdots, U_p$ are POD).*

A proof of Lemma 4.1 may be found in Esary, Proschan, and Walkup (1967).

While the positive dependence properties of many different classes of distributions have been studied, we focus on the results previously obtained for the multivariate normal. Barlow and Proschan (1975) show that the density corresponding to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $TP_2$ in pairs if and only if $\lambda_{ij} \leqq 0$, $i \neq j$, where $\boldsymbol{\Lambda} = \{\lambda_{ij}\} = \boldsymbol{\Sigma}^{-1}$; that is, $\boldsymbol{\Sigma}^{-1}$ is a Stieltjes matrix. Special cases of this results have arisen in a number of other contexts (see, for instance, Ahmed, Leon and Proschan (1978)). Slepian (1962) showed that if $(Y_1, \cdots, Y_p)' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \geqq 0$, then $Y_1, \cdots, Y_p$ are PQD and also POD. (A proof of the POD result may be found in Gupta (1963)).

No corresponding positive dependence results appear to have been obtained for families of distributions with a covariance scale parameter. Basically we show that the new association results that are obtained for the normal distribution hold for covariance scale parameter families, and that the association results do not actually depend on the functional form of the distributions but only the "probabilistic structure."

THEOREM 4.1. *Let $\mathbf{Y} = (Y_1, \cdots, Y_p)' \sim F(\mathbf{y}, \boldsymbol{\Sigma})$, where $F(\mathbf{y}, \boldsymbol{\Sigma})$ is a family of distributions with covariance scale parameter $\boldsymbol{\Sigma}$. If the $(j+1, j)$th entry of $(\boldsymbol{\Sigma}(i, j))^{-1}$ is nonpositive for $i > j$, $1 \leqq j \leqq p-1$, then $Y_1, \cdots, Y_p$ are associated.*

*Proof.* By Theorem 2.3, there exists $\mathbf{T} \in \tau$ and $\mathbf{T} \geqq 0$ such that $\boldsymbol{\Sigma} = \mathbf{TT}'$. Hence, by definition of covariance scale parameter family, there exist independent random variables $X_1, \cdots, X_p$ so that $\mathbf{Y} = \mathbf{TX}$. By Property P4 of Esary, Proschan and Walkup (1967), $X_1, \cdots, X_p$ are associated. Because $\mathbf{T} \geqq 0$, $\mathbf{Y}$ is a nondecreasing function of $\mathbf{X}$, and by their Property P2 it follows that $Y_1, \cdots, Y_p$ are associated random variables. □

Note that when $p = 2$ the condition that $(\boldsymbol{\Sigma}(i, j))^{-1}$ has nonpositive $(j+1, j)$th element for $i > j$, $1 \leqq j \leqq p-1$ reduces to $\rho \geqq 0$, where $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$. For finite variance random variables, association implies $\operatorname{cov}(Y_1, Y_2) \geqq 0$, and, thus for these bivariate covariance scale parameter families we have that $Y_1$, $Y_2$ are associated if and only if $\rho \geqq 0$.

COROLLARY 4.1. *Let* $\mathbf{Y}=(Y_1, \cdots, Y_p)' \sim F(\mathbf{y}, \boldsymbol{\Sigma})$, *where* $F(\mathbf{y}, \boldsymbol{\Sigma})$ *is a family of distributions with covariance scale parameter* $\boldsymbol{\Sigma}$. *If* $\boldsymbol{\Sigma}^{-1}$ *is a Stieltjes matrix then* $Y_1, \cdots, Y_p$ *are associated random variables.*

*Proof.* It follows from Corollary 2.1 that $\boldsymbol{\Sigma}^{-1}$ being a Stieltjes matrix implies for $i>j$, $1 \leqq j \leqq p-1$ that $(\boldsymbol{\Sigma}(i, j))^{-1}$ is a Stieltjes matrix; which by definition yields that the $(j+1, j)$th entry of $(\boldsymbol{\Sigma}(i, j))^{-1}$ is nonpositive. The result follows immediately now from Theorem 4.1. □

COROLLARY 4.2. *Let* $\mathbf{Y}=(Y_1, \cdots, Y_p)' \sim F(\mathbf{y}, \boldsymbol{\Sigma})$, *where* $F(\mathbf{y}, \boldsymbol{\Sigma})$ *is a family of distributions with covariance scale parameter* $\boldsymbol{\Sigma}$. *If the* $(j+1, j)$*th entry of* $(\boldsymbol{\Sigma}(i, j))^{-1}$ *is nonpositive for* $i>j$, $1 \leqq j \leqq p-1$, *then for all* $y_1, \cdots, y_p$

(a) $$P(Y_1 \leqq y_1, \cdots, Y_p \leqq y_p) \geqq \prod_{i=1}^{p} P(Y_i \leqq y_i),$$

(b) $$P(Y_1 > y_1, \cdots, Y_p > y_p) \geqq \prod_{i=1}^{p} P(Y_i > y_i).$$

*Proof.* The proof is immediate from Theorem 4.1 and Lemma 4.1. □

Again the PQD and POD results of Corollary 4.2 hold when $\boldsymbol{\Sigma}^{-1}$ is a Stieltjes matrix.

COROLLARY 4.3. *Let* $\mathbf{Y}=(Y_1, \cdots, Y_p)' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *If* $\boldsymbol{\Sigma} \geqq 0$ *for* $p \leqq 4$, *or for* $p \geqq 5$ *the* $(j+1, j)$*th entry of* $(\boldsymbol{\Sigma}(i, j))^{-1}$ *is nonpositive for* $i>j$, $1 \leqq j \leqq p-1$, *then* $Y_1, \cdots, Y_p$ *are associated random variables.*

*Proof.* For $p \geqq 5$, the association of $\mathbf{Y}-\boldsymbol{\mu}$ and hence of $\mathbf{Y}$ follow from Theorem 4.1. For $p \leqq 4$, Theorem 2.4 and the obvious extension of Theorem 4.1 are required. □

We note that in a similar fashion, a location parameter could be introduced to families with a covariance scale parameter. The corresponding association results for these location and scale families would then be immediate from our previous results. Because triangular matrices are required for parametrizing arbitrary covariance scale families, we are unable to show for general scale families that when $p=3, 4$ and $\boldsymbol{\Sigma} \geqq 0$, the association result holds. To be able to do this would require two things: Lemma 2.2 holding for $p=4$, and a permutational invariance result for covariance scale families. However, in the special case of normality (Corollary 4.3), we are able to obtain the $p=3, 4$ results from the fact that (e.g., Anderson (1958, p. 19)) if $X$ has a normal distribution, then $\mathbf{CX}$ has a normal distribution, where $\mathbf{C}$ is any rectangular matrix.

**5. Comments.** An immediate application of Theorem 4.1 yields that if $(Y_1, Y_2)$ have p.d.f.'s given by (3.1) or (3.2), then $Y_1$ and $Y_2$ are associated when $\rho \geqq 0$. Also we note that for these families $\rho=0$ is equivalent to the independence of $Y_1$ and $Y_2$. Because the variances are finite and suitably scaled for the family of (3.2), it follows that $Y_1$, $Y_2$ are associated for (3.2) if and only if the correlation between $Y_1$ and $Y_2$ is nonnegative. For both examples, when $\rho \geqq 0$, the PQD and POD inequalities for $Y_1$ and $Y_2$ follow from Corollary 4.2.

It is observed that the condition of Theorem 4.1, i.e., that the $(j+1, j)$th element of

$$(\boldsymbol{\Sigma}(i, j))^{-1}=\begin{bmatrix} \boldsymbol{\Sigma}_{jj} & \boldsymbol{\sigma}_{i,j} \\ \boldsymbol{\sigma}_{i,j} & \sigma_{ii} \end{bmatrix}^{-1}$$

is nonpositive, is equivalent to the $j$th element of $\boldsymbol{\sigma}_{i,j}' \boldsymbol{\Sigma}_{jj}^{-1}$ being nonnegative. When $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, the conditional distribution of $Y_i$ given $(Y_1, \cdots, Y_j)' \equiv \mathbf{Y}_j$, is $N(\boldsymbol{\sigma}_{i,j}' \boldsymbol{\Sigma}_{jj}^{-1} \mathbf{y}_j, \sigma_{ii}-\boldsymbol{\sigma}_{i,j}' \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\sigma}_{i,j})$. In this case the nonnegativity of the $j$th element of $\boldsymbol{\sigma}_{i,j}' \boldsymbol{\Sigma}_{jj}^{-1}$ is equivalent to $P(Y_i > y_i \mid Y_1=y_1, \cdots, Y_j=y_j)$ being nondecreasing in $y_j$ for fixed values of $y_1, \cdots, y_{j-1}$. Thus, for the multivariate normal the conditions of Theorem 4.1

are equivalent to

$$(5.1)\qquad \begin{aligned} &P(Y_i > y_i \mid Y_1 = y_1, \cdots, Y_j = y_j) \text{ nondecreasing in } y_j \\ &\quad \text{for fixed values of } y_1, \cdots, y_{j-1}, \text{ and for all } i > j,\ j = 1, \cdots, p-1. \end{aligned}$$

We note that for the multivariate normal, $Y_1, \cdots, Y_p$ being CIS is equivalent to $\boldsymbol{\sigma}'_{j+1,j}\boldsymbol{\Sigma}_{jj}^{-1} \geqq 0$ for $j = 1, \cdots, p$. To see the nonequivalence of CIS and condition (5.1), first define

$$(5.2)\qquad \boldsymbol{\Sigma} = \begin{bmatrix} 3 & 2.2 & 1 \\ 2.2 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

If $(Y_1, Y_2, Y_3)' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is given by (5.2), then $P(Y_2 > y_2 \mid Y_1 = y_1)$ is increasing in $y_1$, $P(Y_3 > y_3 \mid Y_1 = y_1)$ is increasing in $y_1$, $P(Y_3 > y_3 \mid Y_1 = y_1, Y_2 = y_2)$ is increasing in $y_2$ for fixed $y_1$ and is *decreasing in* $y_1$ for fixed $y_2$. Therefore, $Y_1, Y_2, Y_3$ are not CIS and yet they satisfy (5.1) and the conditions of Theorem 4.1.

It is interesting to observe and straightforward to show that $\boldsymbol{\Sigma}^{-1}$ is not a Stieltjes matrix when $\boldsymbol{\Sigma}$ is defined by (5.2).

We conclude by showing that Theorem 2.1 can be proved using some positive dependence results. (A standard algebraic proof is given in Varga (1962)).

*Proof of Theorem* 2.1. If $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is a Stieltjes matrix then there exists $(Y_1, \cdots, Y_p)' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and by Corollary 4.1 (or Example 4.16, Barlow and Proschan (1975, pp. 145–150)), it follows that $Y_1, \cdots, Y_p$ are associated. But this implies $\boldsymbol{\Sigma} \geqq 0$, because the covariances of associated random variables are nonnegative. □

Finally we note that by identical argument, we can show that if the $(j+1, j)$th entry of $(\boldsymbol{\Sigma}(i, j))^{-1}$ is nonpositive for $i > j$, $i \leqq j \leqq p-1$, then $\boldsymbol{\Sigma} \geqq 0$.

**Acknowledgment.** The author wishes to acknowledge a helpful conversation with Professor Werner Rheinbold.

## REFERENCES

M. ABDEL-HAMEED AND A. R. SAMPSON (1978), *Positive dependence of the bivariate and trivariate absolute normal, t, $\chi^2$ and F distributions*, Ann. Statist. 6, pp. 1360–1368.

A. H. N. AHMED, R. LEON AND F. PROSCHAN (1978), *Generalization of Associated Random Variables with Applications*, Department of Statistics Report M468, Florida State University, Tallahassee.

A. H. N. AHMED, N. A. LANGBERG, R. LEON AND F. PROSCHAN (1978), *Two Concepts of Positive Dependence, With Applications in Multivariate Analysis*, Department of Statistics Report M486, Florida State University, Tallahassee.

T. W. ANDERSON (1958), *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York.

B. C. ARNOLD 1967, A note on multivariate distributions with specified marginals, J. Amer. Statist. Assoc., 62, pp. 1460–1461.

R. E. BARLOW AND F. PROSCHAN (1975), *Statistical Theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart and Winston, New York.

P. H. DIANANDA (1962), *On non-negative forms in real variables some or all of which are non-negative*, Proc. Camb. Phil. Soc., 58, pp. 17–25.

J. D. ESARY, F. PROSCHAN AND D. W. WALKUP (1967), *Association of random variables, with applications*, Ann. Math. Statist., 38, pp. 1466–1474.

R. GNANADESIKAN (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley, New York.

L. J. GRAY AND D. G. WILSON (to appear), *Nonnegative factorization of positive semi-definite matrices*, Linear Algebra and Appl.

S. S. GUPTA (1963), *Probability integrals of multivariate normal and multivariate t*, Ann. Math. Statist., 34, pp. 792–828.

M. HALL, JR. AND M. NEWMAN (1963), *Copositive and completely positive quadratic forms*, Proc. Camb. Phil. Soc., 59, pp. 329–339.

K. JOGDEO (1964), *Nonparametric Methods for Regression*, Report S330, Mathematics Center, Amsterdam.

S. KARLIN (1968), *Total Positivity, Vol.* 1, Stanford University Press, Stanford.

D. KELKER (1970), *Distribution theory of spherical distributions and a location–scale parameter generalization* Sankhyā Ser. A, 32, pp. 419–430.

P. R. KRISHNAIAH (1977), *On generalized multivariate gamma distributions and their applications in reliability*, in The Theory and Applications of Reliability Vol. 1, C. Tsokos and I. N. Shimi, ed., Academic Press, New York.

E. L. LEHMANN (1966), *Some concepts of dependence*, Ann. Math. Statist., 37, pp. 1137–1153.

K. V. MARDIA (1970), *Families of Bivariate Distributions*, Hafner Publishing, Darien CT.

C. R. RAO, (1973) *Linear Statistical Inference and its Applications*, 2nd ed., John Wiley, New York.

D. SLEPIAN (1962), *The one-sided barrier problem for Gaussian noise*, Bell System Tech. J., 41, pp. 463–501.

J. F. STEFFENSEN (1922), *A correlation-formula*, Skand. Aktuar., 5, pp. 73–91.

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

J. WILKINSON (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford.

M. WOODBURY (1950), *Inverting Modified Matrices*, Memorandum Report No. 42, Statistics Research Group, Princeton NJ.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0006 $01.00/0

# A NOTE ON THE NP-COMPLETENESS OF VERTEX ELIMINATION ON DIRECTED GRAPHS*

JOHN R. GILBERT†

**Abstract.** A correction is made to Rose and Tarjan's proof [SIAM J. Appl. Math., 1978] that determining a minimum fill-in elimination ordering for a directed graph is NP-complete.

Vertex elimination on directed graphs is a model of Gaussian elimination on sparse systems of linear equations, in which vertices model variables and edges model nonzero coefficients. Eliminating a variable from the system (causing certain zero coefficients to become nonzero) corresponds to deleting a vertex from the graph (causing certain edges to be added to the graph). The zero coefficients which become nonzero while the variables are eliminated in a specified order, or the edges which are added to the graph while the vertices are eliminated in a specified order, are called the *fill-in* associated with that order. Rose and Tarjan [1] present a number of algorithms relating to this model, and prove complexity results about some of the problems involved. The purpose of this note is to correct a flaw in the proof of their Theorem 10, which states that the problem of finding an elimination ordering with the smallest possible fill-in is NP-complete. We will not repeat their entire construction; the reader is assumed to have a copy of [1] at hand, so we will use their notation freely without further explanation.

The difficulty lies in the construction that simulates a 3CNF problem by a minimum fill-in problem. As the construction is given, in the second stage of elimination (eliminating $pqr(i)$ corresponding to false literals), suppose for example that in the clause $xyz$ only $y$ is false, so only vertex $xyz(2)$ is eliminated. This will cause fill-in of size $c^2$ from $xyz(1)$ to $X_{31}(xyz(2))$, which is unacceptably large. This problem can be solved by adding to the construction edges $pqr(1) \to X_{31}(pqr(2))$, $pqr(1) \to X_{31}(pqr(3))$, and $pqr(2) \to X_{31}(pqr(3))$, but now another problem arises. Suppose for example that in the clause $xyz$ only $x$ is false. Then $xyz(1)$ will be eliminated while edges $y \to xyz(1)$ and $xyz(1) \to X_{31}(xyz(2))$ still exist; but edges $y \to X_{31}(xyz(2))$ don't exist, so $c^2$ fill-in will occur.

Here follows a way to fix the construction. First we note that clauses containing more than one instance of the same variable will cause trouble, so we assume that for each variable $x$, no clause contains more than one $x$, more than one $\bar{x}$, or both $x$ and $\bar{x}$. A way to modify the formula so that this is the case is given at the end of this note.

To modify the construction, add the three edges per clause mentioned above:

$$pqr(1) \to X_{31}(pqr(2)), \quad pqr(1) \to X_{31}(pqr(3)), \quad pqr(2) \to X_{31}(pqr(3)).$$

Then remove all edges from vertices $x$ or $\bar{x}$ to $pqr(i)$ or $X_{31}(pqr(i))$, and instead add the edges:

1. From $x$ to $xqr(1)$, $pxr(2)$, and $pqx(3)$.
2. From $\bar{x}$ to $\bar{x}qr(1)$, $p\bar{x}r(2)$, and $pq\bar{x}(3)$.
3. From $x$ to: $\bar{x}qr(i)$ and $X_{31}(\bar{x}qr(i))$ for $i = 1, 2, 3$.
   $p\bar{x}r(i)$ and $X_{31}(p\bar{x}r(i))$ for $i = 2, 3$.
   $pq\bar{x}(3)$ and $X_{31}(pq\bar{x}(3))$.

Examples of the crucial part of the graph for clauses $zxy$ and $z\bar{x}y$ are given in Fig. 1.

* Received by the editors September 10, 1979, and in final revised form January 29, 1980.

† Department of Computer Science, Stanford University, Stanford, California 94305. This research was supported in part by a Hertz fellowship.

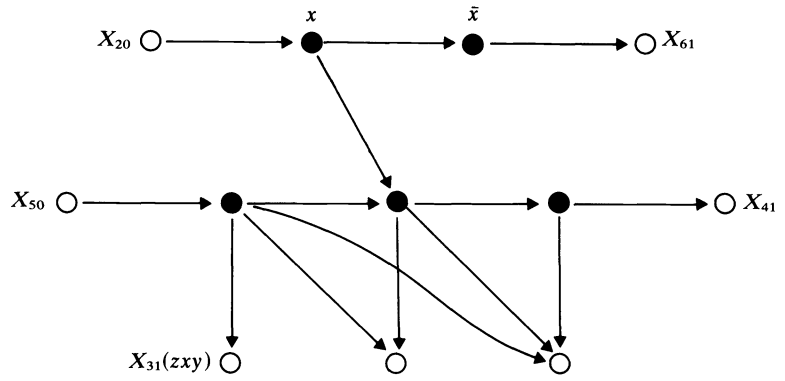



FIG. 1a. *Graph portion for clause zxy.*

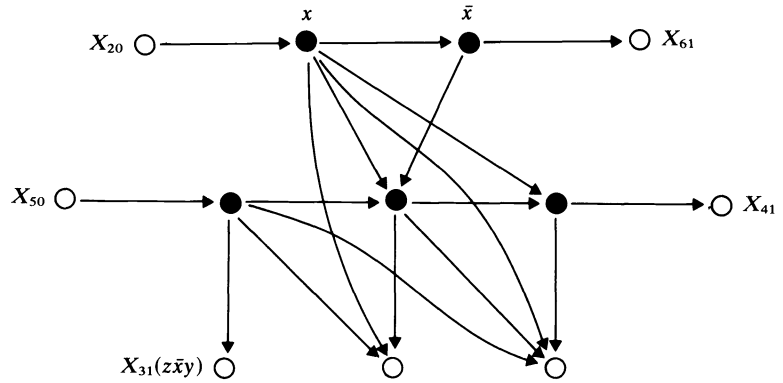



FIG. 1b. *Graph portion for clause zx̄y.*

Now if $F$ is satisfiable the elimination order still works, and fills in as given in [1] (except that $pqr(1) \to pqr(3)$ should be added to the second step; it doesn't increase fill-in because it occurs only if $X_{50} \to pqr(2)$ doesn't occur). It is easy to see that the only places this modification might change fill-in are during elimination of $\bar{x}$ in the first stage, and during the second stage. Since every edge $\bar{x} \to \alpha$ is accompanied by an edge $x \to \alpha$ except $\bar{x} \to X_{61}$, there is no new fill-in when $\bar{x}$ is eliminated. During the second stage we need to see that there is no fill-in from $x$ or $\bar{x}$ to $pqr(i)$ or $X_{31}(pqr(i))$. When eliminating a vertex $pqr(i)$, there is no fill-in from vertex $\bar{x}$ because either the current literal isn't $\bar{x}$ and hence there is no edge $\bar{x} \to pqr(i)$, or it *is* $\bar{x}$ and $\bar{x}$, being false, has already been eliminated. The situation for fill-in from $x$ is similar. Either the current literal is $x$, in which case vertex $x$ has been eliminated; or the vertex $pqr(i)$ has no edge $x \to pqr(i)$; or the current literal is in a clause containing an $\bar{x}$ (and hence no $x$) and every edge $x \to pqr(i)$ is matched by edges $x \to X_{31}(pqr(j))$ for $i \leqq j \leqq 3$, so there is no fill-in.

The proof that fill-in must be at least $(l+3s/2+1)b$ if $F$ is not satisfiable is straightforward, and is not changed by the modification to the construction. The only possible effect could be in case (ii): some $pqr(i)$ is eliminated before its corresponding variable vertex and also before any vertex in $X_{31}(pqr(i))$. With the modified construction this will still give $c^2$ fill-in from the variable vertex to $X_{31}(pqr(i))$.

There is also a trivial typographical error in Table 1 of [1]: $X_{31}(pqr(3))$ has as an adjacency in $X_{33}$, not $X_{23}$.

To modify a formula in 3CNF so that no clause contains more than one $x$ and/or $\bar{x}$ and so that every $x$ occurs the same number of times as $\bar{x}$, we may proceed as follows. First eliminate clauses containing both $x$ and $\bar{x}$, and eliminate duplicate instances of $x$ or $\bar{x}$ in each clause. Now replace each short clause $p \vee q$ by $(p \vee q \vee a) \wedge (p \vee q \vee \bar{a})$, where $a$ is a new variable. Similarly replace each short clause $p$ by $(p \vee a \vee b) \wedge (p \vee \bar{a} \vee b) \wedge (p \vee a \vee \bar{b}) \wedge (p \vee \bar{a} \vee \bar{b})$.

Now all clauses have length three, and none contains more than one $x$ and/or $\bar{x}$. To add an occurrence of $x$ we add clauses $(x \vee a \vee \bar{b}) \wedge (\bar{a} \vee c \vee \bar{d}) \wedge (b \vee \bar{c} \vee d)$, where $a$, $b$, $c$, and $d$ are new variables. The new variables balance their negations, and the new clauses are satisfied by $a = b = c = d = \textit{true}$. Repeating this procedure enough times will put the formula in the desired form, and will lead to at most a linear increase in size.

(In fact Rose and Tarjan's original limitation on the formula, that $x$ may not follow $\bar{x}$ in any clause, seems to be sufficient for the proof to hold; but the details of the argument are extremely messy.)

## REFERENCE

[1] D. J. ROSE AND R. E. TARJAN, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM J. Appl. Math., 34 (1978), pp. 176–197.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0007 $01.00/0

# THE FKG INEQUALITY AND SOME MONOTONICITY PROPERTIES OF PARTIAL ORDERS*

L. A. SHEPP†

**Abstract** Let $(a_1, \cdots, a_m, b_1, \cdots, b_n)$ be a random permutation of $1, 2, \cdots, m+n$. Let $P$ be a partial order on the $a$'s and $b$'s involving *only* inequalities of the form $a_i < a_j$ or $b_i < b_j$, and let $P'$ be an extension of $P$ to include inequalities of the form $a_i < b_j$; i.e., $P' = P \cup P''$, where $P''$ involves *only* inequalities of the form $a_i < b_j$. We prove the natural conjecture of R. L. Graham, A. C. Yao, and F. F. Yao [SIAM J. Alg. Discr. Meth. 1(1980), pp. 251–258] that in particular (*) $\Pr(a_1 < b_1 \mid P') \geqq \Pr(a_1 < b_1 \mid P)$. We give a simple example to show that the more general inequality (*) where $P$ is allowed to contain inequalities of the form $a_i < b_j$ is false. This is surprising because as Graham, Yao, and Yao proved, the general inequality (*) does hold if $P$ totally orders both the $a$'s and the $b$'s separately. We give a new proof of the latter result. Our proofs are based on the FKG inequality.

**1. Introduction.** Suppose $(a_1, a_2, \cdots, a_m, b_1, \cdots, b_n)$ is a random (uniformly distributed) permutation of $1, 2, \cdots, m+n$. Following [GYY], we might think of the permutation as the actual ranking of the tennis skill of players $a_1, \cdots, a_m, b_1, \cdots, b_n$. Here player $x$ always loses to player $y$ in a match if $x < y$. In a contest between two teams $A = \{a_1, \cdots, a_m\}$ and $B = \{b_1, \cdots, b_n\}$, suppose first that the teams have never met before but the players of each team have played some matches among themselves. Thus there is a partial order $P$ between certain $a$'s and certain $b$'s, e.g., $a_1 < a_2$, $a_1 < a_3$, $b_2 < b_1$, $\cdots$, but there is no *direct* information about the relative ranking of $a$'s vs. $b$'s. Denote by $\Pr(a_1 < b_1 \mid P)$ the conditional probability that $a_1$ loses to $b_1$, given the partial order $P$.

After some matches between $a$'s and $b$'s have taken place, in which we shall suppose that the $a$'s have lost each match to the $b$'s so far, we have a new partial ordering $P' = P \cup P''$, where $P''$ contains inequalities of the form $a_i < b_j$; e.g. $P'' = \{a_3 < b_4, a_5 < b_2, \cdots\}$. Note that there are two ways to think about $P$: if $P$ is thought of as a partial order on $\{a_1, \cdots, a_m, b_1, \cdots, b_n\}$, then the *union* $P \cup P'' = P'$ is the *larger* partial order based on the additional information in $P''$. However, we shall think of $P$ as a subset of permutations *defined* by the partial order $P$ so that the *intersection* $P \cap P'' = P'$ is the *smaller* subset of permutations based on the additional information in $P''$. Denote by $\Pr(a_1 < b_1 \mid P')$ the conditional probability that $a_1$ loses to $b_1$ given $P'$. It is tempting to conjecture that, in particular,

$$\Pr(a_1 < b_1 \mid P') \geqq \Pr(a_1 < b_1 \mid P). \tag{1.1}$$

The additional knowledge with $P'$ that $a$'s have lost to $b$'s prompts the belief (prejudice?) that $a$'s are inferior to $b$'s, and seems to make it more likely under $P'$ than under $P$ that $a_1$ loses to $b_1$. This conjecture of R. L. Graham, A. C. Yao, and F. F. Yao [GYY] is true, as we show. However the same intuition makes it even more tempting to conclude that (1.1) holds even if $P$ contains inequalities of the form $a_i < b_j$, because the prejudice under $P$ that $a$'s are inferior to $b$'s is apparently further reinforced by the new inequalities in $P'$. Nevertheless we give a simple example to show this is false. Indeed let $m = n = 2$ and

$$\begin{aligned} P &= \{a_1 < b_2, a_2 < b_1\}, \\ P' &= \{a_2 < b_2\} \cap P. \end{aligned} \tag{1.2}$$

* Received by the editors January 14, 1980, and in final form March 11, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

It is easy to check that $\Pr(P) = \frac{1}{4}$, $\Pr(P') = \Pr(a_1 < b_1,\ P) = \frac{5}{24}$, $\Pr(a_1 < b_1, P') = \frac{1}{6}$ so that (1.1) asserts that $\frac{4}{5} \geqq \frac{5}{6}$, which is of course false. An even simpler example was found by a referee: $m = n = 2$, $P = \{a_2 < b_1\}$, $P' = \{a_2 < b_2\} \cap P$, $\Pr(a_1 < b_1 \mid P') = \frac{5}{8} < \frac{2}{3} = \Pr(a_1 < b_1 \mid P)$.

The example (1.2) is especially surprising because (1.1) is valid even when $P$ contains inequalities of the form $a_i < b_j$, *provided that* $P$ also contains inequalities which give a *total ordering* of each of $A$ and $B$ separately. This was proved by Graham, Yao, and Yao [GYY], and we give a new proof here.

We next give a more general formulation of the two results to be proved in § 2, and discuss the FKG inequality which we will use in their proofs. Let $P_0$ be the subset of permutations for which $A$ and $B$ have the complete order:

$$P_0 = \{a_1 < a_2 < \cdots < a_m\} \cap \{b_1 < b_2 < \cdots < b_n\}. \tag{1.3}$$

Suppose $P_1$, $P_2$, $P_3$ are each subsets of permutations which are intersections of subsets of the form $\{a_i < b_j\}$. Then Graham, Yao, and Yao [GYY] proved:

**Theorem 1.** (Graham, Yao, Yao, [GYY]).

$$\Pr(P_1 \mid P_3 \cap P_2 \cap P_0) \geqq \Pr(P_1 \mid P_2 \cap P_0). \tag{1.4}$$

Note that this is the result stated in the preceding paragraph if $P_1$ is specialized to a single inequality $\{a_i < b_j\}$.

Let $Q_0$ be a subset of permutations defined by intersections of subsets of the form $\{a_i < a_j\}$ and $\{b_i < b_j\}$ *but not of the form* $\{a_i < b_j\}$ or $\{a_i > b_j\}$,

$$Q_0 = \{a_{i_1} < a_{j_1}, \cdots, a_{i_r} < a_{j_r}\} \cap \{b_{k_1} < b_{l_1}, \cdots, b_{k_s} < b_{l_s}\}, \tag{1.5}$$

and let $P_1$, $P_2$ be as in Theorem 1. Then Graham, Yao, and Yao [GYY] conjectured:

**Theorem 2.**

$$\Pr(P_1 \mid P_2 \cap Q_0) \geqq \Pr(P_1 \mid Q_0). \tag{1.6}$$

The FKG (Fortuin, Kasteleyn, Ginibre) inequality was discovered [FKG] in proving "intuitively obvious" conjectures about correlations in a statistical mechanics model. Although as shown in [FKG], the FKG hypothesis (1.7)–(1.10) is only sufficient for the conclusion (1.11), in the present case I found the simple counterexample (1.2) by looking for the simplest case of the general conjecture (1.1) for which the FKG technique does not easily apply. Other applications of the FKG inequality to prove known inequalities in combinatorics have been given in [SW]. D. J. Kleitman and J. B. Schearer [KS] also give an example where (1.1) fails if $P$ is allowed to contain $a_i < b_j$ inequalities, and give a different FKG proof for Theorem 1, but do not obtain Theorem 2.

The setting for the FKG inequality is as follows: Let $\Gamma$ be a finite lattice; i.e., $\Gamma$ is a finite set $\Gamma = \{x, y, z, \cdots\}$ with a partial order $x < y$ for which each pair $x, y \in \Gamma$ has a unique least upper bound $x \vee y$ and a unique greatest lower bound $x \wedge y$,

$$x \vee y \in \Gamma, \qquad x \wedge y \in \Gamma. \tag{1.7}$$

Further, $\Gamma$ is assumed distributive; i.e. for all $x, y, z \in \Gamma$

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z). \tag{1.8}$$

or equivalently, for all $x, y, z \in \Gamma$, $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$. Suppose $\mu, f, g$ are real-valued functions on $\Gamma$ for which for all $x, y \in \Gamma$,

$$\mu(x) \geqq 0, \qquad \mu(x)\mu(y) \leqq \mu(x \wedge y)\mu(x \vee y), \tag{1.9}$$

and $f$ and $g$ are monotonic in the same direction so that either

$$\begin{aligned} &f(x) \leqq f(y), \quad g(x) \leqq g(y), \text{ for all } x \text{ and } y, \text{ or} \\ &f(x) \geqq f(y), \quad g(x) \geqq g(y), \text{ for all } x < y. \end{aligned} \tag{1.10}$$

The FKG inequality [FKG] then asserts

$$\sum_{x\in\Gamma} f(x)g(x)\mu(x) \sum_{y\in\Gamma} \mu(y) \geqq \sum_{x\in\Gamma} f(x)\mu(x) \sum_{y\in\Gamma} g(y)\mu(y). \tag{1.11}$$

**2. Proofs of Theorems 1 and 2.** Let $\Gamma$ be the set of all $\binom{m+n}{m}$ subsets of $\{1, 2, \cdots, m+n\}$ with $m$ elements. For $x = \{x_1 < x_2 < \cdots < x_m\}$, $y = \{y_1 < \cdots < y_m\} \in \Gamma$, say that $x < y$ if $x_i \leqq y_i$, $i = 1, \cdots, m$. Thus the elements of $x \wedge y = \{(x \wedge y)_1 < \cdots < (x \wedge y)_m\}$ and $x \vee y = \{(x \vee y)_1 < \cdots < (x \vee y)_m\}$ are given for $i = 1, \cdots, m$ by

$$(x \wedge y)_i = \min (x_i, y_i), \quad (x \vee y)_i = \max (x_i, y_i). \tag{2.1}$$

Since $x \wedge y, x \vee y \in \Gamma$, (1.7) holds for $\Gamma, <$.

Examining all orderings of any three real numbers $\alpha, \beta, \gamma$ shows that

$$\min (\alpha, \max (\beta, \gamma)) = \max (\min (\alpha, \beta), \min (\alpha, \gamma)). \tag{2.2}$$

From (2.1) and (2.2) we see that (1.8) holds so that $\Gamma, <$ is also distributive.

Let $\bar{P}_1, \bar{P}_2, \bar{P}_3$ each be intersections of subsets of $\Gamma$ of the form $\{x_i \leqq k\}$, $i = 1, \cdots, m$, $k = 1, \cdots, m+n$. Let $\mu, f, g$ be defined by

$$\begin{aligned} &\mu(x) = \begin{cases} 1, & \text{if } x \in \Gamma \cap \bar{P}_2, \\ 0, & \text{else,} \end{cases} \\ &f(x) = \begin{cases} 1, & \text{if } x \in \Gamma \cap \bar{P}_1, \\ 0, & \text{else,} \end{cases}; \qquad g(x) = \begin{cases} 1, & \text{if } x \in \Gamma \cap \bar{P}_3, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{2.3}$$

Since $x_i \leqq k$, $y_i \leqq k$ implies that $\min (x_i, y_i) \leqq \max (x_i, y_i) \leqq k$, we see that $\mu(x) = \mu(y) = 1$ implies that $\mu(x \wedge y) = \mu(x \vee y) = 1$; thus (1.9) holds with equality. If $x \leqq y$ and $f(y) = 1$ then $y \in \bar{P}_1$. But if $y_i \leqq k$ then $x_i \leqq k$, so that $x \in \bar{P}_1$ as well, and $f(x) = 1$. Thus $f$ is decreasing, and similarly so is $g$. Thus (1.10) holds and the hypothesis of the FKG inequality is satisfied. By (1.11), it follows that

$$\#(\Gamma \cap \bar{P}_1 \cap \bar{P}_2 \cap \bar{P}_3)\#(\Gamma \cap \bar{P}_2) \geqq \#(\Gamma \cap \bar{P}_1 \cap \bar{P}_2)\#(\Gamma \cap \bar{P}_2 \cap \bar{P}_3), \tag{2.4}$$

where $\#(A)$ is the cardinality of $A$.

Consider the one-one correspondence $\phi : \Gamma \leftrightarrow P_0$ in (1.3); here $\phi(x) = (a_1, \cdots, a_m, b_1, \cdots, b_n)$, the permutation of $(1, 2, \cdots, m+n)$ which has $a_i(x) = x_i$, $i = 1, \cdots, m$, and $b_j(x) = j$th element of the complement of $x$ in $\{1, 2, \cdots, m+n\}$. Because the $a$'s and $b$'s are totally ordered by (1.3) in $P_0$, we have

LEMMA 2.5. *If* $(a_1, \cdots, a_m, b_1, \cdots, b_n) \in P_0$, *then* $a_i < b_j$ *if and only if* $a_i \leqq i + j - 1$.

It follows from (2.5) that for subsets $P_1, P_2, P_3$ as in Theorem 1 which are each intersections of subsets $\{a_i < b_j\}$, $\bar{P}_k = \phi^{-1}(P_i)$, $i = 1, 2, 3$, are each of the form $\{x_i \leqq k\}$; so that (2.4) holds. Since $\phi$ is one-one we have upon dividing by $((m+n)!)^2$,

$$\begin{aligned} &\Pr (P_0 \cap P_1 \cap P_2 \cap P_3) \Pr (P_0 \cap P_2) \\ &\qquad \geqq \Pr (P_0 \cap P_1 \cap P_2) \Pr (P_0 \cap P_2 \cap P_3), \end{aligned} \tag{2.6}$$

which is the same as (1.4). Theorem 1 is thus proved.

We next prove Theorem 2. For $N = 1, 2, \cdots$, let $\Gamma_N$ be the set of $N^{m+n}$ integer-valued vectors $x = (a_1, \cdots, a_m, b_1, \cdots, b_n)$ where each $a_i$ and $b_j \in \{1, 2, \cdots, N\}$. Denote

$$x_i = a_i = a_i(x), \quad i = 1, \cdots, m; \qquad x_{m+j} = b_j = b_j(x), \quad j = 1, \cdots, n. \tag{2.7}$$

For $x, y \in \Gamma_N$ say that $x < y$ if $x_i = a_i(x) \leqq a_i(y) = y_i$ but $x_{m+j} = b_j(x) \geqq b_j(y) = y_{m+j}$, $j = 1, \cdots, n$. The components of $x \wedge y$ and $x \vee y$ are

$$\begin{aligned} (x \wedge y)_i &= \min(x_i, y_i), & (x \wedge y)_{m+j} &= \max(x_{m+j}, y_{m+j}), \\ (x \vee y)_i &= \max(x_i, y_i), & (x \vee y)_{m+j} &= \min(x_{m+j}, y_{m+j}). \end{aligned} \tag{2.8}$$

Since $x \wedge y$, $x \vee y \in \Gamma_N$, (1.7) holds for $\Gamma_N$.

Because of (2.2), we again have (1.8) so $\Gamma_N$, $<$ is also a finite distributive lattice.

Let $Q_0^*$ be a subset of $\Gamma_N$ defined by intersections of subsets of the form $\{x : a_i(x) < a_j(x)\}$ and $\{x : b_i(x) < b_j(x)\}$, so that

$$\begin{aligned} Q_0^* = \{x : a_{i_1}(x) < a_{j_1}(x), \cdots, a_{i_r}(x) < a_{j_r}(x); \\ b_{k_1}(x) < b_{l_1}(x), \cdots, b_{k_s}(x) < b_{l_s}(x)\}, \end{aligned} \tag{2.9}$$

and let $P_1^*$ and $P_2^*$ be subsets of $\Gamma_N$ defined by intersections of the form $\{x : a_i(x) < b_j(x)\}$. Let $\mu, f, g$ be defined for $x \in \Gamma_N$ by

$$\mu(x) = \begin{cases} 1, & \text{if } x \in Q_0^*, \\ 0, & \text{else,} \end{cases}$$

$$f(x) = \begin{cases} 1, & \text{if } x \in P_1^*, \\ 0, & \text{else,} \end{cases}; \qquad g(x) = \begin{cases} 1, & \text{if } x \in P_2^*, \\ 0, & \text{else.} \end{cases} \tag{2.10}$$

If $x, y \in \Gamma_N$ and $\mu(x)\mu(y) = 1$, then $x, y \in Q_0^*$ so that for $z = x$ or $y$, $a_{i_t}(z) < a_{j_t}(z)$, $t = 1, \cdots, r$, and $b_{k_t}(z) < b_{l_t}(z)$, $t = 1, \cdots, s$. But then

$$\begin{aligned} &\min(a_{i_t}(x), a_{i_t}(y)) < \min(a_{j_t}(x), a_{j_t}(y)), \qquad t = 1, \cdots, r, \\ &\max(b_{k_t}(x), b_{k_t}(y)) < \max(b_{l_t}(x), b_{l_t}(y)), \qquad t = 1, \cdots, s, \end{aligned} \tag{2.11}$$

so that by (2.7) and (2.8), $x \wedge y \in Q_0^*$. Similarly, $x \vee y \in Q_0^*$, so that $\mu(x \wedge y)\mu(x \vee y) = 1$. Thus (1.9) holds. Note that (1.9) would fail if $Q_0^*$ were allowed to contain inequalities $\{a_i < b_j\}$.

If $x < y$ and $f(y) = 1$, then $y \in P_1^*$; so that $a_i(x) \leqq a_i(y) \leqq b_j(y) \leqq b_j(x)$ if $\{a_i < b_j\}$ is any one of the inequalities involved in $P_1^*$. Thus $x \in P_1^*$, and so $f(x) = 1$. Thus $f(x)$ is decreasing and so is $g$. Thus (1.10) holds and the hypothesis of the FKG inequality is satisfied. By (1.11) it follows that

$$\#(Q_0^* \cap P_1^* \cap P_2^*)\#(Q_0^*) \geqq \#(Q_0^* \cap P_1^*)\#(Q_0^* \cap P_2^*). \tag{2.12}$$

Now consider the subset $\Gamma_N^*$ of $\Gamma_N$, for which all $a_i$'s and $b_j$'s are distinct. Since $\#(\Gamma_N^*) = N(N-1)\cdots(N-n-m+1)$, we see that

$$\lim_{N\to\infty} \frac{\#(\Gamma_N^*)}{\#(\Gamma_N)} = 1. \tag{2.13}$$

In $\Gamma_N^*$ however, $a_1(x), \cdots, a_m(x), b_1(x), \cdots, b_n(x)$ are all distinct, and for each $x \in \Gamma_N^*$ a unique ordering of $a_1, \cdots, a_m, b_1, \cdots, b_n$ is obtained by letting the ordering of $a_i(x), b_j(x)$ determine the ordering of $a_i, b_j$. For $N \geqq n + m$, the fraction of

$\Gamma_N^*$ corresponding to each ordering of $a_1, \cdots, a_m, b_1, \cdots, b_n$ is $1/(m+n)!$; so that for $N \geqq m+n$,

$$\frac{\#(A^* \cap \Gamma_N^*)}{\#(\Gamma_N^*)} = \Pr(A). \tag{2.14}$$

Here $A = Q_0$, $Q_0 \cap P_i$, $Q_0 \cap P_1 \cap P_2$ given in Theorem 2 and in (1.5), and the corresponding $A^* = Q_0^*$, $Q_0^* \cap P_i^*$, $Q_0^* \cap P_1^* \cap P_2^*$. Since for any $A$ and the corresponding $A^*$ we have

$$\#(A^* \cap (\Gamma_N - \Gamma_N^*)) \leqq \#(\Gamma_N - \Gamma_N^*) = o(\#(\Gamma_N^*)), \tag{2.15}$$

as $N \to \infty$, it follows that also

$$\lim_{N\to\infty} \frac{\#(A^*)}{\#(\Gamma_N^*)} = \Pr(A), \tag{2.16}$$

for $A = Q_0$, $Q_0 \cap P_i$, $Q_0 \cap P_1 \cap P_2$, and $A^* = Q_0^*$, $Q_0^* \cap P_i^*$, $Q_0^* \cap P_1^* \cap P_2^*$, respectively. Thus from (2.12) and (2.16) we obtain, letting $N \to \infty$,

$$\Pr(Q_0 \cap P_1 \cap P_2) \Pr(Q_0) \geqq \Pr(Q_0 \cap P_1) \Pr(Q_0 \cap P_2), \tag{2.17}$$

which is the same as (1.6). Theorem 2 is thus proved.

## REFERENCES

[1] [FKG]. C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre, *Correlation inequalities on some partially ordered sets,* Comm. Math. Phys., 22 (1970), pp. 89–103.

[2] [GYY]. R. L. Graham, A. C. Yao, and F. F. Yao, *Some monotonicity properties of partial orders,* this Journal, this issue, pp. 251–258.

[3] [KS]. D. J. Kleitman and J. B. Schearer, *Some monotonicity properties of partial orders,* Studies in Appl. Math., to appear.

[4] [SW]. P. D. Seymour and D. J. A. Welsh, *Combinatorial applications of an inequality from statistical mechanics,* Math. Proc. Cambr. Philos. Soc., 77 (1975), pp. 485–495.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0008 $01.00/0

# AN ASYMPTOTIC SOLUTION OF INVENTORY LOT-SIZE MODELS WITH HOMOGENEOUS TIME-DEPENDENT DEMAND FUNCTIONS*

MOSHE FRIEDMAN† AND JEFFREY L. WINTER‡

**Abstract.** This paper considers a deterministic inventory lot-size model without backlogs with a continuous, time-dependent demand function of the form $kt^r$ with $k(>0)$ and $r(>-2)$ known parameters and $t(\geqq t_0 > 0)$ for time. Near closed-form expressions are developed for the asymptotic optimal replenishment times and lot sizes as the time horizon tends to infinity. These expressions are easy to calculate and give good approximations to the finite horizon problem. The results extend those of Barbosa and Friedman (Management Sci., 24 (1980), pp. 819–826), where the considerably restrictive condition $t_0 = 0$ is required. An important special case of the present work is when the demand function is affine ($r = 1$).

**1. Introduction.** This paper considers the classical deterministic inventory lot size model with a continuous homogeneous time-dependent demand function. It derives an asymptotic solution for the optimal replenishment schedule, or equivalently for the optimal lot sizes, since backlogs are prohibited. The solution is easily implementable, involves only very simple calculations, and is a very good approximation even for short time horizons.

The main thrust for taking up the problem is to solve it for the linear time-dependent demand function $kt + k_1$, with $k, k_1 > 0$ the known parameters, and $t \geqq 0$ for time. A simple transformation, i.e., $t \to t - k_1/k$, brings the linear function to the form $kt$, $t \geqq t_0 = k_1/k > 0$. The condition $t_0 > 0$ is crucial since it signifies a general linear demand function versus the time-proportional one, with $t_0 = 0$, that necessitates much simpler analysis. Since we essentially use the homogeneity property of the demand function and not its linearity, we shall develop the solution for general one-variable homogeneous functions $kt^r$, $k > 0$, $r > -2$, $t \geqq t_0 > 0$. Moreover, a general $r$ may still represent viable markets for a commodity, and give modelers greater latitude, like a vanishing market if $r < 0$, or a rapidly increasing one if $r > 1$.

The investigation of the impact of a dynamic demand function on the optimal policy in inventory and production systems has constituted a considerable share of both deterministic and stochastic mathematical inventory theory (see, for instance, Hadley and Whitin [7], Naddor [8], and Veinott [12]). A well known approach is the Wagner-Whitin [13] one, which addresses a periodic review problem with general deterministic demand. An alternative approach examines markets whose development over time can be reasonably estimated by a linear (regression) function $kt + k_1$. For convenience purposes the classical inventory lot-size model has been chosen as a framework.

Special cases of the latter problem have already been solved. The case $k = 0$, $k_1 > 0$, namely, constant demand with infinite time horizon, which in fact signifies the advent of mathematical inventory theory, was solved in 1915 yielding the celebrated Wilsonian "square root law" (see, for example, Naddor [8]). It was further extended to finite time horizons by Carr and Howe [4]. It should be emphasized that the square root law has been extensively used as a rule of thumb in a variety of situations with average demand substituting the perceived fixed demand. Sasieni et al.

* Received by the editors November 13, 1978, and in final revised form April 18, 1980.
† Operations Research Center, Bell Laboratories, Holmdel, New Jersey 07733.
‡ Department of Mathematics, Arizona State University, Tempe, Arizona.

[10] is the earliest source where the case $k > 0$, $k_1 = 0$, namely, time-proportional demand, was alluded to (see problem 2, page 80). Naddor [8, Chapt. 7], suggested elaborate solutions under restrictive conditions, and also considered a good heuristic solution which nonetheless is still not the optimal one. Resh, Friedman, and Barbosa [9] gave the optimal solution for both finite and infinite horizons and established the "cubic root law." Later on, Barbosa and Friedman [2], noting that the solution is based upon the homogeneity property of the demand function, extended it to the family $kt^r$, $k > 0$, $r > -2$, $0 \leqq t \leqq T$, of demand functions, and introduced the general "$(r + 2)$ root law," again for both finite and infinite horizons.

The real problem seems to be, however, with the demand function being off the origin, i.e., when both $k$, $k_1 > 0$, or in the alternative setting, when both $k$, $t_0 > 0$. Until very recently it remained unsolved. Donaldson [5] suggested a good attack method, and Silver [11] gave good heuristics. Barbosa and Friedman [3] established the exact solution for finite time horizons and provided computational means for calculating it. Friedman [6] extended these results to general time-dependent demand and carrying cost rate functions. It turns out, however, that the application of the solution is tedious, and that it is relatively insensitive to the length of the finite time horizon. Furthermore, both Donaldson and Silver remark that a prescribed finite time horizon is a very precarious piece of information. Hence, an asymptotic solution for the problem is of primary relevance for both the theorist and the practitioner.

The importance of the asymptotic solution is two-fold. It first provides a quick, good approximation for finite time horizons, but beyond this it should be also conceived as a rule of thumb that replaces the classical square root law. These two aspects of it will be further discussed in § 3.

Section 2 formally states the problem, while § 3 gives the solution in a concise form, illustrates its rapid convergence, and suggests a general purpose usage of it. Section 4 lists background material for finite time horizons. Sections 5 and 6 prove existence and provide computational means for finding the asymptotic solution, respectively, while § 7 discusses other closely related cases. A heavily technical proof of some uniform convergence can be found in the Appendix.

**2. The problem.** Consider a time-continuous inventory system with demand that is deterministic and of the form $b(t) = kt^r$, where $k > 0$, $r > -2$ are known constants and $t$ stands for time, $t \geqq t_0 > 0$, with $t_0$ known. We shall assume that replenishments are instantaneous and made by lot sizes, backlogs are prohibited, and hence the pertinent costs are the carrying cost $c_1$ dollars per unit per unit time and the replenishment cost $c_3$ dollars per order. Note that the shortage cost is $c_2 = \infty$.

It is easily shown that orders are made when the stock level falls to zero, and that the first order is at $t_0$. The problem is to find the optimal asymptotic replenishment schedule so as to minimize the total carrying and replenishment costs.

Mathematically the model is as follows: Let $m$ be the number of replenishments throughout the finite planning horizon $[t_0, T]$, $m = 1, 2, \cdots$. Let $t_i$, $i = 0, 1, \cdots, m - 1$, denote the replenishment times where $t_0$, $t_m \equiv T$ are known and $t_0 \leqq t_1 \leqq \cdots \leqq t_m$. The stock level at time $t \in [t_i, t_{i+1})$ is given by

$$y(t, t_{i+1}) = \int_t^{t_{i+1}} b(u)\, du, \tag{1}$$

and by definition the lot size quantity $q_i$ at time $t_i$ is $q_i = y(t_i, t_{i+1})$. Let $\tilde{Y}(t_i, t_{i+1})$ denote the total inventory carried throughout the $i$th period, namely,

$$\tilde{Y}(t_i, t_{i+1}) = \int_t^{t_{i+1}} y(t, t_{i+1})\, dt. \tag{2}$$

Hence, for an $m$-replenishment system, the total inventory is

$$Y(m, \mathbf{t}(m)) = \sum_{i=0}^{m-1} \tilde{Y}(t_i, t_{i+1}), \tag{3}$$

where $\mathbf{t}(m) = (t_1, \cdots, t_{m-1})$. For the finite time horizon case the problem is to find an integer $m$ and a vector $\mathbf{t}(m)$ so as to minimize the total cost function

$$C(m, \mathbf{t}(m)) = c_1 Y(m, \mathbf{t}(m)) + c_3 m. \tag{4}$$

Suppose $m^* = m^*(T)$ is optimal for $[t_0, T]$ and $t_i^*(m^*)$, $i = 1, \cdots, m^* - 1$, is the respective optimal replenishment schedule for $m^*$. The problem to be tackled by the manuscript is: Find $t_i^* = \lim_{T\to\infty} t_i^*(m^*)$, $i = 1, 2, \cdots$.

**3. The solution.** Let the sequence $\{\alpha_i\}$ be defined as follows: For $r \neq -1$,

$$\alpha_i = [(r + 2) - (r + 1)\alpha_{i-1}]^{-1/(r+1)}, \qquad i = 1, 2, \cdots, \tag{5}$$

and for $r = -1$,

$$\alpha_i = \exp\{-(1 - \alpha_{i-1})\}, \qquad i = 1, 2, \cdots, \tag{6}$$

for some $\alpha_0 \in [0, 1)$.

Intuitively, $\alpha_i = t_i^*/t_{i+1}^*$. The recursive relation in consecutive ratios of the optimal replenishment times is the exact place where the homogeneity property of the demand function is employed. Note that so far the particular $\alpha_0$ is still unknown.

Let the functions $R_m(\alpha_0)$, $S_m(\alpha_0)$ be

$$\begin{aligned} R_m(\alpha_0) &= \frac{\prod_{i=0}^{m-1} \alpha_i}{[\alpha_{m-1} - \alpha_{m-2}]^{1/(r+2)}}, \\ S_m(\alpha_0) &= \frac{\prod_{i=0}^{m-1} \alpha_i}{[\alpha_m - \alpha_{m-1}]^{1/(r+2)}}, \qquad m = 2, 3, \cdots, \end{aligned} \tag{7}$$

and let $L(\alpha_0)$ be

$$L(\alpha_0) = \lim_{m\to\infty} R_m(\alpha_0) = \lim_{m\to\infty} S_m(\alpha_0). \tag{8}$$

$R_m(\alpha_0)$, $S_m(\alpha_0)$ should be conceived as elaborate lower and upper bounds for the finite horizon solution. Their convergence to the same limit indicates that the finite horizon solution does the same.

Let $\alpha_0$ be the unique solution to the equation

$$L(\alpha_0) = t_0 \Big/ \left[\frac{(r + 2)c_3}{kc_1}\right]^{1/(r+2)}. \tag{9}$$

Then,

$$t_i^* = t_0 \Big/ \prod_{j=0}^{i-1} \alpha_j. \tag{10}$$

The solution is readily implementable since the sequences $\{\alpha_i\}, \prod_{j=0}^{i-1} \alpha_j$, and the function $L(\alpha_0)$ are easily tabulated.

As a matter of fact only $t_1^*$ is of practical relevance since demand will be estimated afresh after each replenishment. Consequently, only $\alpha_0$ should be calculated, via (8), from $L(\alpha_0)$ tables according to the specific parameters of the problem.

To demonstrate the rapid convergence we shall give an extreme example, for a very small $m$, considered first by Donaldson [5] and Silver [11]. Demand is $t$ over the period [6, 17] with cost parameters $c_1 = 0.1$ and $c_3 = 9$. Donaldson [5] finds that the finite horizon optimal solution is $m^* = 3$ and $t_1^*(3) \cong 10.2$. Now, the RHS of (9) is $6 \Big/ \left[\frac{3 \times 9}{1 \times 0.1}\right]^{1/3} \cong 0.9283$. By the tables, $L(\alpha_0) = 0.9283$ yields $\alpha_0 = 0.58167$. Using (10), $t_1^*(\infty) = t_0/\alpha_0 = 6/0.58167 \cong 10.3$, a very close approximation. For bigger $m^*$'s, $t_1^*(m^*)$ and $t_1^*(\infty)$ are hardly distinguishable. Notice also the lack of almost any computational effort compared to the tediousness of the iterative procedures suggested by Donaldson [5] and Silver [11].

The classical EOQ formula (which is obtained as a special case of our model by putting $r = 0$) has provided a good rule of thumb for determining the optimal lot sizes, namely, order the quantity $(2kc_3/c_1)^{1/2}$ whenever the stock level depletes to zero, where $k$ is the average demand per unit time. Intuitively, this method is not reasonable when demand reveals a strong upward trend. In this case our results put forth a simple, easily implementable, rule of thumb for ordering the optimal lot sizes. It renders insight and may be considered as a complement to Wagner and Whitin's approach. The rule is: estimate demand by a statistical linear regression technique, determine $\alpha_0$ from the tables of $L(\alpha_0)$ via (9), compute $t_1^* = t_0/\alpha_0$, and order the lot size quantity $\int_{t_0}^{t_1^*} kt\,dt = \frac{1}{2}k(t_1^{*2} - t_0^2)$. Repeat the procedure afresh when the stock level falls again to zero, etc. Observe that the method tacitly assumes that the variance around the regression function is small, to justify the deterministic approach.

We now proceed to formally prove the asymptotic solution.

**4. Review of finite time horizon results.** The proof relies heavily on the results for the finite time horizon $[t_0, T]$. This background material will be given here with some of the proofs (originally given in [3]) to make the paper self-contained.

Let $H_r$ denote the set of all homogeneous functions of two variables having a degree of homogeneity $r$.

The next lemma is embedded in the entire analysis and hence should be explicitly stated. (Background lemmas are enumerated by Roman numerals, whereas the new lemmas and theorems are enumerated by Arabic numerals.)

LEMMA I. *Let $b(t)$, $y(a_1, a_2)$ and $\tilde{Y}(a_1, a_2)$ be related as in* (1)–(2). *Then the following relations exist:*

$$\{b(t) = kt^r\} \Leftrightarrow \{y(a_1, a_2) \in H_{r+1}\} \Leftrightarrow \{\tilde{Y}(a_1, a_2) \in H_{r+2}\}. \tag{11}$$

LEMMA II. *Let the sequence $\{\alpha_i\}$ be defined by the recursion formula*

$$\frac{1}{k}\, y\left(1, \frac{1}{\alpha_i}\right) = 1 - \alpha_{i-1}, \qquad i = 1, 2, \cdots, \tag{12}$$

with $\alpha_0 \in [0, 1)$, where $y(a_1, a_2)$ is given by (1) and $b(t) = kt^r$, $k > 0$, $r > -2$. Then, $\alpha_i$ is a strictly monotone increasing concave sequence with limit 1.

Equation (12), which is a combined form for (5) and (6), is important since it spares the need to give separate proofs for different values of $r$.

COROLLARY II.1. *$\alpha_i$ is strictly increasing with $\alpha_0$.*

Let the function $\gamma_{m,i}(\alpha_0)$ be defined as

$$\gamma_{m,i}(\alpha_0) = \prod_{j=i}^{m-1} \alpha_j, \quad i = 0, 1, 2, \cdots, m-1; \; m = 1, 2, \cdots \tag{13}$$

where the sequence $\{\alpha_i\}$ is given in (12) for $\alpha_0 \in [0, 1)$, and set

$$\gamma_{m,m}(\alpha_0) = 1. \tag{14}$$

LEMMA III. *The functions $\gamma_{m,i}(\alpha_0)$ are continuous and monotonically increasing with $\alpha_0$ for every $\alpha_0 \in [0, 1)$. In particular this holds for $\gamma_{m,0}(\alpha_0)$, and note that $\gamma_{m,0}(0) = 0$, $\gamma_{m,0}(1) = 1$. ($\alpha_0 = 1$ is excluded from the analysis, however, $\gamma_{m,0}(1)$ is still defined.)*

COROLLARY III.I. *The function $\gamma_{m,0}(\alpha_0)$ has an inverse $\gamma_{m,0}^{-1}$ with domain $[0, 1)$.*

LEMMA IV. *Let $\tau \in [0, 1)$ be an arbitrary number. Then, $\alpha_0(m, \tau) = \gamma_{m,0}^{-1}(\tau)$ which solves the equation $\gamma_{m,0}(\alpha_0) = \tau$, has the following properties: $\alpha_0(m, \tau)$ is unique for a given pair $(m, \tau)$, $\alpha_0(m, 0) = 0$ for every $m$, $\alpha_0(m, \tau)$ is monotonically increasing with $m$ for every $\tau \in (0, 1)$ and bounded by 1, $\lim_{m\to\infty} \alpha_0(m, \tau) = \alpha_0(\tau)$, where $\alpha_0(0) = 0$ and $\alpha_0(\tau) = 1$ for $\tau \in (0, 1)$.*

The optimal replenishment schedule, for a given $m$, is

$$t_i^*(m, t_0, T) = \gamma_{m,i}(\alpha_0)T, \qquad i = 1, \cdots, m-1, \tag{15}$$

with $\alpha_0 \equiv \alpha_0(m, t_0/T) = \gamma_{m,0}^{-1}(t_0/T)$. Note that by the definition of $\alpha_0$, (15) holds also for $i = 0$.

Let $Y^*(m, t_0, T) = \min_{\mathbf{t}(m)} Y(m, \mathbf{t}(m))$, where $Y(m, \mathbf{t}(m))$ is given in (3), and let $\Delta Y^*(m, t_0, T) = Y^*(m, t_0, T) - Y^*(m+1, t_0, T)$. The following properties of these functions will be needed.

LEMMA V. *$Y^*(m, t_0, T)$ is strictly decreasing with $m$, $\Delta Y^*(m, t_0, T)$ is strictly increasing with $T$ and strictly decreasing with $m$, meaning that $Y^*(m, t_0, T)$ is strictly convex with $m$.*

This lemma implies that the unique optimal $m$ for $[t_0, T]$ is the first $m^*$ that satisfies $\Delta Y^*(m^*, t_0, T) \leqq c_3/c_1$.

LEMMA VI. *The explicit form of $Y^*(m, t_0, T)$ is*

$$Y^*(m, t_0, T) = \frac{k}{r+2} P_m T^{r+2}, \tag{16}$$

*with*

$$P_m = (1 - \alpha_{m-1}) - (t_0/T)^{r+2}(1 - \alpha_{-1}), \tag{17}$$

*where $\alpha_0 \equiv \alpha_0(m, t_0/T) = \gamma_{m,0}^{-1}(t_0/T)$, $\alpha_{m-1} \equiv \alpha_{m-1}(\alpha_0(m, t_0/T))$ via (12), and $1 - \alpha_{-1} \equiv (1/k)y(1, (1/\alpha_0))$ (extending (12) to include the index $-1$).*

*Proof.* Using (3), Lemma I (the homogeneity of $\tilde{Y}$) and (15) we obtain

$$Y^*(m, t_0, T) = Q_m T^{r+2}, \qquad m = 1, 2, \cdots, \tag{18}$$

where,

$$Q_m = \sum_{i=0}^{m-1} \tilde{Y}(\gamma_{m,i}(\alpha_0), \gamma_{m,i+1}(\alpha_0)), \quad m = 1, 2, \cdots, \tag{19}$$

with $\alpha_0 \equiv \alpha_0(m, t_0/T) = \gamma_{m,0}^{-1}(t_0/T)$. Noting that $\gamma_{m,m-1}(\alpha_0) = \alpha_{m-1}$ and using again the homogeneity of $\tilde{Y}$ it follows that

$$\sum_{i=0}^{m-1} \tilde{Y}(\gamma_{m,i}(\alpha_0), \gamma_{m,i+1}(\alpha_0))$$

$$= \alpha_{m-1}^{r+2} \left[ \sum_{i=0}^{m-2} \tilde{Y}(\gamma_{m-1,i}(\alpha_0), \gamma_{m-1,i+1}(\alpha_0)) + \tilde{Y}\left(1, \frac{1}{\alpha_{m-1}}\right) \right], \tag{20}$$

$$m = 2, 3, \cdots,$$

and hence

$$Q_m = \alpha_{m-1}^{r+2} \left[ Q_{m-1} + \tilde{Y}\left(1, \frac{1}{\alpha_{m-1}}\right) \right], \qquad m = 2, 3, \cdots. \tag{21}$$

Notice that $Q_m$, $Q_{m-1}$, and $\alpha_{m-1}$ in (21) are based upon the same $\alpha_0(m, t_0/T)$. If we start our calculations with $m - 1$ then $Q_{m-1}$ will be derived from $\alpha_0(m - 1, t_0/T)$, and will have a different value.

Applying to $\tilde{Y}(1, (1/\alpha_{m-1}))$ the fundamental theorem of the homogeneous functions, namely,

$$\frac{\partial \tilde{Y}(a_1, a_2)}{\partial a_1} a_1 + \frac{\partial \tilde{Y}(a_1, a_2)}{\partial a_2} a_2 = (r + 2)\tilde{Y}(a_1, a_2),$$

and employing the fact that

$$\frac{\partial \tilde{Y}(a_1, a_2)}{\partial a_1} = -y(a_1, a_2) \text{ and } \frac{\partial \tilde{Y}(a_1, a_2)}{\partial a_2} = b(a_2)(a_2 - a_1)$$

with (12), yields

$$Q_m - \frac{k}{r+2}(1 - \alpha_{m-1})$$

$$= \alpha_{m-1}^{r+2} \left[ Q_{m-1} - \frac{k}{r+2}(1 - \alpha_{m-2}) \right], \qquad m = 2, 3, \cdots. \tag{22}$$

Applying (22) repeatedly implies

$$Q_m - \frac{k}{r+2}(1 - \alpha_{m-1})$$

$$= \prod_{i=1}^{m-1} \alpha_i^{r+2} \left[ Q_1 - \frac{k}{r+2}(1 - \alpha_0) \right], \qquad m = 2, 3, \cdots, \tag{23}$$

with

$$Q_1 = \tilde{Y}(\alpha_0, 1) = \int_0^1 \int_t^1 k u^r \, du \, dt, \tag{24}$$

and $\alpha_0 \equiv \alpha_0(m, t_0/T) = \gamma_{m,0}^{-1}(t_0/T)$. Observing, via (13), that

$$\prod_{i=1}^{m-1} \alpha_i^{r+2} = \frac{\gamma_{m,0}^{r+2}(\alpha_0)}{\alpha_0^{r+2}} = \frac{(t_0/T)^{r+2}}{\alpha_0^{r+2}},$$

noting (by the homogeneity of $\tilde{Y}(1, (1/\alpha_0))$ that $Q_1/\alpha_0^{r+2} = \tilde{Y}(1, (1/\alpha_0)) = -y(1, (1/\alpha_0))/(r + 2) + k(1 - \alpha_0)/(r + 2)\alpha_0^{r+2}$ and using (12) to define $1 - \alpha_{-1}$ completes the proof. QED

Equation (17) reads explicitly: For $r \neq -1$,

$$P_m = (1 - \alpha_{m-1}) - \frac{(t_0/T)^{r+2}}{r+1}\left(\frac{1}{\alpha_0^{r+1}} - 1\right) \tag{25}$$

and for $r = -1$,

$$P_m = (1 - \alpha_{m-1}) - (t_0/T)\ln(1/\alpha_0). \tag{26}$$

The unique optimal $m^*$ is determined by

$$P_{m^*} - P_{m^*+1} \leqq \frac{(r+2)c_3}{kc_1 T^{r+2}} \leqq P_{m^*-1} - P_{m^*}, \tag{27}$$

and the optimal replenishment schedule is given by (15) for $m = m^*$.

**5. Existence of the asymptotic solution.** In this section we prove that the asymptotic solution exists. This will be shown by proving that $\lim_{T\to\infty} \alpha_0(m^*(T), t_0/T) = a_0 > 0$ exists. Eq. (15), with $m = m^*$, implies that $t_i^*(m^*, t_0, T)/t_{i+1}^*(m^*, t_0, T) = \alpha_i(m^*, t_0/T)$. If $\alpha_0(m^*, t_0/T)$ converges to a strictly positive limit $\alpha_0$, then $t_1^* = t_0/\alpha_0$ exists. The limiting $\alpha_i$'s, $i = 1, 2, \cdots$, are computed via (12), with the limiting $\alpha_0$, due to the continuity, and the limiting $t_i^*$'s are sequentially generated by $t_i^* = t_{i-1}^*/\alpha_{i-1}$. Hence, if a limiting positive $\alpha_0$ exists the entire asymptotic solution also exists.

The next lemma guarantees that there is an optimal $m^*$ for any $c_3/c_1$, as small as it can be.

LEMMA 1. *Let* $Y^*(m, t_0, T)$ *and* $\Delta Y^*(m, t_0, T)$ *be defined as in* § 4. *Then,*

$$\lim_{m\to\infty} \Delta Y^*(m, t_0, T) = 0. \tag{28}$$

*Proof.* Since $Y^*(m, t_0, T)$ decreases with $m$, $\Delta Y^*(m, t_0, T) > 0$. $Y^*(1, t_0, T)$ exists and $\sum_{i=1}^m \Delta Y^*(i, t_0, T) = Y^*(1, t_0, T) - Y^*(m+1, t_0, T)$ is increasing with $m$ and bounded from above by $Y^*(1, t_0, T)$. It follows that $\lim_{m\to\infty} \sum_{i=1}^m \Delta Y^*(i, t_0, T)$ exists and hence $\lim_{m\to\infty} \Delta Y^*(m, t_0, T) = 0$. QED

The following lemma lists some properties of the sequences $\alpha$ that will be needed later on.

LEMMA 2.

a. *For a fixed* $m$, $\alpha_0(m, t_0/T)$ *is decreasing with* $T$, *and* $\alpha_0(m, t_0/T) \equiv \gamma_{m,0}^{-1}(t_0/T)_{T\to\infty} \to 0$.

b. $\lim_{T\to\infty} (\alpha_m(m+1, t_0/T) - \alpha_{m-1}(m, t_0/T)) > 0$, *where* $\alpha_m(m+1, t_0/T)$, $\alpha_{m-1}(m, t_0/T)$ *are computed via* (12) *for* $\alpha_0(m+1, t_0/T)$, $a_0(m, t_0/T)$, *respectively.*

c. $\alpha_0(m, t_0/T) \geqq t_0/T$.

d. $\alpha_0(m+1, t_0/T) < (t_0/T)(\prod_{i=1}^m \alpha_i)^{-1}$, *where the* $\alpha_i$ *are computed via* (12) *for* $\alpha_0 = 0$.

*Proof.*

a. Follows from the definition of $\alpha_0(m, t_0/T)$ and Lemma II.

b. Let $\alpha_m$ be computed by (12) for $\alpha_0 = 0$. By Corollary II.I, $\alpha_m < \alpha_m(m+1, t_0/T)$ for every $T$, by a we can choose $T$ large enough such that $\alpha_{m-1}(m, t_0/T) < \alpha_m$, and combined we can choose $T$ large enough so that $\alpha_{m-1}(m, t_0/T) < \alpha_m < \alpha_m(m+1, t_0/T)$. Since $\alpha_{m-1}(m, t_0/T)$ decreases with $T$ the result follows.

c. Follows from the definition of $\alpha_0(m, t_0/T)$ and the fact that $\alpha_i(m, t_0/T) < 1$.

d. By the definition, $\alpha_0(m+1, t_0/T) = (t_0/T)(\prod_{i=1}^m \alpha_i(m+1, t_0/T)^{-1}$. Since, by Corollary II.I and a, $\alpha_i < \alpha_i(m+1, t_0/T)$, the result follows.

The following lemma is needed for determining the behavior of $m^*$ as a function of $T$.

LEMMA 3. *Let* $Y^*(m, t_0, T)$ *and* $\Delta Y^*(m, t_0, T)$ *be defined as in* § 4. *Then,*

$$\lim_{T\to\infty} \Delta Y^*(m, t_0, T) = \infty. \tag{29}$$

*Proof.* Consider first the case $r \neq -1$. Using (16) and (25), $\Delta Y^*(m, t_0, T)$ is

$$\Delta Y^*(m, t_0, T) = \frac{k}{r+2}\left[ T^{r+2}(\alpha_m(m+1, t_0/T) - \alpha_{m-1}(m, t_0/T)) - \frac{t_0^{r+2}}{r+1}\left(\frac{1}{(\alpha_0(m, t_0/T))^{r+1}} - \frac{1}{(\alpha_0(m+1, t_0/T))^{r+1}}\right)\right]. \tag{30}$$

Using Lemma 2 we shall now find a lower bound to $\Delta Y^*(m, t_0, T)$.

By Part c of Lemma 2,

$$\frac{1}{(\alpha_0(m, t_0/T)^{r+1}} \begin{cases} \leqq (T/t_0)^{r+1}, & r > -1, \\ \geqq (T/t_0)^{r+1}, & -2 < r < -1. \end{cases} \tag{31}$$

By Part d of Lemma 2,

$$\frac{1}{(\alpha_0(m+1, t_0/T))^{r+1}} \begin{cases} > (T/t_0)^{r+1}\left(\prod_{i=1}^{m} \alpha_i\right)^{r+1}, & r > -1, \\ < (T/t_0)^{r+1}\left(\prod_{i=1}^{m} \alpha_i\right)^{r+1}, & -2 < r < -1. \end{cases} \tag{32}$$

Combining (31) and (32) yields

$$\frac{1}{(\alpha_0(m, t_0/T))^{r+1}} - \frac{1}{(\alpha_0(m+1, t_0/T))^{r+1}} \begin{cases} < (T/t_0)^{r+1}\left(1 - \left(\prod_{i=1}^{m} \alpha_i\right)^{r+1}\right), & r > -1 \\ > (T/t_0)^{r+1}\left(1 - \left(\prod_{i=1}^{m} \alpha_i\right)^{r+1}\right), & -2 < r < -1. \end{cases} \tag{33}$$

Employing (33) in (30) implies

$$\Delta Y^*(m, t_0, T) > \frac{k}{r+2}\left[T^{r+2}(\alpha_m(m+1, t_0/T) - \alpha_{m-1}(m, t_0/T)) - T^{r+1}\left(\frac{t_0}{r+1}\left(1 - \left(\prod_{i=1}^{m} \alpha_i\right)^{r+1}\right)\right)\right]. \tag{34}$$

Notice that the inequality in (34) holds for both $r > -1$ and $-2 < r < -1$ since the term $t_0^{r+2}/(r+1)$ takes care of the sign.

Now, the first term in the square brackets in the right-hand side of (34) tends to infinity on order of $T^{r+2}$, since by Part a of Lemma 2, $\alpha_m(m+1, t_0/T) - \alpha_{m-1}(m, t_0/T)$ is bounded away from zero. The second term tends to infinity on order of $T^{r+1}$. Therefore, the difference, and thus $\Delta Y^*(m, t_0, T)$, tends to infinity with $T$.

Take now the case $r = -1$. By (16) and (26), $\Delta Y^*(m, t_0, T)$ is

$$\Delta Y^*(m, t_0, T) = k[T(\alpha_m(m+1, t_0/T) - \alpha_{m-1}(m, t_0/T)) - t_0(\ln(1/\alpha_0(m, t_0/T)) - \ln(1/\alpha_0(m+1, t_0/T)))]. \tag{35}$$

We shall again bound $\Delta Y^*(m, t_0, T)$ from below. By Part c of Lemma 2,

$$\ln (1/\alpha_0(m, t_0/T)) \leqq \ln (T/t_0). \tag{36}$$

By Part d of Lemma 2,

$$\ln (1/\alpha_0(m + 1, t_0/T)) > \ln \left( (T/t_0) \left( \prod_{i=1}^{m} \alpha_i \right) \right). \tag{37}$$

Combining (36) and (37) yields

$$\ln (1/\alpha_0(m, t_0/T)) - \ln (1/\alpha_0(m + 1, t_0/T)) < -\ln \left( \prod_{i=1}^{m} \alpha_i \right). \tag{38}$$

Employing (38) in (35) implies

$$\Delta Y^*(m, t_0, T) > k \left[ T(\alpha_m(m + 1, t_0/T) - \alpha_{m-1}(m, t_0/T)) + t_0 \ln \left( \prod_{i=1}^{m} \alpha_i \right) \right]. \tag{39}$$

Since, by Part a of Lemma 2, $\alpha_m(m + 1, t_0/T) - \alpha_{m-1}(m, t_0/T)$ is bounded away from zero, the lower bound in (39), and thus $\Delta Y^*(m, t_0, T)$, tends to infinity with $T$.

COROLLARY 3.1. *The optimal $m$, $m^* = m^*(T)$, is an increasing, unbounded step function of $T$.*

*Proof.* Follows from the definition of $m^*$, Lemma V and Lemma 3.

The next theorem guarantees the existence of the asymptotic solution.

THEOREM 1. *Let the unique $m^* = m^*(T)$ be determined by* (27), *and let* $\alpha_0(m^*(T), t_0/T) = \gamma^{-1}_{m^*,0}(t_0/T)$. Then

$$\lim_{T\to\infty} \alpha_0(m^*(T), t_0/T) = \alpha_0 > 0. \tag{40}$$

*Proof.* By Corollary 3.1 there exists a closed interval $[T'_m, T''_m]$ such that $m$ is optimal for $[t_0, T]$ for each $T \in [T'_m, T''_m]$. Observe that $T''_m = T'_{m+1}$, meaning that at this point both $m$ and $m + 1$ are optimal. This happens when the left inequality in (27) is satisfied as an equality. Let $\alpha'_0(m) \equiv \alpha_0(m, t_0/T'_m)$ and $\alpha''_0(m) \equiv \alpha_0(m, t_0/T''_m)$. By Part a of Lemma 2, for every $T \in [T'_m, T''_m]$

$$\alpha''_0(m) \leqq \alpha_0(m, t_0/T) \leqq \alpha'_0(m). \tag{41}$$

It therefore suffices to show that both $\alpha'_0(m)$, $\alpha''_0(m)$ converge to the same limit; as $T$, and thus $m$, by Corollary 3.1, tend to infinity.

Since $m + 1$ is optimal for $T'_{m+1}$, it follows by Bellman's Principle of Optimality, that $m$ is optimal for $t^*_m(m + 1, t_0, T'_{m+1})$, and thus $t^*_m(m + 1, t_0, T'_{m+1}) \in [T'_m, T''_m]$. By the definition of $\alpha_0(m, \tau)$, (13), and (15),

$$\begin{aligned} \alpha'_0(m + 1) &\equiv \alpha_0(m + 1, t_0/T'_{m+1}) \\ &= \frac{t_0}{\gamma_{m+1,1}(\alpha_0)T'_{m+1}} = \frac{t_0}{\gamma_{m,1}(\alpha_0)\alpha_m(\alpha_0)T'_{m+1}} \\ &= \frac{t_0}{\gamma_{m,1}(\alpha_0)\gamma_{m+1,m}(\alpha_0)T'_{m+1}} = \frac{t_0}{\gamma_{m,1}(\alpha_0)t^*_m(m + 1, t_0, T'_{m+1})} \\ &= \alpha_0(m, t_0/t^*_m(m + 1, t_0, T'_{m+1})). \end{aligned} \tag{42}$$

Combining (41) and (42), we obtain that

$$\alpha'_0(m + 1) \leqq \alpha'_0(m). \tag{43}$$

Similarly, $\alpha''(m + 1) \equiv \alpha_0(m + 1, t_0/T''_{m+1}) = \alpha_0(m, t_0/t^*_m(m + 1, t_0, T''_{m+1}))$. Since,

again by Bellman's Optimality Principle, $t_m^*(m+1, t_0, T''_{m+1}) \in [T'_m, T''_m]$, it follows that

$$\alpha_0''(m) \leqq \alpha_0''(m+1). \tag{44}$$

Combining (41), (43), and (44) yields

$$\alpha_0''(m) \leqq \alpha_0''(m+1) \leqq \alpha_0'(m+1) \leqq \alpha_0'(m). \tag{45}$$

(45) readily implies that $\lim_{m\to\infty} \alpha_0'(m) \equiv \alpha_0'$, and $\lim_{m\to\infty} \alpha_0''(m) \equiv \alpha_0''$ exist, since $\alpha_0'(m)$ and $\alpha_0''(m)$ are monotone bounded sequences, but not necessarily that $\alpha_0' = \alpha_0''$.

Suppose $\alpha_0' \neq \alpha_0''$. Since $\alpha_0'' < \alpha_0'$, because of (41), Lemma II and Corollary II.I imply that $\prod_{j=0}^{i-1} \alpha_j'' < \prod_{j=0}^{i-1} \alpha_j'$, or alternatively, that

$$\prod_{j=0}^{i-1} (\alpha_j'/\alpha_j'') > 1, \tag{46}$$

where $\alpha_j'$, $\alpha_j''$ are computed via (12) for $\alpha_0'$, $\alpha_0''$, respectively. Since $\alpha_j'/\alpha_j'' > 1$, the product in (46) increases with $i$. Recall also that $\alpha_j'$ converges to 1 with $j$. It is thus possible to choose $m$ large enough such that $\alpha_m' \prod_{j=0}^{m-1} (\alpha_j'/\alpha_j'') > 1$, or alternatively, that

$$\prod_{j=0}^{m} \alpha_j' > \prod_{j=0}^{m-1} \alpha_j''. \tag{47}$$

Employing (43) and (44) yields

$$\prod_{j=0}^{m} \alpha_j'(m+1) > \prod_{j=0}^{m-1} \alpha_j''(m). \tag{48}$$

By (13) and (15),

$$t_0 \Big/ \prod_{j=0}^{m} \alpha_j'(m+1) = T'_{m+1} \quad \text{and} \quad t_0 \Big/ \prod_{j=0}^{m-1} \alpha_j''(m) = T''_m.$$

(48) then implies that $T''_m > T'_{m+1}$, which contradicts Corollary 3.1. Hence, $\alpha_0' = \alpha_0'' = \alpha_0$, which is obviously positive.

The convergence of $\alpha_0'(m)$, $\alpha_0''(m)$ to $\alpha_0$ is illustrated in Fig. 1.

So far we have established the existence of the asymptotic solution. We do not have as yet any computational means to calculate it, however. This; namely, the determination of $\alpha_0$; is the subject matter of the next section.

**6. Calculation of the asymptotic solution.** Isolating $T$ in (27), using (17) to put lower and upper bounds on $T$, and using these bounds in (15), for $i = 1$, implies

$$\begin{aligned} &\left[\frac{(r+2)c_3}{kc_1} + t_0^{r+2}(\alpha_{-1}(m^*) - \alpha_{-1}(m^*-1))\right]^{1/(r+2)} f_{m^*}(\alpha_0(m^*), \alpha_0(m^*-1)) \\ &\leqq (t_1^*(m^*, t_0, T) \\ &\leqq \left[\frac{(r+2)c_3}{kc_1} + t_0^{r+2}(\alpha_{-1}(m^*+1) - \alpha_{-1}(m^*))\right]^{1/(r+2)} \\ &\qquad \cdot g_{m^*}(\alpha_0(m^*+1), \alpha_0(m^*)); \end{aligned} \tag{49}$$

here

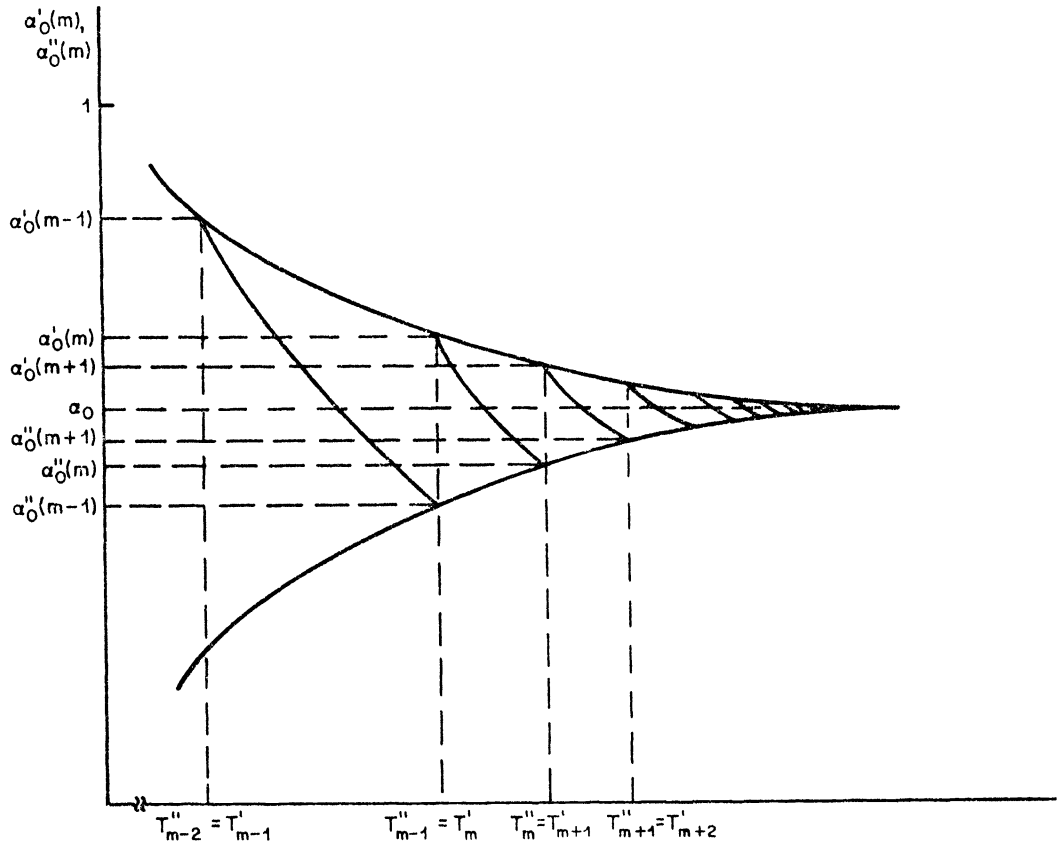



FIG. 1. *The convergence of* $\alpha_0'(m)$, $\alpha_0''(m)$ *to* $\alpha_0$.

$$(50)\qquad f_m(\alpha_0, \beta_0) = \frac{\prod_{i=1}^{m-1} \alpha_i}{[\alpha_{m-1} - \beta_{m-2}]^{1/(r+2)}}, \qquad g_m(\alpha_0, \beta_0) = \frac{\prod_{i=1}^{m-1} \beta_i}{[\alpha_m - \beta_{m-1}]^{1/(r+2)}},$$

are defined on the compact region $0 \leqq \beta_0 \leqq \alpha_0 \leqq 1 - \varepsilon$, for $\varepsilon > 0$, and $\alpha_i$, $\beta_i$ are defined via (12) for $\alpha_0$, $\beta_0$, respectively. Note that $\alpha_{m^*} = \alpha_{m^*}(\alpha_0(m^* + 1, t_0/T))$, $\alpha_{m^*-1} = \alpha_{m^*-1}(\alpha_0(m^*, t_0/T))$, $\alpha_{m^*-2} = \alpha_{m^*-2}(\alpha_0(m^* - 1, t_0/T))$, $\alpha_{-1}(m^* + 1) = \alpha_{-1}(\alpha_0(m^* + 1, t_0/T))$, $\alpha_{-1}(m^*) = \alpha_{-1}(\alpha_0(m^*, t_0/T))$, $\alpha_{-1}(m^* - 1) = \alpha_{-1}(\alpha_0(m^* - 1, t_0/T))$. Since from now on we shall deal only with $m^* = m^*(T)$ that is optimal on $[t_0, T]$, for the sake of notational convenience the dependence of $T$ and the "*" will be suppressed.

Letting $T$ tend to infinity in (49), while employing Corollary 3.1, Theorem 1 and the continuity in (49) and (12), implies

$$(51)\qquad \lim_{m\to\infty} f_m(\alpha_0(m), \alpha_0(m-1)) \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} \leqq t_1^* \leqq \lim_{m\to\infty} g_m(\alpha_0(m+1), \alpha_0(m)) \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)},$$

provided that the limits in (51) exist. We shall now proceed to prove not only that these limits exist but that they are also equal.

The limits of $f_m(\alpha_0(m), \alpha_0(m-1))$, $g_m(\alpha_0(m+1), \alpha_0(m))$ are obfuscated by the intricate dependence on $m$. The functions themselves as well as the arguments depend on $m$. Since we know already that $\alpha_0(m)$ converges, uniform convergence of $f_m(\alpha_0, \beta_0)$, $g_m(\alpha_0, \beta_0)$ of (50) will allow us to write $\lim_{m\to\infty} f_m(\alpha_0(m), \alpha_0(m-1)) = \lim_{m\to\infty} f_m(\alpha_0, \alpha_0)$, where $\alpha_0 = \lim_{m\to\infty} \alpha_0(m-1) = \lim_{m\to\infty} \alpha_0(m)$, and the same for $g_m$ (see, for instance, Apostol [1]).

The following lemma guarantees the uniform convergence of $f_m(\alpha_0, \beta_0)$, $g_m(\alpha_0, \beta_0)$ of (50).

LEMMA 4. *Let the function $h(x)$ be given by*

$$\frac{1}{k}\, y(1, 1/h(x)) = 1 - x, \tag{52}$$

*for $x \in [0, 1)$, where $y(a_1, a_2)$ is given by* (1) *and $b(t) = kt^r$, $k > 0$, $r > -2$, $t \geqq t_0 > 0$. Let $\alpha_i = h(\alpha_{i-1})$, $\beta_i = h(\beta_{i-1})$ and let $f_m(\alpha_0, \beta_0)$, $g_m(\alpha_0, \beta_0)$ be given by* (50) *for the compact region $0 \leqq \beta_0 \leqq \alpha_0 \leqq 1 - \varepsilon$ for $\varepsilon > 0$. Then, $f_m$ and $g_m$ converge uniformly as $m$ tends to infinity.*

The proof of Lemma 4 is given in the Appendix.

The investigation culminates in the following theorem.

THEOREM 2. *Let $\alpha_0 = \lim_{m\to\infty} \alpha_0(m)$, and $f(\alpha_0, \beta_0) = \lim_{m\to\infty} f_m(\alpha_0, \beta_0)$, $g(\alpha_0, \beta_0) = \lim_{m\to\infty} g_m(\alpha_0, \beta_0)$, where $f_m(\alpha_0, \beta_0)$, $g_m(\alpha_0, \beta_0)$ are given in* (50). *Then,*

$$t_1^* = \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} f(\alpha_0, \alpha_0) = \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} g(\alpha_0, \alpha_0). \tag{53}$$

*Proof.* By (51) and Lemma 4,

$$\left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} f(\alpha_0, \alpha_0) \leqq t_1^* \leqq \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} g(\alpha_0, \alpha_0), \tag{54}$$

where $\alpha_0 = \lim_{m\to\infty} \alpha_0(m-1) = \lim_{m\to\infty} \alpha_0(m) = \lim_{m\to\infty} \alpha_0(m+1)$. It is left to show that $f(\alpha_0, \alpha_0) = g(\alpha_0, \alpha_0)$. Take the ratio

$$\begin{aligned} \frac{f_m(\alpha_0, \alpha_0)}{g_m(\alpha_0, \alpha_0)} &= \left[\frac{\alpha_m - \alpha_{m-1}}{\alpha_{m-1} - \alpha_{m-2}}\right]^{1/(r+2)} \\ &= \left[\frac{h(\alpha_{m-1}) - h(\alpha_{m-2})}{\alpha_{m-1} - \alpha_{m-2}}\right]^{1/(r+2)} \underset{m\to\infty}{\to} (h'(1))^{1/(r+2)} = h(1) = 1, \end{aligned} \tag{55}$$

and the result is obtained. QED.

Combining (53) and $t_1^* = t_0/\alpha_0$ yields the following equation for $\alpha_0$,

$$\alpha_0 f(\alpha_0, \alpha_0) = t_0 \Big/ \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)}. \tag{56}$$

Notice that the function $\alpha_0 f(\alpha_0, \alpha_0)$ vanishes for $\alpha_0 = 0$ and is strictly increasing to infinity as $\alpha_0$ approaches 1. Consequently, (56) has always a unique solution.

The best way to summarize the results is given in § 3. The convergence rate of the asymptotic solution as well as its usage as a general rule of thumb were also discussed in § 3.

**7. Discussion.** This paper is basically written for linear functions with a positive slope, and for inventory settings. If the regression line has a negative slope, i.e., $k|t - T|^r$, or the limits of the integral in (1) are switched to read $\int_t^{t_i} b(u)\, du$, to repre-

sent a waste disposal case for example (the mirror image of the currently treated system), the solution looks as follows. For $r \neq -1$, let

$$\delta_i = \frac{r+1}{r+2-\delta_{i-1}^{r+1}}, \qquad i = 1, 2, \cdots \tag{57}$$

and for $r = -1$, let

$$\delta_i = \exp\{1/\delta_{i-1} - 1\}, \qquad i = 1, 2, \cdots. \tag{58}$$

Define

$$\Psi_{m,i}(\delta_0) = \prod_{j=i}^{m-1} \delta_j, \qquad i = 0, 1, \cdots, m-1;\ m = 1, 2, \cdots \tag{59}$$

with $\Psi_{m,m}(\delta_0) \equiv 1$. Let

$$\delta_0 = \delta_0(m, t_0/T) = \Psi_{m,0}^{-1}(t_0/T), \tag{60}$$

and

$$P_m = (1/\delta_m - 1) - (t_0/T)^{r+2}(1/\delta_0 - 1), \quad m = 1, 2, \cdots. \tag{61}$$

The optimal $m$ is uniquely determined by

$$P_m - P_{m+1} \leqq \frac{(r+2)c_3}{kc_1 T^{r+2}} \leqq P_{m-1} - P_m, \tag{62}$$

and the optimal times are

$$t_i = \Psi_{m,i}(\delta_0)T, \qquad i = 1, 2, \cdots, m-1 \tag{63}$$

where $m$ is optimal.

The asymptotic solution of this problem will be identical, mutatis mutandis, to that of the previous one. Theorem 2, for instance, will read

$$\begin{aligned} t_1^* &= \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} \lim_{m\to\infty} \left(\frac{1}{\delta_{m-1}} - \frac{1}{\delta_m}\right)^{1/(r+2)} \prod_{i=1}^{m-1} \delta_i \\ &= \left[\frac{(r+2)c_3}{kc_1}\right]^{1/(r+2)} \lim_{m\to\infty} \left(\frac{1}{\delta_m} - \frac{1}{\delta_{m+1}}\right)^{1/(r+2)} \prod_{i=1}^{m-1} \delta_i, \end{aligned} \tag{64}$$

where $\delta_0 = \lim_{m\to\infty} \delta_0(m)$. Tables may be easily produced also here.

The solution relies heavily on the homogeneity of the function $kt^r$, namely, on the utilization of the sequence $\alpha_i = t_i/t_{i+1}$. Note that the convergence of $\alpha_i$ to 1 is not retained if $r \leqq -2$, and consequently $m^*$ is bounded when $T$ tends to infinity. This makes the analysis for infinite horizons uninteresting.

Asymptotic solutions or algorithms for a general demand function $b(t)$ are yet unavailable, and are currently under study by the authors.

**Appendix.**

*Proof of Lemma 4.* From its first two derivatives it follows that $h(x)$ of (52) is strictly increasing and convex and $h'(x) = (h(x))^{r+2}$. Therefore, $h'(x_2)(x_2 - x_1) > h(x_2) - h(x_1)$ for $x_2 > x_1$. It follows that $\alpha_m^{r+2}(\alpha_{m-1} - \beta_{m-2}) > \alpha_m - \beta_{m-1}$ or $\alpha_m(\alpha_{m-1} - \beta_{m-2})^{1/(r+2)}/(\alpha_m - \beta_{m-1})^{1/(r+2)} > 1$. Thus,

$$f_{m+1}(\alpha_0, \beta_0) = f_m(\alpha_0, \beta_0) \frac{\alpha_m(\alpha_{m-1} - \beta_{m-2})^{1/(r+2)}}{(\alpha_m - \alpha_{m-1})^{1/(r+2)}} > f_m(\alpha_0, \beta_0). \tag{65}$$

By the definition, $f_m(\alpha_0, \beta_0) \leqq f_m(\alpha_0, \alpha_0)$. It was proven in [2] that $\lim_{m\to\infty} f_m(0, 0)$ exists. We shall now show that $\lim_{m\to\infty} f_m(\alpha_0, \alpha_0)$ exists for any $\alpha_0 \in [0, 1)$. Let $\alpha_i'$, $\alpha_i$ be computed for the given $\alpha_0$ and $\alpha_0' = 0$, respectively. Since $\lim_{i\to\infty} \alpha_i' = 1$, it is possible to choose $i$ such that $\alpha_1 < \alpha_i'$. Then, $\alpha_{j+1} < \alpha_{j+i}'$ by Corollary II.I. Consequently, $\prod_{j=1}^{m} \alpha_j < \prod_{j=i}^{m+i-1} \alpha_j'$, and also $\alpha_m - \alpha_{m-1} > \alpha_{m+i-1}' - \alpha_{m+i-2}'$, since $\{\alpha_i\}$ is concave by Lemma II. Hence,

$$\frac{\prod_{j=1}^{m} \alpha_j}{[\alpha_m - \alpha_{m-1}]^{1/(r+2)}} < \frac{\prod_{j=i}^{m+i-1} \alpha_j'}{[\alpha_{m+i-1}' - \alpha_{m+i-2}']^{1/(r+2)}}, \tag{66}$$

which yields

$$f_{m+1}(\alpha_0, \alpha_0) < f_{m+i}(0, 0) \left(\prod_{j=1}^{i-1} \alpha_j'\right)^{-1}. \tag{67}$$

$f_m(\alpha_0, \alpha_0)$ is thus bounded from above by a convergent sequence. By (65), $f_m(\alpha_0, \alpha_0)$ is also increasing with $m$, so $\lim_{m\to\infty} f_m(\alpha_0, \alpha_0)$ exists. It follows that $f_m(\alpha_0, \beta_0)$ increases with $m$ and is bounded from above by a convergent sequence, and hence $\lim_{m\to\infty} f_m(\alpha_0, \beta_0)$ exists and equals $f(\alpha_0, \beta_0)$, say.

We still have to prove that the convergence is uniform. For this end we shall prove that $f(\alpha_0, \beta_0)$ is continuous. This will be accomplished by considering the partial derivatives of $f_m(\alpha_0, \beta_0)$.

$$\begin{aligned}\frac{\partial f_m(\alpha_0, \beta_0)}{\partial \alpha_0} = &- \frac{1}{r+2} (\alpha_{m-1} - \beta_{m-2})^{-(r+3)/(r+2)} \frac{d\alpha_{m-1}}{d\alpha_0} \prod_{i=1}^{m-1} \alpha_i \\ &+ (\alpha_{m-1} - \beta_{m-2})^{-1/(r+2)} \sum_{i=1}^{m-1} \frac{d\alpha_i}{d\alpha_0} \prod_{j\neq i} \alpha_j.\end{aligned} \tag{68}$$

Employing (12), (52), and $h'(x) = (h(x))^{r+2}$ brings about $d\alpha_i/d\alpha_{i-1} = \alpha_i^{r+2}$. Applying the chain rule and (50) implies

$$\frac{d\alpha_i}{d\alpha_0} = \prod_{j=1}^{i} \alpha_j^{r+2} = (f_{i+1}(\alpha_0, \beta_0))^{r+2}(\alpha_i - \beta_{i-1}), \tag{69}$$

which, substituted in (68), yields

$$\frac{\partial f_m(\alpha_0, \beta_0)}{\partial \alpha_0} = - \frac{1}{r+2} (f_m(\alpha_0, \beta_0))^{r+3} + f_m(\alpha_0, \beta_0) \sum_{i=1}^{m-1} \frac{d\alpha_i/d\alpha_0}{\alpha_i}. \tag{70}$$

By (69), $d\alpha_i/d\alpha_0$ decreases with $i$, because $\alpha_j^{r+2} < 1$, and goes to zero since $\lim_{i\to\infty} (f_{i+1}(\alpha_0, \beta_0))^{r+2}$ exists and $\lim_{i\to\infty} (\alpha_i - \beta_{i-1}) = 0$. $\alpha_i$ increases with $i$, so $(d\alpha_i/d\alpha_0)/\alpha_i$ decreases with $i$. To show that $\sum_{i=1}^{\infty} (d\alpha_i/d\alpha_0)/\alpha_i$ exists, define the function

$$\begin{aligned}r(x) = \frac{d\alpha_i/d\alpha_0}{\alpha_i} + (x - i) \left(\frac{d\alpha_{i+1}/d\alpha_0}{\alpha_{i+1}} - \frac{d\alpha_i/d\alpha_0}{\alpha_i}\right), \\ x \in [i, i+1),\ i = 1, 2, \cdots.\end{aligned} \tag{71}$$

Since $(d\alpha_i/d\alpha_0)/\alpha_i$ decreases and goes to zero with $i$, $r(i) = (d\alpha_i/d\alpha_0)/\alpha_i$, and $r(x)$ is linear on $[i, i+1)$, it follows that $\lim_{x\to\infty} r(x) = 0$. Let

$$s(x) = \alpha_i \exp\left\{\int_i^x r(u)\, du\right\}, \qquad x \in [i - \tfrac{1}{2}, i + \tfrac{1}{2}),\ i = 1, 2, \cdots.$$

Since $\lim_{x\to\infty} r(x) = 0$, also $\lim_{x\to\infty} \int_i^x r(u)\,du = 0$, where $x \in [i - \frac{1}{2}, i + \frac{1}{2})$. Thus, $\lim_{x\to\infty} s(x) = \lim_{i\to\infty} \alpha_i = 1$. Note also that $s'(x)/s(x) = r(x)$. Finally,

$$\sum_{i=2}^{\infty} \frac{d\alpha_i/d\alpha_0}{\alpha_i} < \int_1^{\infty} r(x)\,dx = \int_1^{\infty} (s'(x)/s(x))\,dx = [\ln s(x)]_1^{\infty} = \ln 1 - \ln s(1) = -\ln \alpha_1. \tag{72}$$

Eq. (72) asserts the finiteness of $\sum_{i=1}^{\infty}(d\alpha_i/d\alpha_0)/\alpha_i$. This implies, via (70), that $\lim_{m\to\infty} (\partial f_m(\alpha_0, \beta_0)/\partial\alpha_0)$ is finite and in particular

$$\left|\frac{\partial f_m(\alpha_0, \beta_0)}{\partial \alpha_0}\right| < M(\alpha_0, \beta_0), \qquad m = 1, 2, \cdots, \tag{73}$$

for some constant $M(\alpha_0, \beta_0)$. In a similar manner it can be shown that also

$$\left|\frac{\partial f_m(\alpha_0, \beta_0)}{\partial \beta_0}\right| < K(\alpha_0, \beta_0), \qquad m = 1, 2, \cdots, \tag{74}$$

for some constant $K(\alpha_0, \beta_0)$. By the continuity of the upper bound in (73) with respect to $(\alpha_0, \beta_0)$, it may be easily deduced that both partial derivatives of $f_m(\alpha_0, \beta_0)$ are uniformly bounded for all $m$ and $(\alpha_0, \beta_0)$. This immediately implies (see Apostol [1]) that a Lipschitz condition holds for all $m$ and hence that $\lim_{m\to\infty} f_m(\alpha_0, \beta_0) = f(\alpha_0, \beta_0)$ is continuous with respect to $(\alpha_0, \beta_0)$. Dini's theorem now applies (see Apostol [1] again) and the convergence of $f_m(\alpha_0, \beta_0)$ to $f(\alpha_0, \beta_0)$ is uniform. The proof for $g_m(\alpha_0, \beta_0)$ is accommodated along almost identical lines.

## REFERENCES


[1] T. M. Apostol, *Mathematical Analysis,* Addison Wesley, Reading MA, 1974.

[2] L. C. Barbosa and M. Friedman, *Deterministic inventory lot size models–a general root law.* Management Sci., 24 (1978), pp. 819–826.

[3] ———, *Optimal policies for inventory models with some specified markets and finite time horizon,* European J. Oper. Res., to appear.

[4] C. R. Carr and C. W. Howe, *Optimal service policies and finite time horizons,* Management Sci., 9 (1962), pp. 126–140.

[5] W. A. Donaldson, *Inventory replenishment policy for a linear trend in demand–an analytical solution,* Operational Res. Quarterly, 28 (1977), pp. 663–670.

[6] M. Friedman, *The Inventory Lot Size Model with General Time Dependent Demand and Carrying Cost Rate Functions,* Technical Memorandum, Bell Laboratories, Murray Hill, NJ.

[7] G. Hadley and T. Whitin, *Analysis of Inventory Systems.* Prentice-Hall, Englewood Cliffs NJ, 1963.

[8] E. Naddor, *Inventory Systems,* John Wiley, New York, 1966.

[9] M. Resh, M. Friedman and L. C. Barbosa, *On a general solution of the deterministic lot size problem with time proportional demand,* Oper. Res., 24 (1976), pp. 718–725.

[10] M. Sasieni, A. Yaspan, and L. Friedman, *Operations Research: Methods and Problems,* John Wiley, New York, 1959.

[11] E. A. Silver, *A simple inventory replenishment decision rule for a linear trend in demand,* J. Oper. Res. Soc., 30 (1979), pp. 71–75.

[12] A. F. Veinott, Jr., *The status of mathematical inventory theory,* Management Sci., 12 (1966), pp. 745–777.

[13] H. M. Wagner and T. M. Whitin, *Dynamic version of the economic lot size model,* Management Sci., 5 (1958), pp. 81–86.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0009 $01.00/0

# SINGLE FACILITY $l_p$-DISTANCE MINIMAX LOCATION*

Z. DREZNER† AND G. O. WESOLOWSKY‡

**Abstract.** We discuss the problem of locating a new facility among $n$ given demand points on a plane. The maximum weighted distance to demand points must be minimized. The general $l_p$-norm ($p \geqq 1$) is used as distance measure. The method is quite fast computationally: for example, a 3000 demand point problem in $l_2$ is solved in half a second.

**Introduction.** The minimax location problem has been treated extensively in recent literature. A survey is presented in the book by Francis and White [5]. More recent examples are [3] and [4]. In this paper, we present a fast method for locating a single facility on a plane according to the minimax criterion for general or $l_p$-distances. It should be noted that single facility location can be used as a component in multiple facility location; we intend to pursue this approach in a later paper. Related problems, but defined on a network, appear in [6], [7], and [8]. A similar approach to ours, but for the special case of the $l_p$-norm, was discussed by P. M. Dearing [2]. We will refer to his work further. A related methodology for a network was given by Handler [8].

The general $l_p$ formulation for the single facility problem is:

$$\min_{x,y} F(x, y) = \max_{j} \{w_j[|x - a_j|^p + |y - b_j|^p]^{1/p}\} \tag{1}$$

where: $(a_j, b_j)$ is the location of demand point $j$,

$w_j$ is the weight factor associated with demand point $j$;

it converts the distance from the facility to demand point $j$ into cost,

$p \geqq 1$ is the parameter of distance definition.

Note that when $p = 1$, the distance is rectangular, and when $p = 2$, the distance is Euclidean. We will discuss separately the cases $p = 1$ and $p > 1$.

**The Case $p > 1$.** We now present theorems that will be used to prove the convergence of an algorithm for solving (1).

DEFINITION. *p-circle.* A *p-circle* with a given center and radius is the set of all points such that the $l_p$-distance between them and the center is less than or equal to the radius.

It is easy to show that $p$-circles are convex sets, and, in fact, are strictly convex (in the sense that there are no linear segments on the borders) for $p > 1$.

THEOREM 1. *Suppose that three or more p-circles ($p > 1$) with radii $R_i > 0$, $i = 1, \cdots, n$, possess only one common point. Then, one or more subsets of exactly three of these p-circles have only that point in common.*

*Proof.* In contradiction to the theorem, suppose that every three circles possess more than one point in common and hence by convexity, a region of common points. Construct $p$-circles with radii $R_i - \varepsilon$ for $\varepsilon > 0$. If every three $p$-circles with radii $R_i$ possess a region of common points, then for $\varepsilon > 0$ small enough, each three $p$-circles

* Received by the editors January 26, 1978, and in final revised form December 21, 1979. This research was supported in part by the National Research Council of Canada.

† School of Management, University of Michigan–Dearborn, 4901 Evergreen Rd. Dearborn, Michigan 48128.

‡ Production and Management Science Area, Faculty of Business, McMaster University, Hamilton, Ontario, Canada L8S 4M4.

with radii $R_i - \varepsilon$ will have a common point. By the Helly theorem (see [1]), if every subset of three convex sets possess a common point, then all the convex sets possess a common point. Thus, all the $p$-circles with radii $R_i - \varepsilon$ will possess a common point, in contradiction to the fact that all the $p$-circles with radii $R_i$ have only one common point. Note that it is possible that only two $p$-circles define the common point. For the purposes of the theorem any other $p$-circle can then be chosen as the third.

THEOREM 2. *For any $p > 1$, there is a subset of three demand points such that the optimal solution to* (1), *with only these points, is the same as the optimal solution when all points are included.*

*Proof.* The solution to the problem is the minimal $R$ for which $p$-circles with centers at demand points and radii $R/w_i$ have only one common point. The proof follows by Theorem 1.

THEOREM 3. *Consider the solution point for a given triplet of points ($p > 1$). If the weighted distance between a new point and the solution point is greater than the maximum weighted distance between the points of the triplet and the solution point, then one of the triplets, consisting of the new point and two out of the original triplet, possesses a higher value of maximum weighted distance.*

*Proof.* It is clear that the four point solution possesses a higher value of maximum weighted distance. By Theorem 1, there is a triplet out of the four with the same solution. It is not the original triplet, so it must be one of the others.

**Outline of the Algorithm.** Our algorithm for solving the $l_p$ problem is essentially simple: except for the choice of starting point, it is the same as the one proposed by Dearing for $l_2$ distances.

1. Choose a starting point $(x^{(0)}, y^{(0)})$. We propose

$$x^{(0)} = \frac{\sum w_i^p x_i}{\sum w_i^p}, \qquad y^{(0)} = \frac{\sum w_i^p y_i}{\sum w_i^p}, \tag{2}$$

which is the center of gravity of the demand points when weights are $w_i^p$.

2. Find the three points that have the greatest weighted distance from $(x^{(0)}, y^{(0)})$.
3. Solve the minimax location problem for the three points.
4. If the weighted distance to all other points is not greater, stop.
5. Otherwise, find the weighted-furthest point from the solution point, and from among the resulting four points find a new triplet that has a higher maximum weighted distance.
6. Return to step 3.

Since there is a finite number of triplets, and since we do not "pass" any triplet twice because the objective function is increasing, the algorithm must finish in a finite number of steps.

The essential ingredient of the algorithm is step 3. We now turn to solving the three-point problem for $p > 1$.

**The Three-Point Problem.** We first check if any two points define the solution. As can easily be verified, the solution $(x_0, y_0)$ for any two points $(x_1, y_1)$, $(x_2, y_2)$ is found by taking weighted averages as follows:

$$x_0 = \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2}, \qquad y_0 = \frac{w_1 y_1 + w_2 y_2}{w_1 + w_2}. \tag{3}$$

The maximum weighted distance is then

$$F(x_0, y_0) = \frac{w_1 w_2}{w_1 + w_2} [|x_1 - x_2|^p + |y_1 - y_2|^p]^{1/p}. \tag{4}$$

If the weighted distance from $(x_0, y_0)$ to the third point of the triplet $(x_3, y_3)$ is less than or equal to $F(x_0, y_0)$, then $(x_0, y_0)$ is the minimax solution point for the triplet. We can thus check all possible pairs of the triplet. If no pair gives a solution, then we have to find a point *inside* the triangle of points. This point possesses equal weighted distances to the three vertices of the triangle.

**Triplets with $l_p$-distances ($p > 1$).** As suggested by Dearing [2], the locus of equal weighted distances from two points in $l_2$ is a straight line in the equal weights case, and a circle when weights are not equal. The solution to the problem is the intersection between these lines or circles. Explicit formulas appear in Appendix A for reference.

We will now prove that in the $l_p$ ($p > 1$) case, there can be only one point inside the triangle with equal weighted distances to the three vertices.

LEMMA 1. For $p > 1$, an infinitesimal change in the site of a point inside a triangle will increase at least one of the $l_p$-distances to the vertices of the triangle.

*Proof*[1]. Suppose that $x$ is any point in the interior of the triangle and that $y$ is another point such that, in contravention to the lemma its distances from the vertices of the triangle are less than or equal to these distances between $x$ and the vertices. Extend the line segment from $y$ through $x$ until it meets the boundary of the triangle at some point $z$. By the convexity of the $l_p$-norm, $z$ is at least as far from all three vertices as $x$. Consequently, $z$ cannot be one of the vertices. Suppose it lies on the edge joining vertex 1 and vertex 2. Via the triangle inequality, the distance from 1 to 2, which equals the distance from 1 to $z$ plus the distance from $z$ to 2, is less than (since $x$ is interior) the distance from 1 to $x$ plus the distance from 2 to $x$. But, this is impossible if $z$ is at least as far from both vertices 1 and 2 as is $x$. Consequently, the assumption that $y$ has all distances from the vertices less than or equal to the corresponding distances from $x$ to the vertices is untenable.

THEOREM 4. *When $p > 1$, there is at most one point inside the triangle with equal weighted distances to the vertices.*

*Proof.* Such a point is a local minimum by Lemma 1, because the maximum distance increases in any direction. Since $F(x, y)$ in (1) is convex, if there are two distinct local minima, $F(x, y)$ has the same value on the line connecting them; this contradicts Lemma 1.

The following lemma is trivial.

LEMMA 2. *The global minimum of the maximum weighted distances is inside the closed triangle.*

Different computational methods for finding the optimum of a three demand points problem were described in [3]. We now give a very fast heuristic iterative method for solving this problem. Actually, for practical purposes, the method is optimizing. The main idea is based on the fact that when the ratio between $l_p$- and $l_2$-distances is close to 1, it is relatively insensitive to changes in the site of the facility. Recall that the $l_2$ solution can be computed very quickly (formulas in Appendix A). We incorporate this solution in the following iterative method for finding the minimax point for an $l_p$ triplet.

[1] We are grateful to the editor in charge of our paper for suggesting this proof; it is shorter than one we presented originally.

Define $(x^{(k)}, y^{(k)})$ as the $k$th iteration,

$$d_{i,p}^{(k)} = \{|x_i - x^{(k)}|^p + |y_i - y^{(k)}|^p\}^{1/p}. \tag{5}$$

Then, the algorithm can be expressed as follows:

1. Choose a starting point $(x^{(0)}, y^{(0)})$. We suggest the use of (2). Let $k = 0$, $\lambda = 1$, $n(\lambda) = 0$ for all $\lambda$.

2. Compute

$$w_i^{(k)} = w_i d_{i,p}^{(k)}/d_{i,2}^{(k)}; \quad \text{if } d_{i,2}^{(k)} = 0, \text{ then } w_i^{(k)} = w_i. \tag{6}$$

3. Solve the three-point problem in $l_2$ with weights $w_i^{(k)}$ by the method of Appendix A to obtain the solution $(x^*, y^*)$. Let $x^{(k+1)} = \lambda x^* + (1 - \lambda)x^{(k)}$ and $y^{(k+1)} = \lambda y^* + (1 - \lambda)y^{(k)}$. Let $n(\lambda) = n(\lambda) + 1$ for the current $\lambda$.

4. If $\{[x^{(k+1)} - x^{(k)}]^2 + [y^{(k+1)} - y^{(k)}]^2\}^{1/2} < \varepsilon$ for a given $\varepsilon$, then take $(x^{(k+1)}, y^{(k+1)})$ as the solution, declare convergence, and stop. If $n(\lambda) > 50$, set $\lambda = \lambda/2$. Set $k = k + 1$. If $\lambda < \frac{1}{8}$, stop, declaring nonconvergence. Go to step 2.

If the algorithm converges, the following theorem guarantees optimality.

THEOREM 5. *If* $\lim_{k\to\infty} (x^{(k)}, y^{(k)}) = (x, y)$ *then* $(x, y)$ *is the optimal solution.*

*Proof.* In the limit,

$$w_i^{(\infty)} d_{i,2}^{(\infty)} = w_i d_{i,p}^{(\infty)}. \tag{7}$$

As $(x, y)$ is the solution to $l_2$ problem with weights of $w_i^{(\infty)}$, there are two possibilities:

(a) The point is on an edge of the triangle and hence two weighted distances are equal to each other while the third is smaller. Then by (7), the same condition holds for the original weights and $l_p$-distance.

(b) $w_1^{(\infty)} d_{1,2}^{(\infty)} = w_2^{(\infty)} d_{2,2}^{(\infty)} = w_3^{(\infty)} d_{3,2}^{(\infty)}$; then by (8), $w_1 d_{1,p}^{(\infty)} = w_2 d_{2,p}^{(\infty)} = w_3 d_{3,p}^{(\infty)}$.

Applying Lemma 2 to the $l_2$ solution gives that $(x, y)$ is inside the closed triangle. Then, since the point is inside, it is the global optimum for $l_p$ by Lemma 1 and Theorem 4. This algorithm may not always converge: an example is given in Appendix B of nonconvergence for $\lambda = 1$. While the sequence $(x^{(k)}, y^{(k)})$ is bounded inside the triangle, oscillations between points can occur, as is shown in Appendix B. We can not guarantee that instances of nonconvergence for a reasonable number of $\lambda$'s in the sequence $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8} \cdots$, will not occur. As will be seen in the computational results section, we did not find any such examples. If nonconvergence does occur, the "guaranteed" method presented in [3] could be used for the triplet. However it takes about 100 times longer.

**The Minimax problem with $p = 1$.** As is shown in [5] we can (by the $T$ transformation given there) separate the problem into two subproblems, each of which is a one-dimensional problem. It is easy to see that instead of triplets, as in the $p > 1$ case, we deal here with pairs on each axis. The algorithm is otherwise the same.

**Computational results.** Table 1 explores the solution of 5000 "triplets" for each of 7 different values of $p$. The points were uniformly generated within a unit square, and the weights were uniformly generated on the interval [0, 1]. Note that approximately 87% of the triplets had "two-point" solutions and that all converged before $\lambda = \frac{1}{8}$ was reached. The convergence constant, $\varepsilon$, was $10^{-4}$. The average run time for $p = 1.78$ was about .01 seconds on the CDC 6400.

Table 2 gives the results of computational examples with rectangular distances. Four examples were generated randomly for each $n$ (number of demand points). Table 2 gives the worst run time of each four.

Table 3 gives computational results for $l_{1.78}$-distances, as well as the $l_2$ case for

TABLE 1
*Solution of $l_p$ triplets*

| | $p = 1.1$ | $p = 1.2$ | $p = 1.5$ | $p = 1.78$ | $p = 2.5$ | $p = 3.0$ | $p = 4.0$ |
|---|---|---|---|---|---|---|---|
| % 2-point | 87.30 | 87.06 | 86.90 | 86.76 | 86.52 | 86.76 | 87.08 |
| % solved $\lambda = 1$ | 11.48 | 12.08 | 13.08 | 13.24 | 13.48 | 13.24 | 12.58 |
| % solved $\lambda = \frac{1}{2}$ | 1.14 | .82 | .02 | .00 | .00 | .00 | .34 |
| % solved $\lambda = \frac{1}{4}$ | .08 | .04 | .00 | .00 | .00 | .00 | .00 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| $\lambda = 1$ aver. # iter. | 8.23 | 7.11 | 5.03 | 3.66 | 4.24 | 5.32 | 6.40 |
| $\lambda = 1$ max. # iter. | 49 | 48 | 37 | 7 | 9 | 40 | 44 |
| $\lambda = \frac{1}{2}$ aver. # iter. | 55.89 | 55.02 | 55.00 | — | — | — | 53.41 |
| $\lambda = \frac{1}{2}$ max. # iter. | 63 | 67 | 55 | — | — | — | 57 |
| $\lambda = \frac{1}{4}$ aver. # iter. | 105.50 | 104.00 | — | — | — | — | — |
| $\lambda = \frac{1}{4}$ max. # iter. | 111 | 105 | — | — | — | — | — |
| av. time rel. to $p = 1.78$ | 1.8 | 1.5 | 1.1 | 1.0 | 1.1 | 1.3 | 1.4 |

TABLE 2
*Problems with rectangular distances*

| # Demand Points | # Iterations | Time (*sec.*) |
|---|---|---|
| 1000 | 4 | 0.11 |
| 2000 | 3 | 0.17 |
| 3000 | 3 | 0.26 |
| 4000 | 3 | 0.35 |
| 5000 | 3 | 0.43 |

TABLE 3
*Results for $p = 2$ and $p = 1.78$*

| | $p = 2$ | | $p = 1.78$ | |
|---|---|---|---|---|
| # Demand Points | Iterations | Time (*sec.*) | Iterations | Time (*sec.*) |
| 1000 | 1 | 0.10 | 1 | 0.86 |
| 1500 | 2 | 0.19 | 2 | 1.91 |
| 2000 | 2 | 0.26 | 6 | 5.98 |
| 2500 | 5 | 0.57 | 4 | 5.38 |
| 3000 | 3 | 0.48 | 1 | 2.59 |
| 3500 | 2 | 0.45 | 4 | 7.48 |

comparison. Longer times result when $p = 1.78$ mainly because quantities must be raised to the $p$th power.

**Appendix A.**

1. For $w_1 = w_2 = w_3$.

Let:

$$\Delta = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix},$$

$$\Delta_1 = \begin{vmatrix} 1 & x_1^2 + y_1^2 & y_1 \\ 1 & x_2^2 + y_2^2 & y_2 \\ 1 & x_3^2 + y_3^2 & y_3 \end{vmatrix},$$

$$\Delta_2 = \begin{vmatrix} 1 & x_1 & x_1^2 + y_1^2 \\ 1 & x_2 & x_2^2 + y_2^2 \\ 1 & x_3 & x_3^2 + y_3^2 \end{vmatrix}.$$

Then:

$$x_0 = \frac{\Delta_1}{2\Delta}, \qquad y_0 = \frac{\Delta_2}{2\Delta}. \tag{A.1}$$

2. If the weights are not equal, there exists at least one weight, say $w_1$, such that $w_1 \neq w_2$, $w_1 \neq w_3$.

Compute the following:

For $i = 2, 3$,

$$x_{i1} = \frac{w_i^2 x_i - w_1^2 x_1}{w_i^2 - w_1^2}, \qquad y_{i1} = \frac{w_i^2 y_i - w_1^2 y_1}{w_i^2 - w_1^2},$$

$$R_i = \frac{w_1 w_i}{|w_1^2 - w_i^2|}\left[(x_1 - x_i)^2 + (y_1 - y_i)^2\right]^{1/2}. \tag{A.2}$$

Then

$$\bar{x} = x_{31} - x_{21}, \qquad \bar{y} = y_{31} - y_{21}, \tag{A.3}$$

$$\begin{aligned} A &= [(R_2 + R_3)^2 - (\bar{x}^2 + \bar{y}^2)][\bar{x}^2 + \bar{y}^2 - (R_2 - R_3)^2]/4, \\ B &= (\bar{x}^2 + \bar{y}^2 - R_3^2 + R_2^2)/2. \end{aligned} \tag{A.4}$$

If $A < 0$, there is no solution for the triplet. In this case the solution is defined by a pair of points. There are two possible solution points from which we have to find the one with the smaller objective function,

$$x_0 = x_{21} + \frac{\bar{x}B \pm \bar{y}\sqrt{A}}{\bar{x}^2 + \bar{y}^2}, \qquad y_0 = y_{21} + \frac{\bar{y}B \mp \bar{x}\sqrt{A}}{\bar{x}^2 + \bar{y}^2}. \tag{A.5}$$

Note that the sign in $x_0$ is opposite to that in $y_0$.

**Appendix B.** The following is an example of a problem for which the $l_p$ triplet procedure does not converge for $\lambda = 1$ and $1 \leqq p < 2$.

Find $\theta$ $(0 < \theta < 1)$ such that

$$\theta^p > \frac{2\theta^2}{(1 - \theta^2)}. \tag{B.1}$$

We then construct the following problem. Choose $\varepsilon > 0$ small enough such that $O(\varepsilon^2)$ in the computation below will be negligible.

The problem:

| $i$ | $x_i$ | $y_i$ | $w_i$ |
|---|---|---|---|
| 1 | $\frac{-1}{\varepsilon^2}$ | 0 | $\varepsilon^2$ |
| 2 | $\frac{1}{\varepsilon^2}$ | 0 | $\varepsilon^2$ |
| 3 | 1 | $\theta$ | $\frac{1+\varepsilon}{(1+\theta^p)^{1/p}}$ |

Computation yields:

| Iteration $i$ | $x^{(i)}$ | $y^{(i)}$ | $w_1^{(i)} + O(\varepsilon^4)$ | $w_2^{(i)} + O(\varepsilon^4)$ | $w_1^{(i)}(1+\theta^2)^{1/2} + O(\varepsilon^2)$ |
|---|---|---|---|---|---|
| 0 | $\frac{w_3^p}{(w_3^p + 2\varepsilon^{2p})}$ | $\theta x^{(0)}$ | $\varepsilon^2$ | $\varepsilon^2$ | $1+\varepsilon$ |
| 1 | $O(\varepsilon^2)$ | $\frac{\varepsilon(1+\theta^2)}{\theta} + O(\varepsilon^2)$ | $\varepsilon^2$ | $\varepsilon^2$ | $\frac{1-\varepsilon[(1-\theta^2)\theta^{p-2}-2]}{(1+\theta^p)}$ |
| 2 | $O(\varepsilon^2)$ | $O(\varepsilon^2)$ | $\varepsilon^2$ | $\varepsilon^2$ | $1+\varepsilon$ |

The solution oscillates between iterations 1 and 2, while the optimal point is $(O(\varepsilon^2), \varepsilon(1+\theta^p)/\theta^{p-1} + O(\varepsilon^2))$.

## REFERENCES


[1] L. DANZER, B. GRUNBAUM, AND V. KLEE, *Proceedings of Symposium in Pure Mathematics. Convexity,* V. Klee, ed., American Mathematical Society, Providence, RI, 1963, pp. 101–180.

[2] P. M. DEARING, *Minimax location and Voroni diagrams,* presented at Joint National ORSA/TIMS Meeting, Atlanta, November, 1977.

[3] Z. DREZNER AND G. O. WESOLOWSKY, *A new method for the multifacility minimax location problem,* J. Operational Research Society, 29 (1978), pp. 1095–1101.

[4] J. ELZINGA, D. HEARN, AND W. D. RANDOLF, *Minimax multifacility location with Euclidean distances,* Transporation Sci., 10 (1976), pp. 321–336.

[5] R. L. FRANCIS AND J. A. WHITE, *Facility Layout and Location,* Prentice-Hall, Englewood Cliffs, NJ, 1974.

[6] S. L. HAKIMI, *Optimum locations of switching centers and the absolute centers and medians of a graph,* Operations Res., 12 (1964), pp. 450–459.

[7] HANDLER, *Minimax Network Location: Theory and Algorithms,* Technical Report No. 374, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1974.

[8] ———, *The n-center problem: a relaxation approach,* Working Paper No. 5816, Faculty of Management, Tel-Aviv University, Israel. January, 1977.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0010 $01.00/0

# THE CONNECTION PATTERNS OF TWO COMPLETE BINARY TREES*

F. R. K. CHUNG† AND F. K. HWANG†

**Abstract.** We consider the class of channel graphs which can be viewed as compositions of two copies, right and left, of a complete binary tree with terminal nodes of the right tree connected to distinct terminal nodes of the left tree. We study the connection patterns of the two binary trees to minimize the blocking probability of the resulting channel graphs. We show that the connection patterns given by Ikeno are not optimal in general and in fact no optimal connection patterns exist for such graphs with more than 9 stages. We present new connection patterns which uniquely possess certain optimal properties.

**1. Introduction.** We consider the class of graphs which consist of two copies, right and left, of a complete binary tree with terminal nodes of the right tree connected to distinct terminal nodes of the left tree. If the complete binary trees have $n$ levels and $2^{n-1}$ terminal nodes, then there are $(2^{n-1})!$ possible ways to connect these two trees, though some of them might be isomorphic. In Fig. 1 there are two nonisomorphic graphs, each of which is formed by combining two binary trees having three levels. In general such a graph can be viewed as a $2n$-stage network having the roots of the two complete binary trees as the source and sink of the network. We call such a graph a *binary channel graph* or a *binary graph.* We note that a binary graph is a special case of multistage *channel graphs,* also called *linear graphs,* which are often used in studying the blocking performance of switching networks [3]. A channel graph can be viewed as the union of all paths connecting a specified pair of input and output terminals in a switching network. In an *m-stage channel graph,* all vertices are partitioned into a sequence of $m$ subsets, called *stages;* edges, called *links,* exist only between vertices in adjacent stages. The links between stage $i$ and stage $i + 1$ are referred to as the $i$th stage links, and we assume that each $i$th stage link has probability $p_i$ of being occupied. The vector $(p_1, \cdots, p_{m-1})$ is called the *link occupancies* of the channel graph. The *blocking probability* of a channel graph is defined to be the probability that every path from the source to the sink contains at least one occupied edge. Two channel graphs of the same number of stages can be compared in the following way. We say one channel graph is *superior* to the other if the blocking probability of the former never exceeds that of the latter for any given link occupancies. A $2n$-stage binary graph is said to be *optimal* if it is superior to any other $2n$-stage binary graph.

The problem of determining the connection pattern of the two binary trees to minimize the blocking probability of the resulting binary graph is not only interesting on its own right but also useful in designing effective switching networks (see [3]). Ikeno [1] investigated this problem and suggested the following simple and elegant connection: Assign binary numbers to terminal nodes of the right and left trees in an orderly fashion and connect two terminal nodes such that the digits of the corresponding binary numbers are inversions of each other. For example, the connections in the graph in Fig. 1(b) are (0, 0) to (0, 0), (0, 1) to (1, 0), (1, 0) to (0, 1) and (1, 1) to (1, 1).

It can be easily verified that the Ikeno graph is optimal for $n = 3$. However, the optimality for Ikeno graphs for $n > 3$ has not been previously determined in the past,

* Received by the editors October 4, 1979.
† Bell Laboratories, Murray Hill, New Jersey 07974.

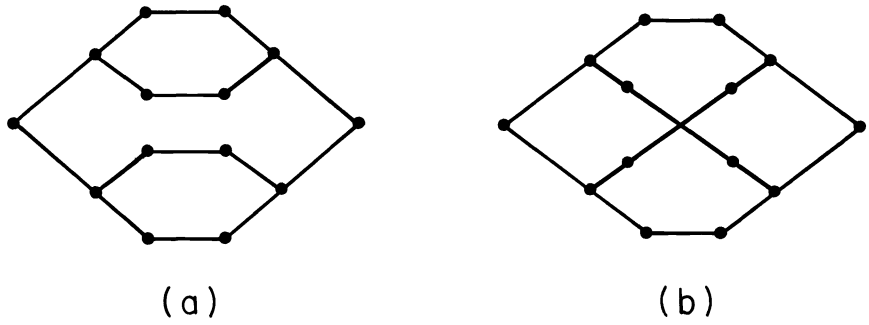



FIG. 1

though there has been a general willingness to believe it (see Neiman [3], for example). In this paper, we show that the optimal binary graph is not the Ikeno graph for $n = 4$, and for $n > 4$ no optimal binary graphs exist. We will give a connection pattern for $2n$-stage binary graphs which uniquely possess certain optimal properties.

**2. Labelings and connection functions for binary graphs.** We first give a labeling for the set of vertices of a binary graph. This labeling is useful in specifying the structure of a binary graph, in particular, in determining whether two binary graphs are isomorphic. Let $T$ be a binary tree in $(n + 1)$ levels. We label the set of vertices of $T$ except the root as follows:

(i) The set of vertices in the level $k + 1$, $k \geqq 1$, are labeled by the set of $k$-tuples with entries either 0 or 1, i.e., $\{\mathbf{l}(u)$: $u$ is a vertex in level $k + 1$ in $T\} = \{0, 1\}^k$.

(ii) Let $u$ be a vertex in level $k + 1$ adjacent to a vertex $v$ in level $k$. Then, the label of $u$, $\mathbf{l}(u)$, is either $(\mathbf{l}(v), 0)$ or $(\mathbf{l}(v), 1)$ for $k \geqq 1$.

For a given $2(n + 1)$-stage binary graph, we label both the left tree and the right tree satisfying properties (i) and (ii). Then this binary graph can be characterized by a function $f$: $\{0, 1\}^n \to \{0, 1\}^n$ such that

$$f(a_1, \cdots, a_n) = (b_1, \cdots, b_n)$$

implies that the vertex with label $(a_1, \cdots, a_n)$ of the left tree is connected to the vertex with label $(b_1, \cdots, b_n)$ of the right tree. $f$ is called *the connection function* of the binary graph. For example, the connection function of the graph in Fig. 1(a) is $f_1(a, b) = (a, b)$. The connection function of the graph in Fig. 1(b) is $f_2(a, b) = (b, a)$. Also, the $2(n + 1)$-stage Ikeno graph has connection function $f(a_1, \cdots, a_n) = (a_n, \cdots, a_1)$.

It is easily verified that $f$ is a bijection (one-to-one and onto). For any given bijection $f$: $\{0, 1\}^n \to \{0, 1\}^n$, we can construct a binary graph having $f$ as its connection function. There are $(2^n)!$ such bijections. However, two binary graphs constructed by using two different connection functions might be isomorphic. Thus, we need some methods for determining whether two binary graphs with distinct connection functions are isomorphic or not. We first introduce some terminology.

Let $C$ be a vector of $2^n - 1$ elements; i.e., $C = \{c_{iS}$: $S \subseteq \{1, \cdots, i - 1\}$, $1 \leqq i \leqq n\}$, where $c_{iS} \in \{0, 1\}$.

Let $\mathbf{C}$ denote the following function from $\{0, 1\}^n$ to $\{0, 1\}^n$, such that the $i$th coordinate of $\mathbf{C}(x_1, \cdots, x_n)$ is $x_i + \sum_{S \subseteq \{1,\cdots,i-1\}} c_{iS}\pi(S)$, where $\pi(S)$ denotes the product of all $x$ in $S$, and $\pi(S) = 1$ if $S$ is the empty set. We note that the numbers in the proceeding expression are in $\{0, 1\}$ reduced modulo 2.

LEMMA 1. *Two binary graphs, $G_1$ and $G_2$, having connection functions $f_1$ and $f_2$*

*respectively, are isomorphic if and only if there exist two* $(2^n - 1)$*-vectors* $C, D$ *such that*

$$\mathbf{C}f_1 = f_2\mathbf{D}.$$

*Proof.* Suppose $\mathbf{C}f_1 = f_2\mathbf{D}$ for some vectors $C$ and $D$. We define a map $\alpha$ from the vertex set of $G_1$ to the vertex set of $G_2$ as follows:

Let $u$ be a vertex of the left tree in $G_1$ with label $(a_1, \cdots, a_i)$. Then we define $\alpha(u)$ to be the vertex of the left tree in $G_2$ with label $\mathbf{D}(a_1, \cdots, a_i)$.

Let $v$ be a vertex of the right tree in $G_1$ with label $(b_1, \cdots, b_i)$. Then we define $\alpha(v)$ to be the vertex of the right tree in $G_2$ with label $\mathbf{C}(b_1, \cdots, b_i)$.

Moreover, the roots of the left tree and the right tree of $G_1$ are mapped to the roots of the left tree and right tree of $G_2$ respectively.

It is easily seen that $\alpha$ is one to one and onto. By straightforward verification based on the fact that $\mathbf{C}f_1 = f_2\mathbf{D}$, it can be shown that for $u, v \in V(G_1)$, $u$ is adjacent to $v$ if and only if $\alpha(u)$ is adjacent to $\alpha(v)$ in $G_2$. Thus $\alpha$ is an isomorphism, and the two graphs $G_1$ and $G_2$ are isomorphic.

Now, we assume $G_1$ and $G_2$ are isomorphic. $f_1$ and $f_2$ can be viewed as connection functions determined by two labelings $\mathbf{l}_1$ and $\mathbf{l}_2$ of the same binary graph $G$. Let us first define another labeling $\mathbf{l}_3$ of $G$ such that $\mathbf{l}_3(u) = \mathbf{l}_1(u)$ for any $u$ in the left tree of $G$ and $\mathbf{l}_3(v) = \mathbf{l}_2(v)$ for any $v$ in the right tree of $G$. Let $f_3$ denote the connection function of $G$ determined by the labeling $\mathbf{l}_3$.

CLAIM. $f_3 = \mathbf{C}f_1$ *for some* $(2^n - 1)$*-vector* $C$.

Let us first consider all the labelings of $G$ with vertices in the left tree labeled as in $\mathbf{l}_1$. There are exactly $2^{2^n-1}$ ways to label the right tree. However, for any vector $C = \{c_{iS}: S \subseteq \{1, \cdots, i-1\}, 1 \leqq i \leqq n, c_{iS} \in \{0, 1\}\}$, we can define a labeling of the vertices in the right tree of $G$ as follows:

$$\mathbf{l}_C(v) = \mathbf{C}(b_1, \cdots, b_i),$$

where $\mathbf{l}_1(v) = (b_1, \cdots, b_i)$.

We note that all labelings $\mathbf{l}_C$, $C \in \{0, 1\}^{2^n-1}$ are distinct. Thus $\{\mathbf{l}_C: C \in \{0, 1\}^{2^n-1}\}$ are exactly the set of all labelings of vertices in the right tree of $G$. Therefore, there exists a vector $C$ such that $\mathbf{l}_3(v) = \mathbf{l}_C(v)$ for any vertex $v$ in the right tree.

Thus, for a terminal vertex $u$ in the left tree which is adjacent to a terminal vertex $v$ in the right tree, we have

$$f_1(\mathbf{l}_1(u)) = \mathbf{l}_1(v),$$

$$f_3(\mathbf{l}_3(u)) = \mathbf{l}_3(v) = \mathbf{l}_C(v) = \mathbf{C}\mathbf{l}_1(v) = \mathbf{C}f_1(\mathbf{l}_1(u)).$$

Since $\mathbf{l}_1(u) = \mathbf{l}_3(u)$ for any $u$ in the left tree, we have

$$f_3 = \mathbf{C}f_1.$$

In a similar way it can be shown that $f_3 = f_2\mathbf{D}$ for some vector $D$. Thus we obtain $\mathbf{C}f_1 = f_2\mathbf{D}$ and the lemma is proved.

Lemma 1 is, in fact, equivalent to the following.

COROLLARY. *Two* $2(n+1)$*-stage binary graphs,* $G_1$ *and* $G_2$*, having connection functions* $f_1$ *and* $f_2$*, respectively, are isomorphic if and only if there exist two* $(2^n - 1)$*-vectors* $C$ *and* $E$ *such that*

$$f_2 = \mathbf{C}f_1\mathbf{E}.$$

*Proof.* It suffices to show that the inverse function of the linear transformation $\mathbf{D}$

is of the form **E** for some vector $E$. This, however, can be done by straightforward calculation to find values of $e_{ij}$, $1 \leqq j \leqq i \leqq n$ such that $\mathbf{DE} = \mathbf{I}$ where **I** is the identity transformation.

Let $f_1, f_2$ be two bijections from $\{0, 1\}^n$ to $\{0, 1\}^n$. We say $f_1$ is *equivalent* to $f_2$ if there exists two $(2^n - 1)$-vectors, say $C$ and $D$ such that

$$f_1 = \mathbf{C} f_2 \mathbf{D}.$$

We note that two connections functions derived from different labelings of a binary graph are equivalent.

**3. The optimal 8-stage binary graph.** It can be easily verified that the Ikeno graph is an optimal $2n$-stage binary graph for $n = 1, 2$ and 3. However, as an immediate result of the following theorem the Ikeno graph is not optimal for $n = 4$.

THEOREM 1. *The optimal 8-stage binary graph $G_8$ is isomorphic to the binary graph determined by the connection function* $f(a_1, a_2, a_3) = (a_3, a_2 + a_1 a_3, a_1)$ *as shown in* Fig. 2.

Before we prove Theorem 1, we state a result of Takagi [6]. Let $T_k$ denote the class of 4-stage channel graphs where there are $k$ each of the second-stage and the third-stage vertices, and each second-stage (third-stage) vertex is connected to two third-stage (second-stage) vertices.

THEOREM (*Takagi*). *A graph in $T_k$ is optimal if and only if the $2k$ middle-stage links form a cycle.*

*Proof of Theorem 1.* Let $L_i$ denote the set of $i$th stage links. It suffices to show that in each of the following four cases, the blocking probability of $G_8$ achieves minimum (since the four cases are mutually disjoint and exhaustive).

(i) Both links in $L_1$ and both links in $L_7$ are idle. In this case, an 8-stage binary graph is not blocked if there exists a path from any second-stage vertex to any seventh-stage vertex. Therefore the graph can be viewed as a 6-stage graph by eliminating the first and last stage and combining vertices in stage 2 and stage 7. By *shrinking* a graph from stage $i$ to stage $j$ we mean replacing every path between the two stages by a link. When all these paths are edge disjoint, then by simply defining the probability of the new link being busy as the probability of the path being busy, the blocking probability of the shrunken graph is the same as the blocking probability of the original graph. By shrinking the reduced $G_8$ from stage 2 to stage 5, we obtain a 4-stage graph in $T_4$. Fig. 3 shows such a $T_4$ obtained from $G_8$:

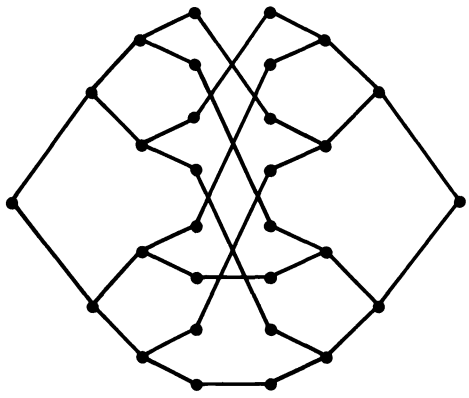


FIG. 2

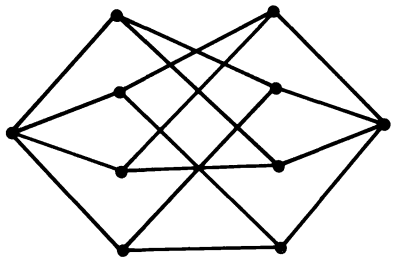


FIG. 3

Note that the eight middle-stage links in this graph form a cycle. Theorem 1 follows immediately from the Takagi Theorem.

(ii) Both links in $L_1$ and one link in $L_7$ are idle. $G_8$ is reduced to the 6-stage graph in Fig. 4. It is straightforward to verify that this is the best reduction.

(iii) One link in $L_1$ and both links in $L_7$ are idle. Same as (ii).

(iv) One link in $L_1$ and one link in $L_7$ are idle. We have to consider the set of four reduced graphs corresponding to the four possible combinations of idle links. There are three possible sets for 8-stage binary graphs. A Type A set consists of two graphs of four paths and two empty graphs. A Type B set consist of two graphs of three paths and two graphs of one path. A Type C set consists of four graphs of two paths. The reduced $G_8$ is a Type C set and it is straightforward to verify that the sum of blocking probabilities of the four graphs in a Type C set is minimal.

From (i), (ii), (iii) and (iv), we conclude that $G_8$ is an optimal 8-stage graph. It is straightforward to verify that $G_8$ is indeed the unique optimal 8-stage graph up to isomorphism since there are only a few distinct connections to check. We omit the details here.

The 8-stage Ikeno graph $I_8$ is illustrated in Fig. 5. It can be easily seen that $G_8$ is superior to the 8-stage Ikeno graph $I_8$ since the blocking probability of $G_8$ is less than or equal to the blocking probability of $I_8$ in case (i); and the blocking probability of $G_8$ is equal to the blocking probability of $I_8$ in case (ii), (iii) and (iv) for any given link occupancies.

**4. Some preliminary results on binary graphs.** In a channel graph $G$, let $u$ be a vertex in stage $i$ and $v$ a vertex in stage $j$ where $j > i$. The *channel subgraph* determined by $u$ and $v$ in $G$ is defined to be the union of all paths of length $j - i$ connecting $u$ and $v$ in $G$. A *maximal channel subgraph* of $G$ is a channel subgraph of $G$ having the number of stages one less than the number of stages in $G$. A maximal channel subgraph is usually determined by a vertex in the second stage and the sink (the vertex in the last stage) or the source (the vertex in the first stage) and a vertex in the stage next to the last.

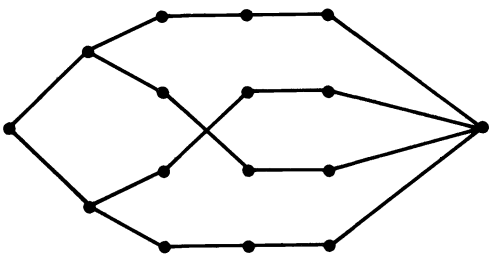


FIG. 4

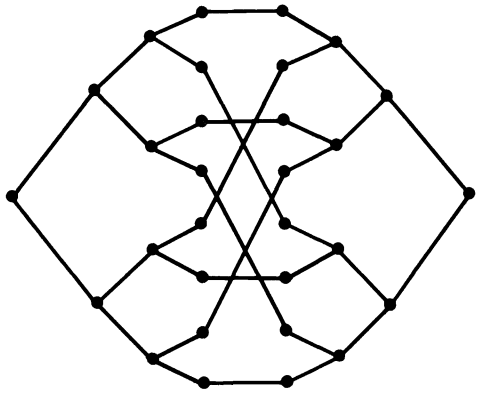


FIG. 5

We say a maximal channel subgraph of a $2n$-stage binary graph is optimal if it is isomorphic to an optimal $2(n-1)$-stage binary graph after shrinking from stage $n-1$ to stage $n+1$ or from stage $n$ to stage $n+2$.

LEMMA 2. *All the maximal channel subgraphs of an optimal 2n-stage binary graph are optimal.*

*Proof.* Let $G'$ be a maximal channel subgraph determined by a vertex in stage 2 and the sink in the optimal $2n$-stage binary graph $G$. Let $H$ be a $2n$-stage binary graph such that all its maximal channel subgraphs determined by a vertex in stage 2 and the sink are, after shrinking from stage $n$ to stage $n+2$, isomorphic to a given arbitrary $2(n-1)$-stage binary graph $H'$. To be precise, let $f$ denote the connection function of a $2(n-1)$-stage binary graph. We can then construct $H$ with the connection function $f^*$ to be $f^*(a_1, \cdots, a_{n-1}) = (b_1, \cdots, b_{n-1})$, where $f(a_2, \cdots, a_{n-1}) = (b_1, \cdots, b_{n-2})$ and $a_1 = b_n$. Now we consider the link occupancies $p_1, \cdots, p_{n-1}$ where $p_1$ is extremely small compared with $p_i$, $i \neq 1$. Then the blocking probabilities of $G$ and $H$ can be approximated by the product of $p_1$ and the blocking probability of $G'$ and $H'$, respectively. Since $G$ is an optimal binary graph, $G'$ must be optimal.

COROLLARY. *If no optimal 2n-stage binary graph exists, then no optimal 2m-stage binary graph exists for* $m \geqq n$.

In a later section we will show that optimal 10-stage binary graphs do not exist. Thus optimal $2n$-stage binary graphs, $n \geqq 5$, do not exist. It is then of interest to investigate binary graphs with certain optimal properties.

We note that the optimal $2n$-stage binary graphs, for $n \leqq 4$, have the *symmetric* property, i.e., the connection function $f$ satisfies the following:

$$f(a_1, \cdots, a_{n-1}) = (b_1, \cdots, b_{n-1}), \quad \text{if and only if}$$
$$f(b_1, \cdots, b_{n-1}) = (a_1, \cdots, a_{n-1}).$$

In other words, the connection function $f$ is *idempotent*, $f = f^{-1}$. Symmetry might be a desirable property for practical considerations since it facilitates the construction and the control of the network.

We define $G_8$ to be an HS-optimal 8-stage binary graph (it is in fact optimal). For $n \geqq 5$, a $2n$-stage binary graph is said to be *hereditary* if all its maximal channel subgraphs, after shrinking, are $2(n-1)$-stage HS-optimal binary graphs. We call a $2n$-stage binary graph HS-optimal if it is both hereditary and symmetric. The main result of this paper is to prove that for $n \geqq 5$, there exists a unique HS-optimal $2n$-stage

binary graph. In this section, we give a set of lemmas concerning HS-optimal $2n$-stage binary graphs.

From the definition of HS-optimal binary graphs, the following observations are immediate.

LEMMA 3. *Let $f$ be the connection function of an* HS-*optimal $2(n+1)$-stage binary graph. We define the function $\bar{f}_{a_1}$: $\{0,1\}^{n-1} \to \{0,1\}^{n-1}$ by the following:*

$$\bar{f}_{a_1}(a_2, \cdots, a_n) = (b_1, \cdots, b_{n-1}),$$

*where $f(a_1, a_2, \cdots, a_n) = (b_1, \cdots, b_n)$. Then $\bar{f}_{a_1}$ is the connection function of the maximal channel subgraph determined by the sink and the vertex of the left tree with labeling $(a_1)$. Similarly $(\bar{f}^{-1})_{b_1}$ is the connection function of the maximal channel subgraph determined by the source and the vertex of the right tree with labeling $(b_1)$.*

LEMMA 4. *A $2(n+1)$-stage binary graph having connection function $f$ is hereditary if and only if $\bar{f}_a$ and $(\bar{f}^{-1})_b$, for $a, b \in \{0,1\}$, are equivalent to connection functions of an* HS-*optimal $2n$-stage binary graph.*

*Proof.* This follows from the definition of hereditary.

LEMMA 5. *Let $f$ be the connection function of an* HS-*optimal $2(n+1)$-stage binary graph $G$. Then $f$ is equivalent to the following function $g$:*

$$g(a_1, \cdots, a_n) = (\bar{f}_{a_1}(a_2, \cdots, a_n), a_1).$$

*Proof.* It suffices to show that we can choose a proper labeling of $G$ with connection function $g$. Because $G$ is HS-optimal, for any labeling $\mathbf{l}$ of $G$ the two vertices with labeling $(b_1, \cdots, b_{n-1}, 0)$ and $(b_1, \cdots, b_{n-1}, 1)$ are connected to two vertices in stage $n+1$ with distinct first components, i.e.,

$$f(a_1, \cdots, a_n) = (b_1, \cdots, b_{n-1}, 0)$$

and

$$f(a_1', \cdots, a_n') = (b_1, \cdots, b_{n-1}, 1)$$

imply $a_1 \neq a_1'$.

Thus we consider a labeling $\mathbf{l}'$ of $G$ such that a vertex $v$ with $\mathbf{l}(v) = (b_1, \cdots, b_{n-1}, b_n)$ will be labeled as $\mathbf{l}'(v) = (b_1, \cdots, b_{n-1}, b_n + 1)$, if $a_1 \neq b_n$, and $\mathbf{l}'(v) = (b_1, \cdots, b_n)$ otherwise, where $f(a_1, \cdots, a_n) = (b_1, \cdots, b_n)$. We note that the connection function $g$ under the labeling $\mathbf{l}'$ can be written as follows:

$$g(a_1, \cdots, a_n) = (\bar{f}_{a_1}(a_2, \cdots, a_n), a_1).$$

Since $f$ and $g$ are connection functions of $G$ in two labelings, we know $f$ and $g$ are equivalent from Lemma 2.

Thus Lemma 5 is proved.

LEMMA 6. *Let $x$ be a number depending on $a$ where $x, a \in \{0,1\}$. Then $x$ can be written as follows:*

$$x = aw_1 + w_2 \pmod 2,$$

*where $w_1, w_2$ are some constants in $\{0,1\}$.*

*Proof.* It is easy to find $w_1, w_2$ by solving two equations for $a = 0$ or 1. Namely, $w_2 = x(0)$, $w_1 = x(1) - x(0)$.

For any vector $C = \{c_{iS}\}$ we let $\mathbf{C}_a$ denote the linear transformation determined by $\{c_{iS}a\}$.

LEMMA 7. Let $C$ be a $(2^n - 1)$-vector $\{c_{iS}: S \subseteq \{1, \cdots, i-1\}, i = 1, \cdots, n\}$ *where $c_{iS}$'s depend on $a$. Then the linear transformation $\mathbf{C}$ can be written as follows:*

$$\mathbf{C} = \mathbf{DE} = \mathbf{FG}$$

*where E and F are* $(2^n - 1)$*-vectors (independent of a) and* $D = D_a$, $G = G_a$.

*Proof.* The values of elements of **E**, and **D** can be obtained by using Lemma 6 and solving the $n(n + 1)$ linear equations derived from $\mathbf{C} = \mathbf{DE}$. The values of elements of **F** and **G** can be obtained similarly.

LEMMA 8. *Let G be a* $2(n + 1)$*-stage hereditary binary graph. There exists a labeling of G such that the connection function can be written as follows:*

$$(*) \qquad f(a_1, \cdots, a_n) = (\mathbf{M}_{a_1} g \mathbf{N}_{a_1}(a_2, \cdots, a_n), a_1),$$

*where g is some connection function of an* HS*-optimal* $2n$*-stage binary graph, and* $n_{i\phi} = 0$, *for* $i = 1, \cdots, n - 1$.

*Proof.* Let $h$ be the connection function of a $2(n + 1)$-stage hereditary binary graph $G$. From Lemma 5 we may assume

$$h(a_1, \cdots, a_n) = (b_1, \cdots, b_n) = (\bar{h}_{a_1}(a_2, \cdots, a_n), a_1).$$

Since $G$ is hereditary, we know that, by Lemma 4, $\bar{h}_{a_1}(a_2, \cdots, a_n) = (b_1, \cdots, b_{n-1})$ is a connection function of the $2n$-stage HS-optimal binary graph. By the induction assumption, we have $\bar{h}_a = \mathbf{R}g\mathbf{S}$ where the elements of $R$ or $S$ depend on $a$. Therefore, by Lemma 7, we have

$$\mathbf{R} = \mathbf{UM}_a,$$

and

$$\mathbf{S} = \mathbf{N}'_a \mathbf{V},$$

where elements of $U$ and $V$ are independent of $a$. We define $U'$ and $V'$ such that $\mathbf{U}'(x_1, \cdots, x_n) = (\mathbf{U}(x_1, \cdots, x_{n-1}), x_n)$ $\mathbf{V}'(x_1, \cdots, x_n) = (x_1, \mathbf{V}(x_2, \cdots, x_n))$, and we note that we can find a $(2^n - 1)$-vector $W$ such that $\mathbf{W}(x_1, \cdots, x_n) = (x_1, \mathbf{W}'_{x_i}(x_2, \cdots, x_n))$ and $\mathbf{N}'_{a_1} = \mathbf{N}_{a_1} \mathbf{W}'_{a_1}$, where $n_{i\phi} = 0$ for $i = 1, \cdots, n$.

It is easy to verify that

$$\mathbf{U}' f \mathbf{W}\mathbf{V}' = h,$$

where $f(a_1, \cdots, a_n) = (\mathbf{M}_{a_1} g \mathbf{N}_{a_1}(a_2, \cdots, a_n), a_1)$. Therefore, Lemma 8 is proved.

From now on, for any HS-optimal binary graph, we will only consider its connection functions in form as described in Lemma 8. We note that several examples of connection functions we give in this paper are in this form.

LEMMA 9. *A* $2(n + 1)$*-stage binary graph G is* HS*-optimal if and only if the connection function of G is equivalent to f satisfying*

$$(1) \qquad \begin{aligned} f(a_1, \cdots, a_n) &= (\mathbf{M}_{a_1} g \mathbf{N}_{a_1}(a_2, \cdots, a_n), a_1) \\ &= (a_n, (\mathbf{N}_{a_n})^{-1} g^{-1} (\mathbf{M}_{a_n})^{-1}(a_1, \cdots, a_{n-1})), \end{aligned}$$

*where g is a connection function of the optimal* $2n$*-stage binary graph and M and N are two* $(2^n - 1)$*-vectors.*

*Proof.* Since $G$ is symmetric and hereditary, we have (1) from Lemma 8. Now, suppose $G$ has a connection function equivalent to $f$. It follows immediately $G$ is symmetric. By Lemma 3 and 4, $G$ is also hereditary. Thus $G$ is HS-optimal.

Now we consider a set of connection functions $P = \{p_i : i = 1, 2, \cdots\}$ where $p_i : [0, 1]^i \to [0, 1]^i$. We set

$$p_2(a_1, a_2) = (a_2, a_1),$$

$$p_3(a_1, a_2, a_3) = (a_3, a_2 + a_1a_3, a_1) \qquad \text{(mod 2)}.$$

The $p_i$ for $i \geqq 4$, will be defined in §§ 5, 6.

We note that $p_2$, $p_3$ are connection functions for optimal 6- and 8-stage binary graphs. The set $P$ denotes the set of the connection functions for HS-optimal binary graphs.

**5. 10-stage binary graphs.** For 10-stage binary graphs, we will show the following.

THEOREM 2. *The* HS-*optimal* 10-*stage binary graph has connection function* (*up to isomorphism*) *as follows* (*see Fig.* 6):

$$p_4(a_1, a_2, a_3, a_4) = (a_4, a_3 + a_2a_4, a_2 + a_1a_3 + a_1a_2a_4, a_1) \qquad \text{(mod 2)}.$$

We remind the reader that all calculations are in $Z_2$. We will, from now on, omit the notation (mod 2).

*Proof.* It is straightforward to verify that $p_4$ is idempotent.

The function $p_4$ can be written as

$$\begin{aligned} p_4(a_1, a_2, a_3, a_4) &= (\mathbf{L}_{a_1}p_3(a_2, a_3, a_4), a_1) \\ &= (a_4, p_3\mathbf{L}_{a_4}(a_1, a_2, a_3)), \end{aligned}$$

where

$$\begin{aligned} \mathbf{L}_a(x_1, x_2, x_3) &= (x_1, x_2, x_3 + ax_2) \\ &= (\mathbf{L}_a)^{-1}(x_1, x_2, x_3). \end{aligned}$$

It follows from Lemma 9 that the binary graph having $p_4$ as the connection function is HS-optimal.

Now, suppose $G$ is an HS-optimal 10-stage binary graph having connection function $f$. From Lemma 9 $f$ satisfies the following

$$f(a_1, a_2, a_3, a_4) = (\mathbf{M}_{a_1}p_3\mathbf{N}_{a_1}(a_2, a_3, a_4), a_1) \tag{2}$$

$$= (a_4, (\mathbf{N}_{a_4})^{-1}p_3(\mathbf{M}_{a_4})^{-1}(a_1, a_2, a_3)). \tag{3}$$

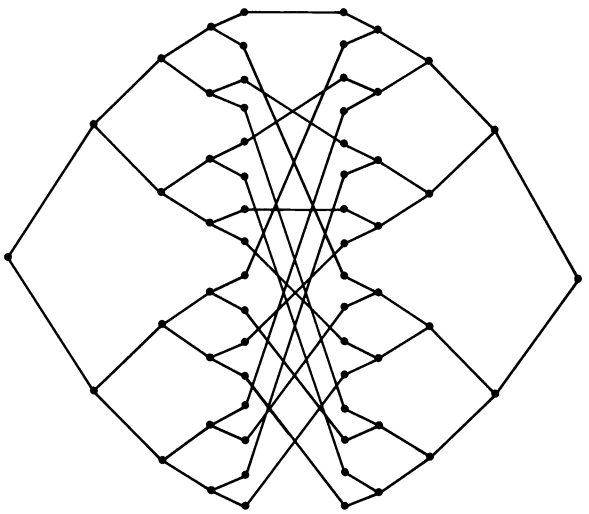


FIG. 6

To simplify the notation we denote the sets $\{1\}$, $\{2\}$, $\{1, 2\}$ by 1, 2, 12, respectively. Equating the values in the first coordinates of (2) and (3), we obtain

$$a_4 = a_4 + a_1a_2n_{3,1} + a_1a_3n_{3,2} + a_1a_2a_3n_{3,12} + a_1m_{1\phi},$$

$$\text{for any} \quad a_i \in \{0, 1\},\ 1 \leqq i \leqq 4.$$

Thus, we have $m_{1\phi} = n_{3,1} = n_{3,2} = n_{3,12} = 0$.

Suppose we replace $a_4$ by 0. The value of the second coordinates in (2) and (3) is

$$a_3 = a_3 + m_{2\phi}a_1 + n_{2,1}a_1a_2.$$

Therefore $m_{2\phi} = n_{2,1} = 0$ and $\mathbf{N}$ is then the identity function. Let us write

$$\mathbf{M}(x_1, x_2, x_3) = (x_1, x_2 + m_{2,1}x_1, x_3 + m_{3\phi} + m_{3,1}x_1 + m_{3,2}x_2 + m_{3,12}x_1x_2).$$

Then we have

$$\mathbf{M}^{-1}(x_1, x_2, x_3) = (x_1, x_2 + m_{2,1}x_1, x_3 + m_{3\phi}$$

$$+ (m_{3,1} + m_{2,1}m_{3,2} + m_{2,1}m_{3,12})x_1 + m_{3,2}x_2 + m_{3,12}x_1x_2).$$

We consider the value of the second coordinates in (2) and (3). We have

$$a_3 + a_2a_4 + m_{2,1}a_1a_4 = a_3 + m_{3\phi}\,a_4 + (m_{3,1} + m_{2,1}m_{3,2} + m_{2,1}m_{3,12})a_4a_1$$

$$+ m_{3,2}a_2a_4 + m_{3,12}a_1a_2a_4.$$

Therefore we have $m_{3,1} = m_{3\phi} = m_{3,12} = 0$, $m_{3,2} = 1$.

Now we consider the third coordinate in (2) and (3). We have $a_2 + a_1a_3 + a_1a_2a_4 = a_2 + m_{2,1}a_1a_4 + a_1a_3 + a_1a_2a_4$. Therefore we have $m_{2,1} = 0$, $\mathbf{M} = \mathbf{L}$ and $f = p_4$. Theorem 2 is proved.

It should be noted that the 10-stage HS-optimal binary graph is not optimal since in the case that $p_1 = p_2 = p_8 = p_9 = 0$, then the graph can be reduced and shrunken to a 4-stage graph in $T_8$ (similar to what we did to $G_8$ in the proof of Theorem 1). However, the 16 middle-stage links form two cycles instead of one. Since there does exist a 10-stage binary graph which, after similar reductions and shrinking, yields a graph in $T_8$ with the middle-stage links forming a cycle, then by Takagi's Theorem the HS-optimal 10-stage binary graph is not optimal. Furthermore, it can be shown (by straightforward arguments) that there exists only one (up to isomorphism) hereditary graph which is not symmetric. Again by using Takagi's Theorem it can be seen that this graph is not optimal. However, by Lemma 2, an optimal 10-stage binary graph must be hereditary. Therefore, no optimal 10-stage binary graph exists.

**6. $2n$-stage binary graphs for $n \geqq 5$.** The main result in this paper is the following.

THEOREM 3. *The* HS*-optimal* $(2n + 2)$*-stage binary graph has the connection function (up to isomorphism) as follows:*

$$p_n(a_1, \cdots, a_n) = (b_1, \cdots, b_n),$$

where

$$b_1 = a_n,$$

$$b_i = a_{n+i-1} + a_{n-i}b_{i-1} \pmod 2, \text{ for } i = 2, \cdots, n - 1,$$

$$b_n = a_1.$$

*Proof.* First we define

$$\mathbf{L}(x_1, \cdots, x_i) = (x_1, \cdots, x_{i-1}, x_i + x_{i-1}).$$

It can be verified by induction that

$$\begin{aligned} p_n(a_1, \cdots, a_n) &= (\mathbf{L}_{a_1} p_{n-1}(a_2, \cdots, a_n), a_1) \\ &= (a_n, p_{n-1}\mathbf{L}_{a_n}(a_1, \cdots, a_{n-1})). \end{aligned}$$

The binary graph determined by $p_n$ is thus symmetric and hereditary.

Now we consider a connection function $g$ of HS-optimal $2(n+1)$-stage binary graph $G$. From Lemmas 8 and 9, $g$ is equivalent to an $f$ which has the following form

$$f(a_1, \cdots, a_n) = (\mathbf{M}_{a_1} p_{n-1} \mathbf{N}_{a_1}(a_2, \cdots, a_n), a_1) \tag{4}$$

$$= (a_n, (\mathbf{N}_{a_n})^{-1} p_{n-1} (\mathbf{M}_{a_n})^{-1}(a_1, \cdots, a_{n-1})). \tag{5}$$

We want to show by induction on $n$ that:

(*): if $f$ can be written in the form of (4) and (5), then $\mathbf{N}$ is the identity mapping and $\mathbf{M} = \mathbf{L}$.

(*) is true for $n = 4$ from Theorem 2. Suppose (*) is true for all $n'$ with $4 \leqq n' < n$. We will prove that (*) holds for $n$.

The value in the first coordinates in (4) and (5) is

$$a_n = a_n + \sum_{S \subseteq \{0, \cdots, n-1\}} c_{n,S} \pi(S), \quad \text{for all} \quad a_i \in \{0, 1\},\ 1 \leqq i \leqq n.$$

Thus we have $c_{n,S} = 0$, for all $S$.

We consider, $\mathbf{M}'$, $\mathbf{N}'$ and $\mathbf{H}$ to be mappings from $\{0, 1\}^{n-2} \to \{0, 1\}^{n-2}$ such that

$$\begin{aligned} \mathbf{M}(0, x_2, \cdots, x_{n-1}) &= (0, \mathbf{M}'(x_2, \cdots, x_{n-1})), \\ \mathbf{N}(x_1, \cdots, x_{n-1}) &= (\mathbf{N}'(x_1, \cdots, x_{n-2}), x_{n-1}), \\ \mathbf{H}(x_1, \cdots, x_{n-2}) &= (x_1 + m_{21}, x_2 + m_{31}, \cdots, x_{n-2} + m_{n-1,1}). \end{aligned}$$

It is easy to see that

$$\begin{aligned} \mathbf{M}(x_1, \cdots, x_{n-1}) &= (x_1, \mathbf{H}_{x_1}\mathbf{M}'(x_2, \cdots, x_{n-1})), \\ \mathbf{M}^{-1}(x_1, \cdots, x_{n-1}) &= (x_1, \mathbf{M}'^{-1}\mathbf{H}_{x_1}(x_2, \cdots, x_{n-1})), \\ \mathbf{N}^{-1}(x_1, \cdots, x_{n-1}) &= (\mathbf{N}'^{-1}(x_1, \cdots, x_{n-2}), x_{n-1}). \end{aligned}$$

By induction $p_{n-1}(x_1, \cdots, x_{n-1}) = (x_{n-1}, p_{n-2}\mathbf{L}_{x_{n-1}}(x_1, \cdots, x_{n-2}))$. Therefore (4) is equal to the following:

$$\begin{aligned} &((\mathbf{M}_{a_1}(a_n, p_{n-2}\mathbf{L}_{a_n}\mathbf{N}'_{a_1}(a_2, \cdots, a_{n-1}))), a_1) \\ &\quad = (a_n, \mathbf{H}_{a_1 a_n}\mathbf{M}'_{a_1} p_{n-2}\mathbf{L}_{a_n}\mathbf{N}'_{a_1}(a_2, \cdots, a_{n-1}), a_1). \end{aligned}$$

On the other hand, (5) is equal to the following:

$$\begin{aligned} &(a_n, \mathbf{N}_{a_n}^{-1}(\mathbf{L}_{a_1} p_{n-2}\mathbf{M}_{a_n}'^{-1}\mathbf{H}_{a_1 a_n}(a_2, \cdots, a_{n-1}), a_1)) \\ &\quad = (a_n, \mathbf{N}_{a_n}^{-1}\mathbf{L}_{a_1} p_{n-2}\mathbf{M}_{a_n}'^{-1}\mathbf{H}_{a_1 a_n}(a_2, \cdots, a_{n-1}), a_1). \end{aligned}$$

Therefore we have

$$\mathbf{H}_{a_1 a_n}\mathbf{M}'_{a_1} p_{n-2}\mathbf{L}_{a_n}\mathbf{N}'_{a_1} = \mathbf{N}_{a_n}^{-1}\mathbf{L}_{a_1} p_{n-2}\mathbf{M}_{a_n}'^{-1}\mathbf{H}_{a_1 a_n}. \tag{6}$$

By setting $a_n = 0$ in (6), we have

$$\mathbf{M}'_a p_{n-2} \mathbf{N}'_{a_1} = \mathbf{L}_a p_{n-2}.$$

From the above equality and the definition of $p_{n-2}$ and $\mathbf{L}$ we define $f'$ as follows:

$$\begin{aligned} f'(x_1, \cdots, x_{n-1}) &= (\mathbf{M}'_{x_1} p_{n-2} \mathbf{N}'_{x_1}(x_2, \cdots, x_{n-1}), x_1) \\ &= (\mathbf{L}_{x_1} p_{n-2}(x_2, \cdots, x_{n-1}), x_1) \\ &= (x_{n-1}, p_{n-2} \mathbf{L}_{x_{n-1}}(x_1, \cdots, x_{n-2})) \\ &= (x_{n-1}, \mathbf{N}'^{-1}_{x_{n-1}} p_{n-1} \mathbf{M}'^{-1}_{x_{n-1}}(x_1, \cdots, x_{n-2})). \end{aligned}$$

By the induction assumptions, we have $\mathbf{M}'_x = \mathbf{L}_x$ and $\mathbf{N}'_x$ is the identity mapping. Therefore $\mathbf{N}$ is the identity mapping and (6) is equivalent to the following:

$$\mathbf{H}_{a_1 a_n} \mathbf{L}_{a_1} p_{n-2} \mathbf{L}_{a_n} = \mathbf{L}_{a_1} p_{n-2} \mathbf{L}_{a_n} \mathbf{H}_{a_1 a_n}. \tag{7}$$

By setting $a_1 = a_n = a$ in (7), we have

$$\mathbf{H}_a \mathbf{L}_a p_{n-2} \mathbf{L}_a = \mathbf{L}_a p_{n-2} \mathbf{L}_a \mathbf{H}_a,$$

i.e.,

$$\mathbf{H}_a \mathbf{L}_a p_{n-2} \mathbf{L}_a \mathbf{H}_a \mathbf{L}_a = \mathbf{L}_a p_{n-2}.$$

Now we consider $f''$ as follows:

$$\begin{aligned} f''(x_1, \cdots, x_{n-1}) &= (\mathbf{H}_{x_1} \mathbf{L}_{x_1} p_{n-2} \mathbf{L}_{x_1} \mathbf{H}_{x_1} \mathbf{L}_{x_1}(x_2, \cdots, x_{n-1}), x_1) \\ &= (\mathbf{L}_{x_1} p_{n-2}(x_2, \cdots, x_{n-1}), x_1) \\ &= (x_{n-1}, p_{n-2} \mathbf{L}_{x_{n-1}}(x_1, \cdots, x_{n-2})) \\ &= (x_{n-1}, \mathbf{L}_{x_{n-1}} \mathbf{H}_{x_{n-1}} \mathbf{L}_{x_{n-1}} p_{n-2} \mathbf{L}_{x_{n-1}} \mathbf{H}_{x_{n-1}}(x_1, \cdots, x_{n-2})). \end{aligned}$$

As before, we have, by the induction assumptions, $\mathbf{H}_a \mathbf{L}_a = \mathbf{L}_a$ and $\mathbf{L}_a \mathbf{H}_a \mathbf{L}_a$ is the identity. This implies $\mathbf{H}_a$ is the identity and $\mathbf{M} = \mathbf{L}$. Therefore we have

$$\begin{aligned} f(a_1, \cdots, a_n) &= (\mathbf{L}_{a_1} p_{n-1}(a_2, \cdots, a_n), a_1) = (a_n, p_{n-1} \mathbf{L}_{a_n}(a_1, \cdots, a_{n-1})) \\ &= (b_1, \cdots, b_n). \end{aligned}$$

Theorem 3 is then proved.

**7. Construction of switching networks with prescribed channel graphs.** We will present several constructions of switching networks with prescribed channel graphs.

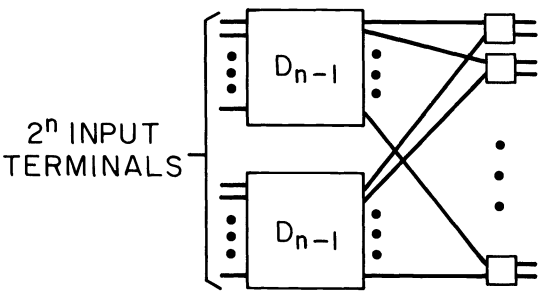



FIG. 7

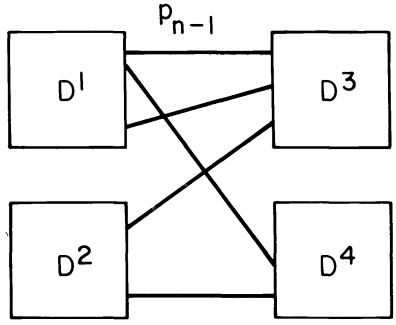



FIG. 8

Let us first consider a $2n$-stage network $N_{2n}$ consisting of $2 \times 2$ switches. Let $D_1$ denote a $2 \times 2$ switch and $D_n$ be the network built recursively as shown in Fig. 7. The $2n$-stage network $N_{2n}$ consists of 2 copies of $D_n$ denoted by $D^1$, $D^2$ together with 2 copies denoted by $D^3$ and $D^4$, of the mirror image of $D_n$ as shown in Fig. 8.

The linking pattern between the middle two stages is determined by the connection function of the channel graph $p_{n-1}$. Namely, the $i$th output switch of $D_1$ or $D_2$ is connected to the $j$th input switch of $D^3$ or $D^4$ if and only if $i - 1$ and $j - 1$ have the binary expressions $(a_1, \cdots, a_{n-1})$, $(b_1, \cdots, b_{n-1})$ respectively, and $p_{n-1}(a_1, \cdots, a_{n-1}) = (b_1, \cdots, b_{n-1})$.

It can be easily verified that the channel graph of $N_{2n}$ is the HS-optimal binary graph $G_{2n}$.

Let us now consider networks with an odd number of stages. We note that a binary graph of $2n + 2$ stages can be viewed as a binary graph of $2n + 1$ stages by shrinking the two middle stages. We can then derive HS-optimal binary graphs $G_{2n+1}$ from the HS-optimal graphs $G_{2n+2}$. For example $G_5$ is as shown in Fig. 9.

Now, we want to construct a $2n + 1$ stage network with channel graph $G_{2n+1}$. $G_{2n+1}$ consists of a copy of $D_{n+1}$ together with two copies, denoted by $D^3$, $D^4$, of the mirror image of $D_n$ as shown in Fig. 10.

The linking pattern between stage $n + 1$ and stage $n + 2$ is determined by $p_n$. Namely, the $i$th output switch of $D_{n+1}$ is connected to the $j$th input line of $D^3$ or $D^4$ if and only if $i - 1$ and $j - 1$ have the binary expressions $(a_1, \cdots, a_n)$, $(b_1, \cdots, b_n)$, respectively, and $p_n(a_1, \cdots, a_n) = (b_1, \cdots, b_n)$. It is easily verified that $N_{2n+1}$ has channel graph $G_{2n+1}$.

**8. Concluding remarks.** We can generalize the ideas in this paper and consider the connection pattern of two $t$-ary graphs. A $(2n + 2)$-stage Ikeno graph is the $t$-ary graph determined by the connection function $f(x_1, \cdots, x_n) = (x_n, \cdots, x_1)$

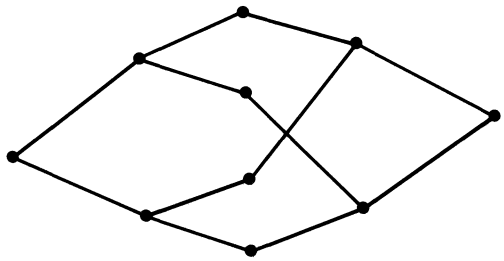


FIG. 9

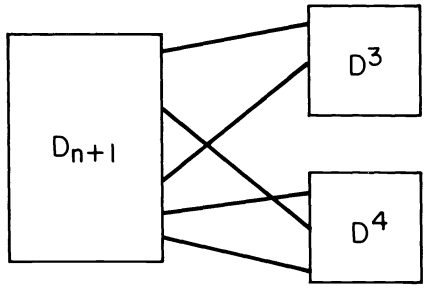



FIG. 10

where the leaves of the left and right $t$-ary trees are labeled by elements in $\{0, \cdots, t-1\}$. Again it can be shown that the Ikeno graph is not optimal for $t \geqq 3$, $n \geqq 3$ by considering an alternative connection function defined in a similar way as we defined $p_n$ in previous sections. The connection patterns of $t$-ary graphs are probably more complicated than the binary case.

## REFERENCES

[1] N. IKENO, *A limit on crosspoint number,* IRE Trans. Inform. Theory, Vol. IT-5 (1959), pp. 187–196.

[2] P. LEGALL, *Etude du blocade dans les systèmes de commutation téléphonique,* Ann. Télécommun., 11, No. 9 (1956).

[3] V. I. NEIMAN, *Structural properties of connecting networks,* 8th ITC, Milbourne (1976), pp. 131-1 to 131-8.

[4] K. TAKAGI, *Design of multi-stage link systems by means of optimal channel graphs,* Electronics and Communications in Japan, Vol. 51-A (1968), pp. 37–46.

[5] ———, *Optimum channel graph of link system,* Electronics and Communications in Japan, Vol. 54-A (1971), pp. 1–10.

[6] ———, *A comparison of channel graphs for link system design,* Trans. IECE Japan, Vol. E61, No. 7 (1978), pp. 538–539.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0011 $01.00/0

# PLANAR LEAST-SQUARES INVERSE POLYNOMIALS. PART II: ASYMPTOTIC BEHAVIOR*

PH. DELSARTE,† Y. GENIN,† AND Y. KAMP†

**Abstract.** This paper contains a study of the limit function $a(z_1, z_2)$ of the PLSI polynomials relative to a given $H_2$-function $b(z_1, z_2)$. It is shown that $a(z_1, z_2)$ is analytic in the unit bicylinder, and that $b(z_1, z_2)$ admits a canonical weakly inner-strongly outer factorization if and only if $a(z_1, z_2)$ enjoys a well-defined property of stability. The theory is illustrated by a detailed example.

**1. Introduction.** A preceding and companion paper [2] has been devoted to a study of the algebraic properties of the planar least-squares inverse (PLSI) polynomials $a_{k,l}(z_1, z_2)$ relative to a given polynomial $b(z_1, z_2)$. These $a_{k,l}$ appear as solutions of doubly Toeplitz systems built on the Fourier coefficients of $|b(e^{i\theta}, e^{i\phi})|^2$. The aim of the present paper is to discuss the convergence of the PLSI polynomials when both degrees $k$ and $l$ tend to infinity; thus to show the existence and to study the properties of the function $a(z_1, z_2) = a_{\infty,\infty}(z_1, z_2)$. In fact, the Hardy class $H_2$ provides the natural setting for the problem, so that $b(z_1, z_2)$ is herein allowed to be any element of $H_2$ (not necessarily a polynomial as usual).

This asymptotic PLSI problem, which turns out to be very difficult, is of primary importance both for theoretical and for practical reasons as explained hereafter. For the purpose of stabilization in two-dimensional digital filtering, Shanks, Treitel, and Justice [15] have made the conjecture that the PLSI polynomials are always stable, i.e., devoid of zeros in $|z_{1,2}| \leqq 1$. (Throughout this paper, a notation like $|z_{1,2}| \leqq 1$ stands for $|z_1| \leqq 1$ and $|z_2| \leqq 1$.) However, this property has been proved to fail [6] (although it holds for one-variable polynomials). Further analysis has led to weaker conjectures [2], [12], the weakest and most tractable of which seems to be the following: the limit function $a(z_1, z_2)$ has no zeros in $|z_{1,2}| < 1$.

Let us now turn to more theoretical aspects. In the one-variable situation the asymptotic solution of positive definite Toeplitz systems is directly related to the spectral factorization problem, both in the scalar case [7] and in the matrix case [1], [18]. It is not known, however, in what form these results might extend to the two-variable situation. In particular, as the classical inner-outer factorization [10], [17] of a one-variable function $b(z) \in H_2$ can be obtained from the asymptotic solution of the corresponding least-squares inverse problem, the question naturally arises of a possible extension of this property in two dimensions.

The above remarks put into light the fact that the study of PLSI polynomials tackles fundamental unsolved questions in the theory of two-variable functions. It should be observed that the subject dealt with is closely related to the factorization problem for bivariate spectra recently introduced by Strintzis [16], but reduces neither to the "halfplane factorization" of Helson and Lowdenslager [9] nor to the "four factors factorization" studied by Ekstrom and Woods [5] in the context of two-dimensional recursive filtering. Another topic which is connected to the present study is that of the invertibility of quarterplane Toeplitz operators [3], [4], [13]; the results obtained in this framework have a high degree of generality but do not seem to yield answers to the specific questions of the convergence and asymptotic stability of the PLSI polynomials.

* Received by the editors July 16, 1979, and in final form February 5, 1980.
† Philips Research Laboratory, Av. Van Becelaere 2, B-1170, Brussels, Belgium.

Let us now briefly summarize the content of this paper. Some classical material from the theory of two-variable functions of class $H_2$ is given in § 2. The PLSI problem relative to a given function $b(z_1, z_2) \in H_2$ is described in § 3; the minimizing polynomials $a_{k,l}(z_1, z_2)$ are defined and characterized. The main result of § 4 is a proof of the convergence of the $a_{k,l}(z_1, z_2)$ towards a function $a(z_1, z_2)$ analytic in the bicylinder $|z_{1,2}| < 1$. In § 5, the significance of the concept of strongly outer functions (in the sense of Helson [8]) is emphasized; $b$ turns out to be strongly outer if and only if $a$ equals $b^{-1}$. A "dual" definition of weakly inner functions is introduced and discussed in § 6; it is shown that the product $ab$ is weakly inner provided it is bounded. These concepts are used in § 7, where the function $b(z_1, z_2)$ is proved to admit a canonical weakly inner-strongly outer factorization if and only if the corresponding $a(z_1, z_2)$ enjoys a well-defined property of stability. As an illustration, a two-parameter class of first degree polynomials $b$ is studied in § 8; the desired canonical factorization is proved to exist and explicitly obtained, except for a small region of the whole parameter space where the question remains open.

As a conclusion, it is fair to say that this paper gives only partial answers to the many questions it raises. Nevertheless, the authors think that their contribution at least provides an adequate mathematical setting for the difficult problem of asymptotic PLSI stability and shows the primary importance of this problem in connection with canonical factorization of $H_2$-functions.

**2. Definitions and preliminaries.** To start with, let us recall certain standard concepts and results from the theory of analytic functions in two variables [14], [19]. Let $H_0$ denote the vector space of the complex functions $f(z_1, z_2)$ analytic in the *unit bicylinder* $|z_{1,2}| < 1$. The subset of $H_0$ consisting of the functions $f$ satisfying

$$\|f\| = \sup_{0 \leqq r_{1,2} < 1} \left\{ \frac{1}{4\pi^2} \int\int_0^{2\pi} |f(r_1 e^{i\theta}, r_2 e^{i\phi})|^2 \, d\theta \, d\phi \right\}^{1/2} < \infty, \tag{1}$$

is classically denoted by $H_2$. In fact, $H_2$ is a linear subspace of $H_0$ and, for the definition (1) of the $H_2$-norm, it has the structure of a Banach space. Thus any Cauchy sequence out of $H_2$ converges in $H_2$. Note that the $H_2$-convergence implies the usual $H_0$-convergence on every compact subset of the unit bicylinder; this immediately follows from the *Poisson inequality,*

$$|f(z_1, z_2)|^2 \leqq \|f\|^2 \frac{1 + |z_1|}{1 - |z_1|} \cdot \frac{1 + |z_2|}{1 - |z_2|}. \tag{2}$$

Let $P$ denote the space of all polynomials in $z_1, z_2$. A fundamental property of the space $H_2$ is the fact that $P$ is dense in $H_2$. Indeed, for any $f \in H_2$, the truncated Maclaurin expansion $f_{m,n}$ converges to $f$.

By the Fatou theorem, every function $f$ in $H_2$ has a *radial limit* almost everywhere (a.e.) on the *torus* $z_1 = e^{i\theta}$, $z_2 = e^{i\phi}$; thus

$$f(e^{i\theta}, e^{i\phi}) = \lim_{r_{1,2} \to 1^-} f(r_1 e^{i\theta}, r_2 e^{i\phi}) \tag{3}$$

exists a.e. Moreover, $f(e^{i\theta}, e^{i\phi})$ is square integrable on the torus. In view of this property, a scalar product can be defined in $H_2$ by

$$\langle f, g \rangle = \frac{1}{4\pi^2} \int\int_0^{2\pi} f(e^{i\theta}, e^{i\phi}) g^*(e^{i\theta}, e^{i\phi}) \, d\theta \, d\phi, \tag{4}$$

for all $f, g \in H_2$. It is well known that the norm attached to (4) is the same as (1), i.e., $\|f\| = \langle f, f \rangle^{1/2}$. Let us also recall that $\log |f|$ is integrable on the torus provided $f$

is not identically zero. The notation $f \perp g$ will be used to mean that the functions $f$ and $g$ are *orthogonal,* in the sense that their scalar product (4) vanishes.

The space $H_\infty$ is defined to consist of the functions $f \in H_0$ which are bounded in the unit bicylinder, i.e., satisfy

$$N(f) = \sup_{|z_{1,2}|<1} |f(z_1, z_2)| < \infty. \tag{5}$$

It is clear from (1) and (5) that $H_\infty$ is a subspace of $H_2$, with $\|f\| \leqq N(f)$, enjoying the properties $H_\infty H_\infty = H_\infty$ and $H_\infty H_2 = H_2$. In fact, the supremum norm (5) gives $H_\infty$ the structure of a Banach algebra (for ordinary multiplication).

**3. The PLSI problem.** The planar least-squares inverse polynomial problem can be stated in the following form. Given a nonzero function $b(z_1, z_2)$ of class $H_2$, it is required to minimize the norm $\|1 - pb\|$, i.e., the distance between 1 and $pb$, for $p(z_1, z_2)$ varying over the space $P_{k,l}$ consisting of all polynomials of formal degree $k$ in $z_1$ and $l$ in $z_2$. It can be easily shown that this problem has a unique solution $p(z_1, z_2)$, denoted by $a_{k,l}(z_1, z_2)$ in the sequel. In fact, the coefficients of $a_{k,l}$ appear as the solution of a linear system, the matrix of which is the positive definite *doubly Toeplitz matrix* built on the Fourier coefficients of $|b(e^{i\theta}, e^{i\phi})|^2$. For details, the reader is referred to [2]. (Only the case $b \in P$ is treated there but the generalization to $b \in H_2$ is straightforward.)

As pointed out in [2], the minimizing polynomial $a_{k,l}$ is characterized by the orthogonality of the function $1 - a_{k,l}b$ to the space of functions $pb$ with $p \in P_{k,l}$, i.e.,

$$1 - a_{k,l}b \perp P_{k,l}b. \tag{6}$$

Geometrically, this property says that $a_{k,l}b$ is the nearest point to 1 in the space $P_{k,l}b$, i.e., the orthogonal projection of 1 onto that space. An immediate consequence of (6) is the *Pythagoras identity*

$$\|1 - pb\|^2 = \|1 - a_{k,l}b\|^2 + \|(a_{k,l} - p)b\|^2, \tag{7}$$

for all $p(z_1, z_2) \in P_{k,l}$. Note that $a_{k,l}(z_1, z_2)$ is identically zero when $b(0, 0)$ vanishes (which means $1 \perp b$). Henceforth we shall exclude this case, and always assume $b(0, 0) \neq 0$.

Let $\mu_{k,l}(b)$ be the squared distance between 1 and the space $P_{k,l}b$, i.e., by definition,

$$\mu_{k,l}(b) = \|1 - a_{k,l}b\|^2. \tag{8}$$

From (7) and (8) one deduces $\mu_{k,l}(b) \leqq \|1 - \lambda b\|^2$, for all constant $\lambda$. Since $P_{s,t}$ is a subspace of $P_{k,l}$ for $s \leqq k$ and $t \leqq l$, one has $\mu_{k,l}(b) \leqq \mu_{s,t}(b)$, which proves that the double sequence $[\mu_{k,l}(b)]$ is monotonically decreasing and hence convergent. The limit value will be denoted by $\mu(b)$. By definition,

$$\mu(b) = \lim_{k,l\to\infty} \mu_{k,l}(b) = \inf_{u\in P} \|1 - ub\|^2. \tag{9}$$

Thus $\mu(b)$ appears as the *residual squared error* in the polynomial inverse approximation of $b$.

**4. Convergence of the minimizing polynomials.** Let us first study the convergence properties of the double sequence of functions $g_{k,l} \in H_2$ defined by

$$g_{k,l}(z_1, z_2) = a_{k,l}(z_1, z_2)b(z_1, z_2). \tag{10}$$

The following theorem is an almost immediate consequence of the fact that $H_2$ is a Banach space.

THEOREM 1. *The double sequence* $[g_{k,l}(z_1, z_2)]$ *has a limit* $g(z_1, z_2)$ *in* $H_2$, *satisfying* $\|1 - g\|^2 = \mu(b)$.

*Proof.* Let $k \geqq s \geqq 0$ and $l \geqq t \geqq 0$. Applying (7) with $p = a_{s,t}$, one deduces from (8) and (10) the identity

$$\|g_{k,l} - g_{s,t}\|^2 = \mu_{s,t}(b) - \mu_{k,l}(b). \tag{11}$$

Since $[\mu_{k,l}(b)]$ converges, the right member of (11) can be made arbitrarily small for $k$, $l$, $s$, and $t$ sufficiently large. As a consequence, the double sequence $[g_{k,l}]$ has the Cauchy property; hence it converges (for the $H_2$-norm) to a well-defined function $g \in H_2$. Then $\mu(b) = \|1 - g\|^2$ immediately follows from taking the limit in (8). □

THEOREM 2. *The function* $g = \lim g_{k,l}$ *enjoys the following property:* $1 - g$ *is orthogonal to every product* $vg$ *with* $v$ *polynomial, i.e.,*

$$1 - g \perp Pg. \tag{12}$$

*Proof.* Let $u(z_1, z_2)$ be a polynomial. From (6) one derives $1 - g_{k,l} \perp ub$, provided $P_{k,l}$ contains $u$; hence $1 - g \perp ub$ in the limit. Specializing this to $u = va_{s,t}$ for a given $v \in P$, one has $1 - g \perp vg_{s,t}$ and finally $1 - g \perp vg$ by letting $s \to \infty$, $t \to \infty$. □

As a direct consequence of (12), one deduces $\|1 - g\|^2 = 1 - g(0, 0)$. In view of Theorem 1, this yields

$$\mu(b) = 1 - g(0, 0). \tag{13}$$

We next prove the convergence of the minimizing polynomials $a_{k,l}$. It should be noted that the property is weaker than for the $g_{k,l}$; in general, $[a_{k,l}]$ is not $H_2$-convergent.

THEOREM 3. *The double sequence* $[a_{k,l}(z_1, z_2)]$ *converges in the unit bicylinder* $|z_{1,2}| < 1$ *to an analytic function* $a(z_1, z_2) \in H_0$, *satisfying* $a(z_1, z_2)\, b(z_1, z_2) = g(z_1, z_2)$.

*Proof.* The argument consists in showing that, for any fixed $\zeta_1$, with $|\zeta_1| < 1$, the double sequence $[a_{k,l}(\zeta_1, z_2)]$ is uniformly convergent on every closed disk $|z_2| \leqq r_2 < 1$. Hence the limit $a(\zeta_1, z_2)$ is analytic in $|z_2| < 1$. Since the dual result holds for $[a_{k,l}(z_1, \zeta_2)]$ with $|\zeta_2| < 1$, it then follows from the Hartogs theorem [11] that $a(z_1, z_2)$ is analytic in the domain $|z_{1,2}| < 1$.

For the given $\zeta_1$ there exists a function $b'(z_1, z_2)$ in $H_2$, satisfying $|b'(e^{i\theta}, z_2)| = |b(e^{i\theta}, z_2)|$ a.e., such that $b'(\zeta_1, z_2)$ is not identically zero in $z_2$. (This results from a well-known one-variable property.) Using (11) and the one-variable version of the Poisson inequality (2) one readily obtains

$$\frac{1}{2\pi}\int_0^{2\pi} |b'(z_1, e^{i\phi})[a_{k,l}(z_1, e^{i\phi}) - a_{s,t}(z_1, e^{i\phi})]|^2\, d\phi \leqq \frac{1 + |z_1|}{1 - |z_1|}\,[\mu_{s,t}(b) - \mu_{k,l}(b)]. \tag{14}$$

Next, let $f(z_2)$ be a function of class $H_2$, devoid of zeros in $|z_2| < 1$, satisfying $|f(e^{i\phi})| = |b'(\zeta_1, e^{i\phi})|$ a.e. (cf. [10]). Applying (14) with $z_1 = \zeta_1$ and again the Poisson inequality, one deduces

$$|f(z_2)|^2|a_{k,l}(\zeta_1, z_2) - a_{s,t}(\zeta_1, z_2)|^2 \leqq \frac{1 + |z_2|}{1 - |z_2|} \cdot \frac{1 + |\zeta_1|}{1 - |\zeta_1|} [\mu_{s,t}(b) - \mu_{k,l}(b)]. \tag{15}$$

Since $|f(z_2)|$ is bounded away from zero in $|z_2| \leqq r_2(<1)$, it follows from (15) and from the Cauchy criterion that the double sequence $[a_{k,l}(\zeta_1, z_2)]$ is uniformly convergent on $|z_2| \leqq r_2$. As explained before, this leads to the desired conclusion. □

**5. Outer functions.** A function $f(z_1, z_2) \in H_2$, with $f(0, 0) \neq 0$, will be called *strongly outer* if it satisfies $\mu(f) = 0$, i.e., if it can be inversely approximated by polynomials with arbitrarily small error. This concept coincides with the definition of an outer function as given by Helson [8]: a function $f \in H_2$ is (strongly) outer if and only if the space $Pf$ is dense in $H_2$.

It is known that, if $f$ is strongly outer, then $\log |f(e^{i\theta}, e^{i\phi})|$ has nonzero Fourier coefficients in the first and third quadrant only. Moreover, $f$ can be written as

$$f(z_1, z_2) = \gamma \exp \left\{ \frac{1}{4\pi^2} \int\int_0^{2\pi} \left[ \frac{2e^{i(\theta+\phi)}}{(e^{i\theta} - z_1)(e^{i\phi} - z_2)} - 1 \right] \cdot \log |f(e^{i\theta}, e^{i\phi})| \, d\theta \, d\phi \right\}, \tag{16}$$

for $|z_{1,2}| < 1$, where $\gamma$ is a constant of unit modulus [14]. In the sequel a function $f \in H_2$ is called *weakly outer* if it satisfies (16); this is what Rudin calls an outer function [14].

In the one-variable setting the definitions of strongly and weakly outer functions are equivalent [8]. This property is lost in the two-variable situation. In fact, Rudin has given an example of weakly outer functions which are not strongly outer [14, p. 76].

As a direct consequence of (16), a weakly outer function is devoid of zeros in the unit bicylinder. The following weak converse of this result holds true: if $f \in H_\infty$ and $f^{-1} \in H_2$, then $f$ is strongly outer. Indeed, one has $\mu(f) \leqq N(f)\|f^{-1} - u\|$, for all $u \in P$, and $\|f^{-1} - u\|$ can be made arbitrarily small, so that $\mu(f)$ vanishes. For example, any polynomial $f(z_1, z_2)$ devoid of zeros in $|z_{1,2}| \leqq 1$ is a strongly outer function. On the other hand, it is known that a polynomial $f$ devoid of zeros in $|z_{1,2}| < 1$ is weakly outer [14]. It would be interesting to know whether, in the latter situation, $f$ is strongly outer. Let us finally point out an obvious consequence of Theorem 1.

THEOREM 4. *A given function $b(z_1, z_2)$ is strongly outer if and only if $g(z_1, z_2)$ identically equals* 1.

**6. Inner functions.** A function $g(z_1, z_2) \in H_2$, with $g(0, 0) \neq 0$, will be called *weakly inner* if it is bounded (i.e., $g \in H_\infty$) and satisfies

$$1 - g \perp H_2 g. \tag{17}$$

In view of the characterization (6) of PLSI polynomials, this property can be formulated as follows: a function $b \in H_\infty$ is weakly inner if and only if all minimizing polynomials relative to $b$ reduce to $a_{k,l} = 1$ or, equivalently, if and only if $\mu(b)$ equals $\|1 - b\|^2$. Thus weakly inner and strongly outer functions have opposite properties: they make $\mu(b)$ respectively maximum $(= \|1 - b\|^2)$ and minimum $(=0)$. Note that the only weakly inner polynomial is the constant $b = 1$.

A function $g(z_1, z_2) \in H_2$, with $g(0, 0) \neq 0$, will be said to be *strongly inner* if it satisfies $|g(e^{i\theta}, e^{i\phi})|^2 = g(0, 0)$ a.e. So, within a constant factor, $g$ is an inner function in the usual sense [14]. A strongly inner function obviously is weakly inner. In

the one-variable situation both definitions turn out to be exactly equivalent. However, this is not true in the two-variable problem; as indicated in § 8, there do exist weakly inner functions which are not strongly inner. The product of a strongly inner function by a weakly inner function clearly is weakly inner, but examples show that the product of two weakly inner functions is generally not weakly inner.

The importance of the concept of weakly inner functions in the present study is clear. Indeed, when $g$ belongs to $H_\infty$, the condition (17) is equivalent to $1 - g \perp Pg$, so that Theorem 2 can be formulated as follows.

THEOREM 5. *The function* $g = \lim a_{k,l}b$ *is weakly inner if and only if it is bounded.*

**7. Canonical factorization of $H_2$-functions.** The celebrated Beurling theorem says that any nonzero one-variable function $b(z) \in H_2$ can be canonically written as $b(z) = f(z)g(z)$, with $f$ outer and $g$ inner [10]. In addition, when $b(0) \neq 0$, it is known that $f^{-1}$ is the limit of the minimizing polynomials relative to $b$ (see [7]). The next theorem tackles the problem of a possible extension of these properties to the two-variable situation.

THEOREM 6. *A function* $b(z_1, z_2) \in H_2$, *with* $b(0, 0) \neq 0$, *can be expressed as the product of a strongly outer function by a weakly inner function if and only if the limit* $a$ *of the minimizing polynomials* $a_{k,l}$ *relative to* $b$ *has a strongly outer inverse* $a^{-1} \in H_2$ *and, in addition, makes the product* $g = ab$ *bounded. In this case, the strongly outer and weakly inner factors of* $b$ *are uniquely determined as* $a^{-1}$ *and* $g$, *respectively.*

*Proof.* Let us first assume $b = fh$ with $f$ strongly outer and $h$ weakly inner. Since $1 - h$ is assumed to be orthogonal to the space $H_2h$, one immediately obtains the identity

$$\|1 - ub\|^2 = \|1 - h\|^2 + \|h - ub\|^2, \tag{18}$$

valid for all functions $u$ of class $H_\infty$. Applying (18) successively to $u = a_{k,l}$ and to $u =$ any $v_{k,l} \in P_{k,l}$, one has

$$\|h - a_{k,l}b\| \leqq \|h - v_{k,l}b\| \leqq N(h)\|1 - v_{k,l}f\|, \tag{19}$$

as a consequence of $\|1 - a_{k,l}b\| \leqq \|1 - v_{k,l}b\|$ and $h \in H_\infty$. Taking for $v_{k,l}$ the minimizing polynomial relative to $f$, one deduces $\|h - g_{k,l}\|^2 \leqq N^2(h)\mu_{k,l}(f)$ from (19). Therefore, since $\mu_{k,l}(f)$ tends to zero, $g_{k,l}$ converges to $h$, which yields $g = h$; thus $afh = h$, i.e., $a = f^{-1}$.

Conversely, let us assume $a^{-1}$ to be strongly outer and $ab$ to be bounded. Define $f = a^{-1}$ and $g = ab$, hence $b = fg$. Since $g$ belongs to $H_\infty$, it is weakly inner on the strength of Theorem 5. This completes the proof. □

Thus the question of the appropriate "*Beurling factorization*" of $b$ is equivalent to that of the "*stability*" of $a$ (relative to $b$) defined as follows: $a^{-1}$ strongly outer and $ab$ bounded. It is then natural to ask whether any $b \in H_2$ admits a factorization of the required type. Unfortunately, the answer to this question is negative for the simple reason that there exist functions $b \in H_2$ for which it is impossible to find any nonzero function $h \in H_\infty$ vanishing at the zeros of $b$ in the unit bicylinder [14, p. 60]. In particular, $g = ab$ is not bounded in this situation. Moreover, although this is generally conjectured to be true at least for $b \in P$, the fact that $a$ is devoid of zeros in the unit bicylinder is still today an open question.

On the basis of these rather disappointing observations, one is led to the following problem: to characterize the subclass of $H_2$ for which the canonical factorization

exists. The example studied in the last section not only shows the difficulty of the problem but also reveals the interest of the present approach.

**8. Example.** For arbitrary complex numbers $\alpha$ and $\beta$ let us consider the following two-variable polynomial:

$$b(z_1, z_2) = 1 + \alpha^* z_1 + \beta(\alpha + z_1)z_2. \tag{20}$$

Since a change of variables $z_1 \to e^{iu}z_1$, $z_2 \to e^{iv}z_2$ on $b$ induces the same change on the minimizing polynomials $a_{k,l}$ relative to $b$, it is sufficient to treat the cases where $\alpha$ and $\beta$ are real and nonnegative.

*Case 1.* $0 < \alpha < 1$, $0 < \beta \leqq 1$. Except for $\beta = 1$, the polynomial (20) is devoid of zeros in $|z_{1,2}| \leqq 1$, since $b(\pm 1, \pm 1) > 0$ holds, so that $b$ is strongly outer. It turns out that this conclusion is still valid when $\beta = 1$. As a consequence, one deduces

$$a = \frac{1}{b}, \qquad g = 1, \qquad \mu(b) = 0. \tag{21}$$

*Case 2.* $0 < \alpha < 1$, $\beta \geqq 1$. The reciprocal $b(z_1, z_2) = \beta[1 + \alpha z_1 + \beta^{-1}(\alpha + z_1)z_2]$ of $b$ is strongly outer and has same modulus as $b$ on the torus. Hence the results of Case 1 can be used to yield

$$a = \frac{1}{\beta \acute{b}}, \qquad g = \frac{b}{\beta \acute{b}}, \qquad \mu(b) = 1 - \frac{1}{\beta^2}, \tag{22}$$

with the help of (13). Thus here, as well as in Case 1, the functions $a^{-1}$ and $g$ obviously are strongly outer and strongly inner, respectively.

*Case 3.* $\alpha > 1$, $0 < \beta < 1$. Both $b$ and $\acute{b}$ have zeros in the unit bicylinder, so that the function $a$ is not obtainable by a direct argument. Using the approach contained in [2] one can explicitly determine $a_{\infty,l}(z_1, z_2)$, and then let $l$ tend to infinity in the result. Due to the particular form of the polynomial $b$, the matrix spectral factorization involved in this method can be easily performed. After somewhat tricky but elementary computations one finds

$$a = \sum_{t=1}^{\infty} \frac{\beta^{2t-2}(1 - \beta^{2t})}{\alpha - \alpha^{-1}\beta^{2t} + (1 - \beta^{2t})z_1} \left( \frac{1}{\alpha + \beta^{2t-1}z_2} - \frac{\beta^2}{\alpha + \beta^{2t+1}z_2} \right), \tag{23}$$

$$g = 1 - (\alpha^2 - 1)(1 - \beta^2) \sum_{t=1}^{\infty} \frac{\beta^{2t-2}}{\alpha - \alpha^{-1}\beta^{2t} + (1 - \beta^{2t})z_1} \cdot \frac{1}{\alpha + \beta^{2t-1}z_2}, \tag{24}$$

$$\mu(b) = (\alpha^2 - 1)(1 - \beta^2) \sum_{t=1}^{\infty} \frac{\beta^{2t-2}}{\alpha^2 - \beta^{2t}}. \tag{25}$$

The functions $a$ and $g$ clearly are analytic in $|z_1| < \alpha$, $|z_2| < \alpha/\beta$. Hence $g$ is weakly inner in view of Theorem 5. But $g$ is not strongly inner, as shown by direct numerical computation. Because of the lack of appropriate tools to localize the zeros of a series like (23), it is not easy to study the properties of $a^{-1}$. However, it can be shown that $a(z_1, z_2)$ is devoid of zeros in $|z_{1,2}| \leqq 1$, hence that $a^{-1}$ is strongly outer, provided $\alpha \geqq \beta^2 \sqrt{1 + \beta^2}$ holds; the argument is explained in the Appendix. In the small part of the region $\beta < 1 < \alpha$ where $\alpha < \beta^2 \sqrt{1 + \beta^2}$ holds, $a^{-1}$ is "likely" to be strongly outer as well, for this is true on the boundaries $\alpha = 1$ and $\beta = 1$ (see Cases 5 and 6).

*Case 4.* $\alpha > 1$, $\beta > 1$. By considering the polynomial $\hat{b}$ instead of $b$, one can apply the results of Case 3 and obtain the formulas

$$a = \sum_{t=1}^{\infty} \frac{\beta^{2t-1} - \beta^{-1}}{\alpha\beta^{2t} - \alpha^{-1} + (\beta^{2t} - 1)z_1} \left( \frac{1}{\alpha\beta^{2t-1} + z_2} - \frac{1}{\alpha\beta^{2t+1} + z_2} \right), \tag{26}$$

$$g = 1 - (\alpha^2 - 1)(\beta^2 - 1) \sum_{t=0}^{\infty} \frac{\beta^{2t-1}}{\alpha\beta^{2t} - \alpha^{-1} + (\beta^{2t} - 1)z_1} \cdot \frac{1}{\alpha\beta^{2t+1} + z_2}, \tag{27}$$

$$\mu(b) = (\alpha^2 - 1) \frac{\beta^2 - 1}{\beta^2} \sum_{t=0}^{\infty} \frac{1}{\alpha^2\beta^{2t} - 1}. \tag{28}$$

The main conclusions are that $g$ is always weakly inner, and that $a^{-1}$ is strongly outer when $\alpha\beta^3 \geqq \sqrt{1 + \beta^2}$ holds.

*Case 5.* $\alpha > 1$, $\beta = 1$. The solution can be obtained by the method sketched for Case 3. The results turn out to be rather simple:

$$a = \frac{1}{b} \left[ 1 - \frac{\alpha^2 - 1}{b} \log \left( 1 + \frac{b}{\alpha^2 - 1} \right) \right], \tag{29}$$

$$g = 1 - \frac{\alpha^2 - 1}{b} \log \left( 1 + \frac{b}{\alpha^2 - 1} \right), \tag{30}$$

$$\mu(b) = (\alpha^2 - 1) \log \frac{\alpha^2}{\alpha^2 - 1}. \tag{31}$$

In fact, although this is not obvious, the above formulas are the limits of (23)–(25) and of (26)–(28) for $\beta \to 1$. Note that $a$ and $g$ are analytic in $|z_{1,2}| < \alpha$. The function $g$ is readily seen not to be strongly inner, but $g$ is of course weakly inner in view of Theorem 5. On the other hand, it is not difficult to establish that $a(z_1, z_2)$ is devoid of zeros in $|z_{1,2}| \leqq 1$, so that $a^{-1}$ is strongly outer. To see this it is sufficient to observe that the equation $\log(1 + x) = x$ has no complex root other than $x = 0$. This means that $a$ can only vanish at the zeros of $b$. But $b = 0$ yields $a = 1/2(\alpha^2 - 1) \neq 0$, so that $a$ has no zeros.

*Case 6.* $\alpha = 1$, $\alpha = 0$, *and* $\beta = 0$. The remaining situations are degenerate because $b$ is either a one-variable polynomial or a product of such polynomials, so that the well-known results of the one-variable theory apply [2]. For instance, $\alpha = 1$, $\beta \leqq 1$ yields $a = b^{-1} = (1 + z_1)^{-1}(1 + \beta z_2)^{-1}$, hence $g = 1$, $\mu(b) = 0$.

To sum up, it has been shown that the canonical weakly inner-strongly outer factorization of the polynomial (20) exists in the whole parameter space $(\alpha, \beta)$, with the exception of a small region defined by $|\alpha| > 1$, $|\beta| \neq 1$, $|\alpha| < |\beta|^2 \sqrt{1 + |\beta|^2}$ and $|\alpha\beta^3| < \sqrt{1 + |\beta|^2}$. In this region, the canonical factorization has not been established but its existence is firmly conjectured by the authors.

**Appendix.** Let us outline the argument showing that the function $a(z_1, z_2)$ defined by (23) is devoid of zeros in the closed unit bicylinder for $\alpha \geqq \beta^2 \sqrt{1 + \beta^2}$ (with $0 < \beta < 1 < \alpha$). Instead of $a$ we shall consider the function

$$a'(z_1, z_2) = [\alpha^2 - \beta^2 + \alpha(1 - \beta^2)z_1](\alpha + \beta^3 z_2)a(z_1, z_2), \tag{A.1}$$

which has the same zeros as $a(z_1, z_2)$ in $|z_{1,2}| \leqq 1$. Defining the one-variable functions

$$f_t(z_1) = \frac{\alpha^2 - \beta^2 + \alpha(1 - \beta^2)z_1}{\alpha^2 - \beta^{2t} + \alpha(1 - \beta^{2t})z_1},$$

(A.2)

$$g_t(z_2) = \frac{\alpha + \beta^3 z_2}{(\alpha + \beta^{2t-1}z_2)(\alpha + \beta^{2t+1}z_2)},$$

one can write (A.1) as follows:

$$a'(z_1, z_2) = \alpha^2(1 - \beta^2) \sum_{t=1}^{\infty} \beta^{2t-2}(1 - \beta^{2t}) f_t(z_1) g_t(z_2). \tag{A.3}$$

The proof then consists in establishing that each term of (A.3) has a positive real part in $|z_{1,2}| \leqq 1$, which immediately leads to the desired conclusion. As can be verified after elementary but rather tedious calculation, the inequality

$$|\operatorname{Im} f_t| \cdot |\operatorname{Im} g_t| < \operatorname{Re} f_t \cdot \operatorname{Re} g_t \tag{A.4}$$

is satisfied in $|z_{1,2}| \leqq 1$, for all $t \geqq 1$, provided $\alpha \geqq \beta^2 \sqrt{1 + \beta^2}$ holds. Now (A.4) yields $\operatorname{Re}(f_t g_t) > 0$, and hence $\operatorname{Re} a' > 0$.

## REFERENCES


[1] 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.

[2] ———, *Planar least-squares inverse polynomials, part* I. *algebraic properties,* IEEE Trans. Circuits and Systems, CAS-26 (1979), pp. 59–66.

[3] R. G. Douglas, *On the invertibility of a class of Toeplitz operators on the quarter-plane,* Indiana Univ. Math. J., 21 (1972), pp. 1031–1035.

[4] R. G. Douglas and R. Howe, *On the C*-algebra of Toeplitz operators on the quarter-plane,* Trans. Amer. Math. Soc., 158 (1971), pp. 203–217.

[5] M. P. Ekstrom and J. W. Woods, *Two-dimensional spectral factorization with applications in recursive digital filtering,* IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-24 (1976). pp. 115–128.

[6] Y. Genin and Y. Kamp, *Counterexample in the least-squares inverse stabilization of* 2 D *recursive filters,* Electronics Letters, 11 (1975), pp. 330–331.

[7] L. Ya. Geronimus, *Orthogonal Polynomials,* Consultants Bureau, New York, 1961.

[8] H. Helson, *Lectures on Invariant Subspaces,* Academic Press, New York, 1964.

[9] H. Helson and D. Lowdenslager, *Prediction theory and Fourier series in several variables,* Acta Mathematica, 99 (1958), pp. 165–202.

[10] K. Hoffman, *Banach Spaces of Analytic Functions,* Prentice-Hall, Englewood Cliffs NJ, 1962.

[11] L. Hörmander, *An Introduction to Complex Analysis in Several Variables,* North-Holland, Amsterdam, 1973.

[12] E. I. Jury, *An overview of Shanks' conjecture and comments on its validity,* in Proc. Tenth Asilomar Conf. Circuits, Systems and Computers, Asilomar, CA, 1976.

[13] V. A. Malyšev, *On the solution of discrete Wiener-Hopf equations in a quarter-plane,* Soviet Math. Dokl., 10(1969), pp. 1032–1036.

[14] W. Rudin, *Function Theory in Polydiscs,* Benjamin, New York, 1969.

[15] J. L. Shanks, S. Treitel and J. H. Justice, *Stability and synthesis of two-dimensional recursive filters,* IEEE Trans. Audio Electroacoust., AU-20 (1972), pp. 115–128.

[16] M. G. Strintzis, *On the spacially causal estimation of two-dimensional processes,* IEEE Proceedings, 65 (1977), pp. 979–980.

[17] B. Sz-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space,* North-Holland, Amsterdam, 1970.

[18] 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.

[19] A. Zygmund, *Trigonometric Series,* Cambridge University Press, Cambridge, 1959.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0012 $01.00/0

# CHARACTERIZATION OF DISTRIBUTIONS BY RANDOM SUMS*

LUDWIG BARINGHAUS†

**Abstract.** Let $N, X_1, X_2, \cdots$, be independent random variables with $X_1, X_2, \cdots$, being nonnegative and identically distributed. Let $N$ have a power series distribution. Considering the random sum $S = \sum_{i=1}^{N} X_i$, the present paper gives a characterization of the distributions of $N$ and $X_i$ by means of the property that, up to a scale parameter, $S$ has the same distribution as $X_i$. If the expectation of $X_i$ is finite, one obtains a characterization of the gamma distribution.

**1. Introduction.** It is known that a geometric sum of independent identically distributed (i.i.d.) exponential random variables is exponentially distributed. Arnold (1973) proved that if for every $p \in (0, 1)$, $p$ times a geometric $(p)$ sum of independent i.i.d. nonnegative random variables has the same distribution as the individual random variables, then the random variables are exponentially distributed. A more general characterization problem is as follows. Suppose that $X_1, X_2, \cdots$ is a sequence of i.i.d. nonnegative random variables. Let $N$ be a random variable independent of the $X_i$'s and having a power series distribution. Then we ask for the distributions of $X_1$ and $N$ for which up to a scale parameter the random sum $S = \sum_{i=1}^{N} X_i$ has the same distribution as the individual random variables. We give the complete solution of this problem by determining the Laplace transform and the generating function of $X_1$ and $N$, respectively. Assuming that $X_1$ has a finite expectation, we obtain that $X_1$ has a gamma distribution.

**2. A functional equation.** Let $\varphi(z) = E(\exp(-zX_1))$, $z \geqq 0$, be the Laplace transform of $X_1$ and let

$$\mathfrak{P} = \{P_\vartheta;\ P_\vartheta(\{x\}) = c(\vartheta)\vartheta^x h(x),\ x = 0, 1, \cdots,\ 0 < \vartheta < \vartheta_1\}, \qquad \vartheta_1 > 0, \tag{2.1}$$

be the power series family containing the distribution of $N$. Note that $h(x) \geqq 0$, and $c(\vartheta) = \left(\sum_{k=0}^{\infty} h(k)\vartheta^k\right)^{-1}$ for $0 < \vartheta < \vartheta_1$. If $G_\vartheta$ denotes the generating function of $P_\vartheta \in \mathfrak{P}$, we obtain $G_\vartheta(t) = G_{\vartheta_0}(\vartheta t/\vartheta_0)/G_{\vartheta_0}(\vartheta/\vartheta_0)$, $|t| \leqq 1$, for parameters $\vartheta$, $\vartheta_0 \in (0, \vartheta_1)$. It is easily seen that the radius of convergence of $G_\vartheta$ is greater or equal to $\vartheta_1/\vartheta$. Assuming that $N$ has the distribution $P_\vartheta \in \mathfrak{P}$, $\vartheta \in (0, \vartheta_1)$, the Laplace transform of the random sum $S = \sum_{i=1}^{N} X_i$ is given by $G_\vartheta(\varphi(z))$, $z \geqq 0$. Hence, for any $\vartheta \in (0, \vartheta_1)$, the distribution of $S$ differs from that of $X_1$ only up to a scale parameter iff there exists a real valued positive function $g$ on the interval $(0, \vartheta_1)$, such that the equality

$$G_\vartheta(\varphi(z)) = \varphi(g(\vartheta)z) \tag{2.2}$$

holds for all $z \geqq 0$ and all $\vartheta \in (0, \vartheta_1)$. Now for some fixed $\vartheta_0 \in (0, \vartheta_1)$ we can write (2.2) in the form

$$\frac{G_{\vartheta_0}(\vartheta\varphi(z)/\vartheta_0)}{G_{\vartheta_0}(\vartheta/\vartheta_0)} = \varphi(g(\vartheta)z), \qquad z \geqq 0, \quad \vartheta \in (0, \vartheta_1). \tag{2.3}$$

* Received by the editors April 9, 1979, and in final form March 4, 1980.

† Lehrgebiet Mathematische Stochastik, Universität Hannover, Welfengarten 1, Federal Republic of Germany.

Putting $G = G_{\vartheta_0}$, $t = \vartheta/\vartheta_0$, $h(t) = g(\vartheta_0 t)$, we get the functional equation

$$\frac{G(t\varphi(z))}{G(t)} = \varphi(h(t)z), \qquad z \geqq 0, \quad t \in (0, 1], \tag{2.4}$$

where $G$ is a probability generating function with a radius of convergence greater than 1, and where $h$ is a real valued positive function on $(0, 1]$. In the following we assume that the distribution of $N$ and $X_1$, respectively, is nondegenerate. Then it follows from (2.4) that $\lim_{z\to\infty} \varphi(z) = 0$, and so $G(0) = \lim_{t\to 0} G(t) = 0$. Putting $z = \exp(-y)$, $F(y) = \varphi(\exp(-y))$, $\gamma(t) = -\log h(t)$ for $-\infty < y < +\infty$, $t \in (0, 1]$, we obtain from (2.4),

$$\frac{G(tF(y))}{G(t)} = F(y + \gamma(t)), \qquad -\infty < y < +\infty, \quad t \in (0, 1]. \tag{2.5}$$

This functional equation occurs in a different connection in Baringhaus (1979). The general solution given there is as follows:

(i) $G(t) = \left(\dfrac{pt^k}{1 - (1 - p)t^k}\right)^{1/k}$, $p \in (0, 1)$, $k$ a positive integer;

(ii) $F(y) = (1 + \exp(-ay + b))^{-1/k}$, $a > 0$, $-\infty < b < +\infty$;

(iii) $\gamma(t) = \dfrac{1}{a} \log(1 - (1 - p)t^k)$.

Note that $G(t) = tG^*(t^k)$, where $G^*(t) = (p/(1 - (1 - p)t))^{1/k}$ denotes the generating function of a negative binomial distribution. From (ii) we obtain

(ii)′ $\varphi(z) = (1 + \alpha z^a)^{-1/k}$, $z \geqq 0$,

with $\alpha = \exp(b)$, $a > 0$. However, this is the Laplace transform of a probability distribution on $[0, \infty)$ iff $a \leqq 1$, because $\varphi(z)$ must be completely monotone (see Feller (1971)). We obtain that

$$E(X_1) = -\lim_{z \downarrow 0} \frac{d}{dz} \varphi(z)$$

is finite iff $a = 1$.

**3. The result.** Summarizing the results of § 2 we can state the following

THEOREM. *Let $N, X_1, X_2, \cdots$, be independent random variables with $X_1, X_2, \cdots$, nonnegative and identically distributed. Let the distribution of $N$ belong to the power series family $\mathfrak{P} = \{P_\vartheta; 0 < \vartheta < \vartheta_1\}$, $\vartheta_1 > 0$. The distributions $P^N$ of $N$ and $P^{X_1}$ of $X_1$ are assumed to be nondegenerate. Then for any $\vartheta \in (0, \vartheta_1)$ with $P^N = P_\vartheta$ there exists a positive number $g(\vartheta)$ such that the random variables $\sum_{i=1}^N X_i$ and $g(\vartheta)X_1$ are identically distributed iff*

(a) $\varphi(z) = E(\exp(-zX_1)) = (1 + \alpha z^a)^{-1/k}$, $z \geqq 0$,

(b) $$G_\vartheta(t) = E(t^N) = \left(\frac{\left(1 - (1 - p)\left(\dfrac{\vartheta}{\vartheta_0}\right)^k\right) t^k}{1 - (1 - p)\left(\dfrac{\vartheta}{\vartheta_0}\right)^k t^k}\right)^{1/k}, \qquad |t| \leqq 1,$$

*where $\alpha > 0$, $a \in (0, 1]$, $p \in (0, 1)$, $\vartheta_0 \in (0, \vartheta_1)$, $\vartheta_0 \geqq \vartheta_1(1 - p)^{1/k}$, and $k$ is a positive integer. Moreover, $E(X_1)$ is finite iff $a = 1$.*

Hence, for random variables $X_i$ with a finite expectation $E(X_i)$, the theorem given above yields a characterization of the gamma distribution. As a consequence of our theorem we have the following.

COROLLARY. *Let $N$, $X_i$ and $\mathfrak{P}$ be as in the theorem. Suppose that $X_i$ has a finite expectation and that the random variable $N$ takes on the value 2 with positive probability. Then for any $\vartheta \in (0, \vartheta_1)$ with $P^N = P_\vartheta$, there exists a positive number $g(\vartheta)$ such that the random variables $\sum_{i=1}^{N} X_i$ and $g(\vartheta)X_1$ are identically distributed iff $X_1$ is exponentially distributed and $N$ has a geometric distribution. In this case $\mathfrak{P}$ is a family of geometric distributions.*

For the proof of the corollary it suffices to note that the random variable $N$ takes on the value 2 with positive probability iff the integer $k$ specified in the theorem is equal to 1.

*Remark.* Distributions with a Laplace transform of the form $\varphi(z) = (1 + \alpha z^a)^{-1}$, $\alpha > 0$, $a \in (0, 1]$, occur as limit distributions of normalized sums of a random number of positive independent random variables (see Gnedenko (1972)).

## REFERENCES

B. C. ARNOLD, (1973). *Some characterizations of the exponential distribution by geometric compounding,* SIAM J. Appl. Math. 24, pp. 242–244.

L. BARINGHAUS, (1979). *Eine simultane Charakterisierung der geometrischen Verteilung und der logistischen Verteilung.* To appear in METRIKA.

W. FELLER, (1971). *An Introduction to Probability Theory and its Applications,* Vol. II, John Wiley, New York.

B. V. GNEDENKO, (1972). *Limit theorems for sums of a random number of positive independent random variables,* in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 2, pp. 537–549.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0013 $01.00/0

# SINGLE COMMODITY REPRESENTATION OF MULTICOMMODITY NETWORKS*

Y. SOUN† AND K. TRUEMPER†

**Abstract.** An efficient algorithm is described that transforms a directed multicommodity flow network to an equivalent single commodity network, provided such a transformation is possible.

**Introduction.** This paper establishes necessary and sufficient conditions under which the linear constraints of a multicommodity flow problem can be transformed to those of a single commodity network. An efficient algorithm tests whether or not the conditions are satisfied. If the answer is affirmative, the equivalent single commodity network is then efficiently constructed. Transformability requires that the constraint coefficient matrix of the multicommodity flow problem be unimodular. It is shown that the latter property is not sufficient to induce transformability; in fact, an elementary scheme constructs all nontransformable problems with unimodular coefficient matrix. In matroid terminology (see Welsh [13] for definitions) this construction establishes an infinite class of connected regular matroids that are neither graphic nor cographic.

The only previous work addressing the transformation problem for general directed multicommodity networks is due to Evans [4]. He shows that the transformation is possible if the underlying graph observes a recursively defined condition. In a related survey [5] Evans shows that his condition is not necessary for transformability. This is also demonstrated by the example problem presented below.

Tutte [11], [12], Iri [6], Bixby and Cunningham [2], as well as several others (see [2], [6] for surveys) have investigated the problem of transforming an arbitrary matrix to the node-arc incidence matrix of a directed graph. The four references present efficient algorithms that establish whether or not such a transformation is possible. The procedure given here for settling the transformation question for the multicommodity problem as well as the method of construction of the equivalent single commodity problem are substantially faster than the general methods of [2], [6], [11], [12]. For example, the transformation question is resolved in $O(n)$ additions/subtractions, where $n$ is the number of arcs of the underlying graph.

The remainder of this section introduces definitions and preliminary results. We will consider multicommodity min-cost flow problems of the form

$$\text{(1)} \qquad \text{max:} \quad \sum_{i=1}^{k} (a^i)^t w^i,$$

$$\text{(2)} \qquad \text{s.t.} \quad E_G w^i = d_G^i, \qquad i = 1, 2, \cdots, k,$$

$$\text{(3)} \qquad \sum_{i=1}^{k} w^i + s = c,$$

$$\text{(4)} \qquad w^i \geqq 0 \quad \forall i; \quad s \geqq 0,$$

where $E_G$ is the node-arc incidence matrix of the underlying directed graph $G$ with finite node and arc sets. All vectors are column vectors of appropriate dimension. In

* Received by the editors March 3, 1978, and in final form January 10, 1980. This paper is a revised version of part of [7].

† University of Texas at Dallas, P.O. Box 688, Richardson, Texas 75080.

particular, $w^i$ represents flow of commodity $i$, so (1)–(4) involves $k$ commodities. Equation (2) enforces flow conservation for each commodity, while (3) represents the capacity restriction for each arc of $G$. We assume $\sum_j (d_G^i)_j = 0$, $\forall i$, since else (2) does not have a solution.

If we collect the $a^i$ ($w^i$) into vector $a$ ($w$), and the $d_G^i$ and $c$ into vector $b$, then (1)–(4) can be written as

$$\text{(5)} \qquad \max: \quad a^t w,$$

$$\text{(6)} \qquad \text{s.t.} \quad M \begin{bmatrix} w \\ \hline s \end{bmatrix} = b,$$

$$\text{(7)} \qquad w, s \geqq 0,$$

where

$$\text{(8)} \qquad M = \left[\begin{array}{ccccc|c} E_G & & & & & \\ & E_G & & & & \\ & & \ddots & & & 0 \\ & & & & E_G & \\ \hline I & I & \cdots & & I & I \end{array}\right]$$

contains $k$ copies of $E_G$.

If columns of a matrix $B$ are independent, and if they span all columns of a matrix $A$, then $B^{-1}\{A\}$ will denote the solution $Z$ of $BZ = A$. Let $B$ now be a basis of $M$, and define $\hat{M}$ to be $B^{-1}\{\tilde{M}\}$, where $\tilde{M}$ consists of the nonbasic columns of $M$. Assume (5)–(7) is brought into basic (feasible or infeasible) form

$$\text{(9)} \qquad \max: \quad \hat{a}^t x,$$

$$\text{(10)} \qquad \text{s.t.} \quad \hat{M}x \leqq \hat{b},$$

$$\text{(11)} \qquad x \geqq 0,$$

with related dual problem

$$\text{(12)} \qquad \min: \quad \hat{b}^t y,$$

$$\text{(13)} \qquad \text{s.t.} \quad (-\hat{M}^t)y \leqq (-\hat{a}),$$

$$\text{(14)} \qquad y \geqq 0.$$

We will say that (1)–(4) is *transformable to a single commodity network problem* (or simply: (1)–(4) is *transformable*), if $[I \mid \hat{M}]$ or $[I \mid -\hat{M}^t]$ is equal to $T^{-1}\{F\}$, where $F$ is the node-arc incidence matrix of a directed graph, and $T$ is a basis of $F$. Transformability implies that (9)–(11) or its dual (12)–(14) may be solved by very efficient network flow methods (see [3] for survey). One can easily prove that transformability holds for every basis of $M$ if it holds for a particular basis, so the definition does not depend on the choice of $B$. Further, the definition agrees with conclusions by Iri [6] regarding transformation of a general linear program to a (single commodity) network problem. Iri's test for existence of such a transformation includes a scaling of variables if the updated coefficient matrix (corresponding to $\hat{M}$ here) is not a $\{0, \pm 1\}$ matrix. Such scaling is never needed for (1)–(4), since $M$ always contains a basis $B$ such that $B^{-1}\{\tilde{M}\}$ is a $\{0, \pm 1\}$ matrix (see [9]).

A matrix is *totally unimodular* if every square submatrix has determinant 0 or $\pm 1$. It is *unimodular* if every basis $B$ satisfies g.c.d.$_i$ $\{\det B^i\} = 1$, where the $B^i$ are the maximal square submatrices of $B$.

LEMMA 1. *$M$ of* (8) *is unimodular if* (1)–(4) *is transformable.*

*Proof.* In [9] it is shown that $M$ always has a unimodular basis $B$. Under transformability $\hat{M}$ or $(-\hat{M}^t)$, defined via $B^{-1}\{\tilde{M}\}$, is a submatrix of $T^{-1}\{F\}$ for a node-arc incidence matrix $F$ with basis $T$. Matrix $T^{-1}\{F\}$ is totally unimodular, so $\hat{M}$ is totally unimodular. Then $M = B \cdot [I \mid \hat{M}]$ is unimodular since $B$ is unimodular. Q.E.D.

All graphs considered here are directed and may have parallel arcs (i.e., arcs with the same endpoints). *Cycle* refers to a simple cycle whose arcs have arbitrary directions. A graph is defined to be *2-connected* if it is connected, and if every pair of arcs is contained in at least one cycle. A 2-connected graph is called a *suspension graph* if it contains a node such that removal of that node and its incident arcs reduces the graph to a tree. We will say that nodes $i$ and $j$ without connecting arc are *identified* when the two nodes are to be considered as just one node, say $l$. Thus any arc with $i$ or $j$ as endpoint becomes incident at $l$.

In the description of the algorithm and the proof of validity we will always assume that graph $G$ underlying (1)–(4) is 2-connected, since this assumption greatly simplifies the presentation. In the final section results are then extended to the general case. By the above definition $G$ consisting of just one arc and its two endpoints is 2-connected. The corresponding problem (1)–(4) is trivially transformable, so we will always suppose that $G$ has two or more arcs.

Following Whitney [14] we define two graphs to be 2-*isomorphic* if there exists a one-to-one mapping $\phi$ between the arc sets such that each cycle of one graph corresponds to a cycle in the other graph. This correspondence includes the relative orientation of arcs of each cycle. The following two lemmas are an immediate consequence of this definition.

LEMMA 2. *Let $G$ and $H$ be 2-isomorphic, and assume arc $e$ of $G$ corresponds to arc $e'$ of $H$ under mapping $\phi$. Then the two graphs derived from $G$ and $H$ by* (a) *or* (b) *below, are also 2-isomorphic.*

a) *Replace arc $e$ ($e'$) by two parallel arcs $f$ and $g$ ($f'$ and $g'$) in $G$ ($H$). Relative orientation of $f$ and $g$ to $e$ is the same as that of $f'$ and $g'$ to $e'$.*

b) *Replace $e$ of $G$ ($e'$ of $H$) by a simple path with two arcs $f$ and $g$ ($f'$ and $g'$). Let $C$ ($C'$) be the cycle defined by arcs $e$, $f$, and $g$ ($e'$, $f'$ and $g'$) and their endpoints. (Note that $C$ ($C'$) is not a subgraph of $G$ ($H$) since $e$ ($e'$) is no longer present in $G$ ($H$)). Relative orientation of $f$ and $g$ to $e$ in $C$ must be the same as for $f'$ and $g'$ to $e'$ in $C'$. (The simple path in $G$ ($H$) introduces an additional node of degree two into $G$ ($H$), having $f$ and $g$ ($f'$ and $g'$) incident).*

LEMMA 3. *Suppose $E_G$ and $E_H$ are node-arc incidence matrices of 2-isomorphic graphs $G$ and $H$, where columns of $E_H$ are ordered in the same way as columns of $E_G$ under mapping $\phi$. Let $T_G$ be a basis of $E_G$, and $T_H$ be the corresponding basis of $E_H$. Then $T_G^{-1}\{E_G\} = T_H^{-1}\{E_H\}$. For given $d_G$ suppose $E_G \cdot z = d_G$ has a solution $z$. Then $E_G \cdot z = d_G$ and $E_H \cdot z = d_H$ have same solution set iff $d_H = T_H \cdot T_G^{-1}\{d_G\}$.*

The next section presents the algorithm that transforms (1)–(4) to a single commodity network problem iff such a transformation is possible. The final section establishes validity of the algorithm.

**Transformation algorithm.** The algorithm consists of two parts. The analysis part (step 1) checks whether or not (1)–(4) can be transformed. This step is taken from [8], where it is used to test a $\bar{k}$-commodity version of $M$ for unimodularity, for

$\bar{k} \geqq 3$. At the time work on [8] was completed, it was not recognized that this step may also be employed to test $M$ for transformability (regardless of the number of commodities).

Suppose now that step 1 does indicate transformability. The synthesis part first constructs a suspension graph $H$ that is 2-isomorphic to $G$ (step 2), then derives the desired single commodity network problem via graph $H$ (step 3). In [4] Evans proves that (1)–(4) is transformable if graph $G$ observes a certain recursively defined condition. It is easily seen that 2-connected $G$ satisfies Evans' condition iff $G$ is a suspension graph. In a recent survey paper [5], Evans also mentions this relationship. It is also easily demonstrated that the algorithm given below can be simplified to Evans' transformation rules [4] if $G$ is a suspension graph.

STEP 1 (Analysis). Initially all arcs of $G$, the underlying graph of (1)–(4), are marked nonspecial. Repeatedly apply the following rules (in any order), starting with graph $G$.

a) If two arcs are parallel, replace them by a single special arc with same end points and arbitrary direction.

b) If the graph has a node of degree two, and if at most one of the two arcs incident at that node is special, then remove that node and merge the two incident arcs into a single arc of arbitrary direction. The resulting arc is marked special if one of the two arcs just merged was special.

Once $G$ has been reduced to a cycle with two arcs, go to step 2. If $G$ has been reduced to a graph that is not a cycle of two arcs, and if no further reduction via 1a) or 1b) is possible, then (1)–(4) cannot be transformed.

STEP 2 (Synthesis: 2-isomorphic graph $H$). Let $G = G_0, G_1, G_2, \cdots, G_R$ (= cycle with two arcs) be the graphs obtained in step 1 by repeated application of rules of step 1a), 1b). Construct a sequence of suspension graphs $H = H_0, H_1, H_2, \cdots, H_R$ as follows:

a) Let $H_R$ be the same graph as $G_R$, but mark both arcs as special, and designate one of the nodes of $H_R$ as $v$. Each arc of $H_R$ corresponds to an arc of $G_R$ by the isomorphism between the two graphs.

b) For $n = R, R-1, \cdots, 1$ derive $H_{n-1}$ from $H_n$, $G_n$ and $G_{n-1}$ as follows:

ba) If step 1a) reduced $G_{n-1}$ to $G_n$, i.e., two parallel arcs $f$ and $g$ of $G_{n-1}$ were replaced by special arc $e$ of $G_n$, then replace the corresponding special arc $e'$ of $H_n$ by two arcs $f'$ and $g'$ with the same end points as $e'$. Direction of $f'$ ($g'$) is selected such that relative orientation of $f'$ ($g'$) to $e'$ agrees with that of $f$ ($g$) to $e$. Mark both $f'$ and $g'$ as special. For subsequent steps $f'$ ($g'$) corresponds to $f$ ($g$).

bb) If step 1b) reduced $G_{n-1}$ to $G_n$, i.e., two arcs $f$ and $g$ incident at a node of degree two were merged into arc $e$, then replace the corresponding arc $e'$ of $H_n$ by a simple path with two arcs $f'$ and $g'$. (This introduces an additional node of degree two as well). As in Lemma 2b let $C$ ($C'$) be the cycle defined by arcs $e, f$, and $g$ ($e', f'$, and $g'$) and their endpoints. Direction of $f'$ ($g'$) is selected such that relative orientation of $f'$ ($g'$) to $e'$ in $C'$ agrees with that of $f$ ($g$) to $e$ in $C$. For subsequent steps, $f'(g')$ corresponds to $f$ ($g$).

Special markers for $f$ and $g$ (if any) are determined as follows:

bba) If $e'$ is not special, then $f'$ and $g'$ are not special.

bbb) If both $e$ and $e'$ are special, then one of $f$ and $g$ must be special. Mark the corresponding $f'$ or $g'$ as special. Without loss of generality suppose $f'$ is now marked special. If $f'$ is not incident to node $v$, ex-

change it with $g'$ so that $f'$ becomes incident to $v$. This exchange is done in such a manner that the relative orientation of $f'$ ($g'$) to $e'$ (as defined above via $C'$) remains unchanged.

bbc) If $e'$ is special and $e$ is not, one of the arcs $f'$, $g'$ must be incident at node $v$. Mark that arc as special.

STEP 3 (Synthesis: Single Commodity Network).

a) Let $N$ and $A$ be node and arc set of $H$, and $A_v$ be the set of arcs incident at node $v$. Order columns of $E_H$, the node-arc incidence matrix of $H$, such that columns of arcs of $A_v$ are listed first. If necessary, rearrange (1)–(4) so that columns of $E_G$ and elements of vectors $a$, $c$, $w^i$ are ordered in the same way as columns of $E_H$ under the correspondence of arcs determined in step 2. Below, the notation "$c_{jl}$" will refer to element $c_{xy}$ of $c$ in (3), where $xy$ is the arc of $G$ corresponding to arc $jl$ of $H$. Analogous relationships hold for vectors $a$, $w^i$, and $s$ of (1)–(4).

b) Select an arbitrary basis $T_H$ from $E_H$, and let $T_G$ be the corresponding basis of $E_G$. Compute $d_H^i = T_H \cdot T_G^{-1}\{d_G^i\}$, $i = 1, 2, \cdots, k$. Delete from $E_H$ ($d_H^i$, $i = 1, 2, \cdots, k$) the row (element) representing node $v$, getting $\bar{E}_H$ ($\bar{d}_H^i$, $i = 1, 2, \cdots, k$). Define $\bar{d}_H^0$ by $(\bar{d}_H^0)_j = \sum_{l \in N}[c_{jl} - c_{lj}] - \sum_{i=1}^{k} (d_H^i)_j$, $j \in N - \{v\}$.

c) Let $|A_v| \times |A|$ matrix $L$ be equal to $[J \mid 0]$, where the columns correspond to arcs of $A$ as in $E_H$. $J$ is drived from the identity by replacing some $+1$ entries by $-1$. Specifically, $J_{l,l}$ is $+1$ ($-1$) if the arc of $A_v$ corresponding to column $l$ is $vj$ ($jv$), some $j \in N$. Finally let $h$ be a vector of length $|A_v|$. Element $h_l$ represents the same arc of $A_v$ as column $l$ of $L$, and it has value $c_{vj}$ ($-c_{jv}$) if that arc is $vj$ ($jv$).

d) (1)–(4) is then equivalent to the single commodity network problem

$$\text{max:} \quad \sum_{i=1}^{k} (a^i)^t w^i, \tag{15}$$

$$\text{s.t.} \quad \bar{E}_H w^i = \bar{d}_H^i, \qquad i = 1, 2, \cdots, k, \tag{16}$$

$$\bar{E}_H s = \bar{d}_H^0, \tag{17}$$

$$\sum_{i=1}^{k} L w^i + L s = h, \tag{18}$$

$$w^i \geqq 0 \quad \forall i; \quad s \geqq 0. \tag{19}$$

We briefly examine the computational complexity of the algorithm and discuss an example, deferring the proof of validity to the next section. We count as one computational step addition or subtraction of two real numbers that are part of the input, or that are generated by additions or subtractions from the input by the algorithm. Assume (1)–(4) with $k$ commodities is specified by the node set (with $m$ nodes) and the arc set (with $n$ arcs) of graph $G$, and by vectors $a$, $c$, and $d_G^i$, $i = 1, 2, \cdots, k$. Steps 1 and 2, suitably implemented, involve $O(n)$ steps. Computation of $\bar{d}_H^i$, $i = 1, \cdots, k$ for (15)–(19) in step 3 requires $O(k \cdot m)$ steps since $T_H \cdot T_G^{-1}\{d_G^i\}$ is of order $O(m)$. Finally, $\bar{d}_H^0$ is found in $O(n)$ steps (assuming $\sum_{i=1}^{k} d_H^i$ is computed along with $T_H \cdot T_G^{-1}\{d_G^i\}$), and $L$ and $h$ are determined in $O(m)$ steps. Hence overall complexity is $O(\max\{k \cdot m, n\})$.

The following observations result in a minor improvement of the algorithm. It is easily seen that a suspension graph $H$ with $m$ nodes and no parallel arcs has at most $2m - 3$ arcs. Now $G$ and $H$ of steps 1 and 2 have the same number of arcs and nodes, and one of these graphs has parallel arcs only if the other graph has such arcs as well.

Suppose for the moment that $G$ has no parallel arcs. By the preceding arguments $n \leqq 2m - 3$ is a necessary condition for transformability, and overall complexity of the algorithm becomes $O(k \cdot m)$ once this test on $n$ is added to step 1. A similar refinement is possible if $G$ has parallel arcs. In that case one first applies step 1a until all parallel arcs have been eliminated. Suppose that the graph at that time has $n_1$ arcs, of which $n_2$ arcs are marked special, and that (1)–(4) is transformable. If we disregard parallel arcs, then $H$ of step 2 will have $n_1$ arcs, of which at least $n_2$ arcs will be special and incident at one node. Thus inequalities $n_1 \leqq 2m - 3$ and $n_2 \leqq m - 1$ must be satisfied, and we should verify this before proceeding with step 1. However, this change does not improve the order of complexity, which remains as $O(\max\{k \cdot m, n\})$. It seems instructive to compare the latter formula with the complexity of general methods (e.g., the ones cited in the Introduction). Every method published to date requires $\hat{M}$ of (10) as input. For arbitrary $m$, transformable problems exist where $\hat{M}$ averages about $m$ nonzero entries per column, so any general method that uses $\hat{M}$ has complexity of at least $O(k \cdot m \cdot (n - m))$, which is larger than $O(\max\{k \cdot m, n\})$, the order of the algorithm presented here.

We conclude this section with presentation of an illustrative example involving two commodities. Fig. 1 displays graph $G$ and all relevant data of (1)–(4). Note that $G$ is not a suspension graph, so this problem cannot be transformed by the method of [4]. The pair $(d_1, d_2)$ next to a node specifies that $d_i$ units of external inflow of commodity $i$ must occur at that node for $i = 1, 2$. The triple $[\omega_0, \omega_1, \omega_2]$ alongside an arc establishes $\omega_0$ as capacity of that arc and $\omega_i$ as cost per unit of flow of commodity $i$, for $i = 1, 2$. The letters $a, b, \cdots$ alongside arcs are labels that allow easy identification of the reductions and expansions in steps 1 and 2. The latter steps produce graphs $G_0 = G, G_1, \cdots, G_7$, and $H_7, H_6, \cdots, H_0 = H$. Fig. 2 displays a representative subset of this list of graphs. For example, arcs $e$ and $d$ of $G_0$ are replaced by arc $m$ of $G_3$. For clarity we have used the same arc labels in $H_j$ as in $G_j$, for all $j$. Special arcs are designated by the symbol $\otimes$.

We now turn to step 3. Some convention is needed to assign external flows of $H$ to entries of $d_H^1$ and $d_H^2$. It seems simplest to assign external flow at node $j$ to entry $j$, and with this convention we have $d_H^1 = T_H \cdot T_G^{-1}\{d_G^1\} = (-2, 2, -3, 6, -3)^t$ and $d_H^2 = (-13, -2, 9, -5, 11)^t$, where we used arc set $\{a, e, g, h\}$ in $G$ and $H$ to define $T_G$ and $T_H$. Since $v = 1$, deletion of the first entry in $d_H^1$ and $d_H^2$ gives $\bar{d}_H^1$ and $\bar{d}_H^2$, respectively. Then $\bar{d}_H^0 = (1, -2, 1, 1)^t$ by the formula of step 3b). The equivalent

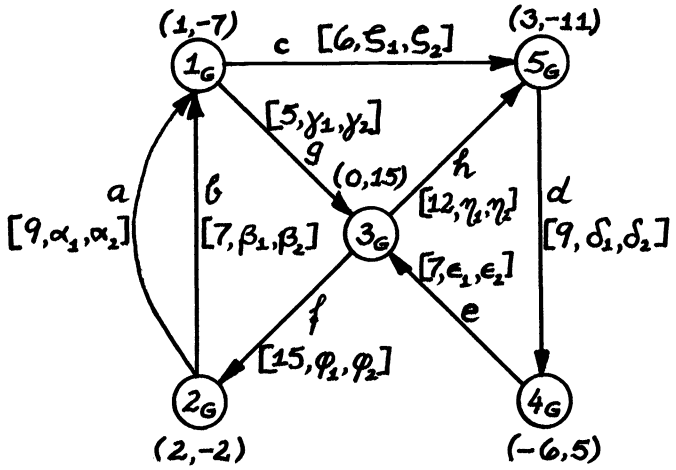



FIG. 1. *Graph G and data of example problem.*

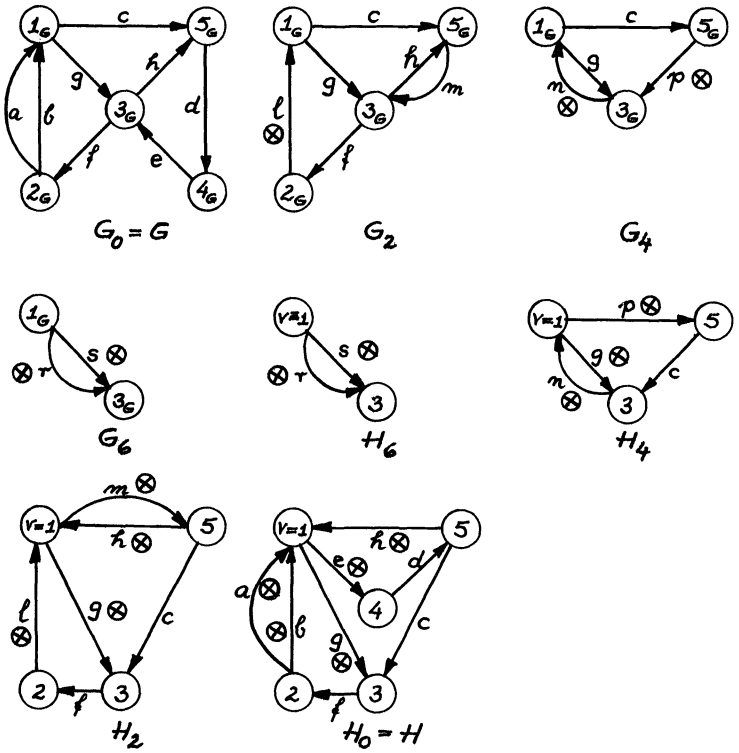



FIG. 2. *Graphs of steps 1 and 2.*

single commodity problem (15)–(19) is now easily found since $\bar{E}_H$, $L$ and $h$ are uniquely determined by $H$. The corresponding network is shown in Fig. 3. Here each number next to a node specifies the external inflow into that node, and a Greek letter on an arc denotes the cost per unit of flow on that arc. Equation (16) generates nodes 2.$i$, 3.$i$, 4.$i$, and 5.$i$, for $i = 1, 2$, while (17) results in nodes 2.0, 3.0, 4.0, and 5.0. Nodes $a$, $b$, $e$, $g$, and $h$ arise from (18). Flows in the network of Fig. 3 are easily related to flows in $G$ with the aid of the cost coefficients. For example,

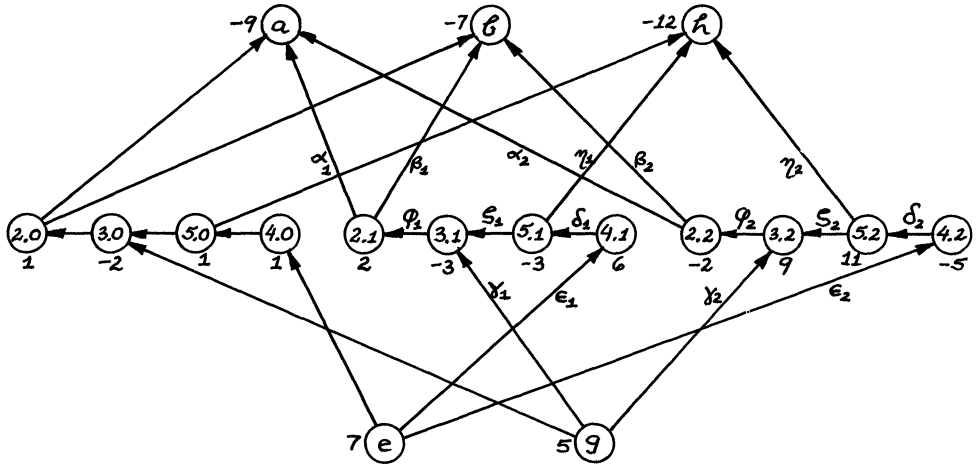



FIG. 3. *Equivalent Single Commodity Network Problem.*

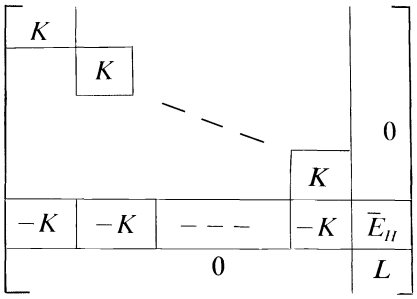
the arc with endpoints 5.1 and $h$ in Fig. 3 has cost coefficient $\eta_1$; hence its flow corresponds to the flow of commodity 1 on arc $h$ with endpoints $3_G$ and $5_G$ of $G$ in Fig. 1.

**Validity of the algorithm.** First it is proved that the algorithm does transform (1)–(4) to a single commodity network, provided underlying graph $G$ is reduced to a cycle with two arcs in step 1. For ease of exposition we define the following arrays.

(20) $\bar{E}_G(\bar{E}_H)$: derived from $E_G(E_H)$ by deletion of an arbitrary row (of row $v$).
$\bar{M}$, $\bar{b}$: derived from $M$ of (8) by replacing each copy of $E_G$ by $\bar{E}_G$. Deletion of related elements of $b$ produces $\bar{b}$.

Problem (1)–(4) can obviously be transformed to a single commodity network problem if one finds a matrix $D$ with independent columns such that $D\bar{M}$ is the node-arc incidence matrix of a directed network. Vector $D\bar{b}$ specifies external flows for that network. Let $E_G$ and $E_H$ be the matrices of step 3; i.e., columns of $E_G$ are ordered in the same way as those of $E_H$ (under the correspondence of arcs of $G$ and $H$ as determined in step 2), and columns of arcs in $A_v$, the set of arcs incident to $v$ in $H$, are listed first in $E_H$. Define $\bar{T}_G$ to be a basis of $\bar{E}_G$ and $\bar{T}_H$ to be the corresponding basis of $\bar{E}_H$, where $\bar{E}_G$ and $\bar{E}_H$ are defined by (20). Using $K = \bar{T}_H \cdot \bar{T}_G^{-1}$ and $L$ of step 3, let

(21) $D =$

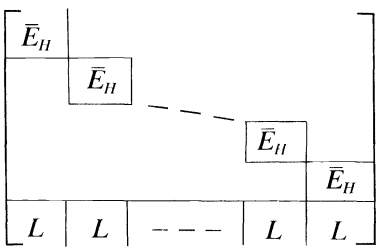



where $D$ contains $k$ copies of $K$. With the aid of the following lemma we will show that $D$ has independent columns, and that product $D \cdot \bar{M}$ is defined and leads to the desired node-arc incidence matrix. As in step 3, $N$ denotes the node set of $H$.

LEMMA 4. *$H$ of step 2 is 2-isomorphic to $G$ of* (1)–(4). *Further, every special arc of $H$ is incident at node $v$ and vice versa, while the nonspecial arcs form a tree that spans nodes of $N - \{v\}$.*

The proof of Lemma 4 is easily achieved by induction, using the sequence $H_R$, $H_{R-1}, \cdots, H_0$ of step 2 and Lemma 2. So let $H$ have $m$ nodes, and suppose $r = |A_v|$ arcs are incident at node $v$ of $H$. By Lemma 4, $\bar{E}_H$ is $(m-1) \times (r+m-2)$, and the last $m-2$ columns, which correspond to nonspecial arcs, are linearly independent. Then $D$ of (21) is $(r + (k+1) \cdot (m-1)) \times (r + (k+1) \cdot (m-1) - 1)$ and has linearly independent columns. Moreover, product $D \cdot \bar{M}$ is defined, since $\bar{M}$ is $(r + (k+1) \cdot (m-1) - 1) \times ((k+1) \cdot (r+m-2))$, and

(22) $D\bar{M} =$



The definition of $L$ and $\bar{E}_H$ implies that $D\bar{M}$ is the node-arc incidence matrix of a single commodity network. $D\bar{M}$ is also the constraint coefficient matrix of (16)–(18). The right-hand side of the latter equations is easily seen to be equal to $D\bar{b}$. We conclude that the algorithm achieves the desired transformation of (1)–(4) as claimed.

Now suppose that (1)–(4) is transformable, but that step 1 of the algorithm does not reduce $G$ to a cycle with two arcs. By Lemma 1, $M$ of (8) must be unimodular. In [8] it is demonstrated that under these conditions (1)–(4) involves $k = 2$ commodities, and that $G$ contains a subgraph $\bar{G}$ that is homeomorphic to (i.e., isomorphic to, within nodes of degree two) a graph constructed as follows.

One starts with a cycle $C$ having between three and six arcs. Then one parallel arc is added to each of three distinct arcs of $C$, such that the resulting graph has no node of degree two. Thus $\bar{G}$ is either the graph of Fig. 4 (where each dashed line represents a simple path with one or more arcs), or it is obtained from that graph by removal of one or more paths of $\{e, f, g\}$ and identification of the endpoints of any path so removed.

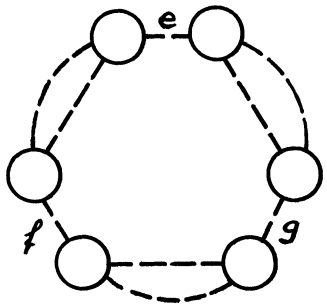



FIG. 4. *Subgraph $\bar{G}$ of $G$.*

The following easily established lemma reduces the potentially large number of graphs that must be examined to just one graph.

LEMMA 5. *Let* (1)–(4) *with underlying graph $G$ be transformable. Then* (1)–(4) *with graph $G'$ is also transformable, where $G'$ is obtained from $G$ by one of the following operations.*

(i) *Change direction of an arc.*

(ii) *Delete an arc.*

(iii) *Delete arc $ij$ and all arcs parallel to it, then identify nodes $i$ and $j$.*

By repeated applications of steps (i)–(iii) of Lemma 5, we conclude that transformability of (1)–(4) with graph $G$ and subgraph $\bar{G}$ of Fig. 4, implies transformability of (1)–(4) with graph $\hat{G}$ of Fig. 5. Let $B$ be the following basis of $M$ *of* (8) arising from $k = 2$ commodities and graph $\hat{G}$: For commodities one and two select columns of

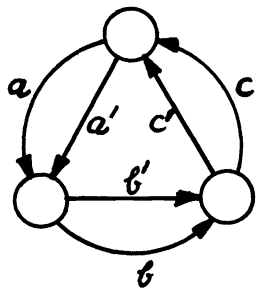



FIG. 5. *Graph $\hat{G}$.*

arcs $b'$ and $c'$, then use all columns of slack vector $s$. Let $\tilde{M}$ consist of the nonbasic columns of $M$. With these definitions $\hat{M} = B^{-1}\{\tilde{M}\}$ is

(23)

$$\hat{M} = \begin{array}{c} \begin{array}{cccccccc} a & a' & b & c & a & a' & b & c \end{array} \\ \left[\begin{array}{cccc|cccc} -1 & -1 & 1 & & & & & \\ -1 & -1 & & 1 & & & & \\ \hline & & & & -1 & -1 & 1 & \\ & & & & -1 & -1 & & 1 \\ \hline 1 & 1 & -1 & & 1 & 1 & -1 & \\ 1 & 1 & & -1 & 1 & 1 & & -1 \\ 1 & & & & 1 & & & \\ & 1 & & & & 1 & & \\ & & 1 & & & & 1 & \\ & & & 1 & & & & 1 \end{array}\right] \\ \underbrace{\qquad\qquad\qquad}_{\text{commodity 1}} \quad \underbrace{\qquad\qquad\qquad}_{\text{commodity 2}} \end{array}$$

Let $D^1$ ($D^2$) be the submatrix of $\hat{M}$ of (23) defined by columns 1–5 (1, 3, 5, 7), and rows 1, 2, 5–7 (1, 3, 5, 7, 9). That is,

$$D^1 = \begin{bmatrix} -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 \\ 1 & 1 & -1 & 0 & 1 \\ 1 & 1 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}; \quad D^2 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Let $K_{3,3}$ denote any complete bipartite graph with three source nodes and three sink nodes, regardless of the directions of arcs. Then $[I \mid (D^1)^t] = (B^1)^{-1} \cdot E^1$, where $E^1$ is the node/arc incidence matrix of a graph consisting of $K_{3,3}$ and one additional arc, and where $B^1$ is a basis of $E^1$. Similarly $[I \mid D^2] = (B^2)^{-1} \cdot E^2$, where this time $E^2$ corresponds to just $K_{3,3}$. By Tutte's characterization of graphic matroids [10], the conclusions about $D^1$ and $D^2$ imply that (1)–(4) with graph $\hat{G}$ is not transformable. By Lemma 5 this holds as well for (1)–(4) with graph $G$, which contradicts the initial assumption. It follows that reduction of $G$ to a cycle with two arcs in step 1 of the algorithm is achieved iff (1)–(4) is transformable.

We briefly relate the above results to the theory of regular matroids (see Welsh [13] for definitions). In [8] it is proved that $M$ with $k = 2$ commodities may be unimodular even if step 1 does not reduce underlying graph $G$ to a cycle with two arcs. By the results of [8] such matrices form an infinite class, and any member of that class can be efficiently constructed. Matrix $\hat{M}$ of (10) corresponding to such $M$ is totally unimodular, so by the above discussion the linear independence matroid defined from $[I \mid \hat{M}]$ is connected and regular, but not graphic or cographic. To our knowledge (1)–(4) so far is the only known combinatorial problem of practical significance that gives rise to an infinite class of such matroids. However, one should not attach too much importance to this claim. It seems likely that a number of practical combinatorial problems have the same property—they just have not been discovered as yet.

The following theorem combines the above conclusions with results of [8].

THEOREM 1. *The following statements are equivalent for problem* (1)–(4) *with underlying graph G.*

(i) (1)–(4) *is transformable.*
(ii) *The algorithm transforms* (1)–(4) *to* (15)–(19).
(iii) *G is 2-isomorphic to a suspension graph.*
(iv) *Every $\bar{k}$-commodity version of M is unimodular, for $\bar{k} \geqq 2$.*

*Proof.* Equivalence of (i) and (ii), as well as (i) $\Rightarrow$ (iii), follow from the above discussion. To show (iii) $\Rightarrow$ (i), let $G$ be 2-isomorphic to suspension graph $H$. Matrices $D_G$ and $D_H$ are easily determined via Lemma 3 such that $D_H \cdot D_G^{-1}\{M\} = M'$, where each copy of $E_G$ of $M$ has become a copy of $E_H$ in $M'$. A trivial inductive proof establishes that step 1 of the algorithm must reduce $H$ to a cycle with two arcs, so step 3 finds the equivalent single commodity network problem. Hence (1)–(4) has been shown to be transformable. Finally, equivalence of (ii) and (iv) is a consequence of results of [8]. Q.E.D.

Finally we address the general case of (1)–(4) when $G$ is connected but not 2-connected. In that case $G$ can be decomposed into its 2-connected components, say $G^1, G^2, \cdots, G^r$, in $O(n)$ steps, where $n$ is the number of arcs of $G$ (see [1, p. 185]). We then separate (1)–(4) with $G$ into $r$ multicommodity flow problems of type (1)–(4), where the $j$th problem has $G^j$ as underlying graph. External flows are found as follows. One of the $G^j$, say $G^1$, must be joined to the remaining $\bar{G} = \cup_{j=2}^{r} G^j$ at exactly one node, say at $x$. We now separate $G$ into $G^1$ and $\bar{G}$. External flows at nodes not equal to $x$ are those of $G$. External flows at $x$ of $G^1$ ($\bar{G}$) are chosen so that external flows for that graph sum to zero for each commodity. Proceeding iteratively, we remove one $G^j$ at a time from $\bar{G}$ until $r$ multicommodity flow problems have been determined. It is trivial to show that each such subproblem is transformable if the original problem is transformable. The converse statement follows from theorem 1.

**Acknowledgment.** We thank the anonymous referees for the time and effort they devoted to the evaluation of the paper. In addition one referee suggested several changes that resulted in an improved exposition.

## REFERENCES

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms,* Addison-Wesley, Reading, MA, 1974.
[2] R. E. Bixby and W. H. Cunningham, *Converting linear programs to network problems,* Working Paper, Carleton University, 1978.
[3] G. B. Bradley, G. G. Brown and G. W. Graves, *Design and implementation of large scale primal transshipment algorithms,* Management Sci., 24 (1977), pp. 1–34.
[4] J. R. Evans, *A single commodity transformation for certain multicommodity networks,* Operations Research, 26 (1978), pp. 673–680.
[5] ———, *A survey of integrality in multicommodity networks,* Working Paper, Dept. of Quantitative Analysis, University of Cincinnati, 1977.
[6] M. Iri, *On the synthesis of loop and cutset matrices and the related problems,* RAAG Memoirs, 4 (1968), pp. 376–410.
[7] Y. Soun, *Transformable Multicommodity Networks,* Ph.D. dissertation, University of Texas, Dallas, 1978.
[8] K. Truemper and Y. Soun, *Minimal Forbidden Subgraphs of Unimodular Multicommodity Networks,* Math. Oper. Res., 4 (1979), pp. 379–389.
[9] K. Truemper, *Unimodular matrices of flow problems with additional constraints,* Networks, 7 (1977), pp. 343–358.
[10] W. T. Tutte, *Matroids and graphs,* Trans. Amer. Math. Soc., 90 (1959), pp. 527–552.
[11] ———, *An algorithm for determining whether a given binary matroid is graphic.* Proc. Amer. Math. Soc., 11 (1960), pp. 905–917.
[12] ———, *From matrices to graphs.* Canad. J. Math., 16 (1964), pp. 108–127.
[13] D. J. A. Welsh, *Matroid Theory,* Academic Press, London, 1976.
[14] H. Whitney, *2-Isomorphic Graphs,* Amer. J. of Math., 55 (1933), pp. 245–254.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 3, September 1980


0196-5212/80/0103-0014 $01.00/0

# A GRAPH THEORETICAL APPROACH TO HANDICAP RANKING OF TOURNAMENTS AND PAIRED COMPARISONS*

KENNETH A. BERMAN†

**Abstract.** In this paper, the players in a tournament $T$ are ranked according to the number of out arborescences rooted at the vertices of a digraph $D$ associated with $T$. This ranking is shown to be equivalent to handicap ranking of tournaments. The tournaments considered in this paper may involve any number of games or no games between particular pairs of players and thus are more general than round-robin tournaments. In fact, these tournaments can be interpreted in statistics as systems of paired comparisons.

**1. Introduction.** By a "tournament" we will mean a competition involving players $P_1, P_2, \cdots, P_n$ where all games are between two players. The outcome of a game between two players $P_i$ and $P_j$ is one of three possibilities: (i) $P_i$ defeats $P_j$ (ii) $P_j$ defeats $P_i$ (iii) draw. In a *tieless* tournament the third possibility is not allowed. A tieless tournament $T$ corresponds to a digraph $D$ where multiple directed edges are allowed but no loops; the vertices of $D$ correspond to the $n$ players of $T$ and a directed edge of $D$ joins a vertex $i$ to a vertex $j$ for each game of $T$ in which player $P_i$ defeats player $P_j$. A tournament with draws may be transformed into a tieless tournament by replacing each game in which $P_i$ defeats $P_j$ by two games in which $P_i$ defeats $P_j$ and replacing each draw between $P_i$ and $P_j$ with two games, one in which $P_i$ defeats $P_j$ and the other in which $P_j$ defeats $P_i$.

The vertices of digraph $D$ are partitioned by the strongly connected components. (A *strongly connected* digraph is a digraph such that there is a dipath from every vertex to every other vertex. A *strongly connected component* of $D$ is a maximal subdigraph which is strongly connected.) It is easily shown that these components can be ranked such that every player in a component of higher rank defeats every player in a component of lower rank. Now by ranking the players in each component we obtain a ranking of all the players. Thus, the problem of ranking tournaments reduces to the problem of ranking *strong* tournaments, that is, tieless tournaments whose associated digraph is strongly connected.

In the papers of Moon and Pullman [3], [4] and the independent paper of Daniels [1] equitable systems of handicapping are devised for the players. In one method, discussed in these papers, each player $P_i$ pays the amount $s_i$ $(i = 1, 2, \cdots, n)$ to the winner of every game he loses, where $s_i$ is a positive real number. A fairness condition is imposed on the *payoff* vector $\mathbf{s} = (s_1, s_2, \cdots, s_n)$. Payoff vector $\mathbf{s}$ is *fair* if each player has a net gain of zero. The component $s_i$ of $\mathbf{s}$ can be thought of as the "strength" of player $P_i$. The players of greater strength are ranked above those of lesser strength.

An *out arborescence* at a vertex $i$ of $D$ is a spanning tree such that there is a dipath in the tree from $i$ to every other vertex. The *arborescence vector* is the vector $\mathbf{a} = (a_1, a_2, \cdots, a_n)$ where $a_i$ is the number of out arborescences at vertex $i$ of $D$. In this paper, we show that for a strong tournament the arborescence vector $\mathbf{a}$ is a fair payoff vector and every fair payoff vector is a scalar multiple of $\mathbf{a}$.

In § 2 we prove this result and in § 3 we obtain, as a corollary, a result on the existence of nowhere-zero $k$-flows.

* Received by the editors September 13, 1979 and in revised form February 28, 1980.

† Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England.

**2. Handicap ranking and the arborescence vector.** In this section we prove the following main theorem.

THEOREM 2.1. *Let $T$ be a strong tournament with associated digraph $D$. Then, the arborescence vector* $\mathbf{a}$ *of $D$ is a fair payoff vector and every fair payoff vector is a scalar multiple of* $\mathbf{a}$.

Since $D$ is strongly connected, $a_i > 0$ for all $i$. To prove the first part of the theorem we must show that

$$a_i d_i = \sum_{\vec{ij} \in D} a_j \tag{1}$$

for each $i$, where $d_i$ denotes the in-degree of $i$ and the sum is over all directed edges of the type $\vec{ij}$ in $D$.

Let us define a $(*)$-graph as any spanning subgraph of $D$ in which each vertex has in-degree one and which contains exactly one cycle. (It is the converse of what is sometimes called a connected functional digraph.) Let $S(i)$ denote the set of $(*)$-graphs of $D$ in which vertex $i$ belongs to the cycle. If from any graph $G$ in $S(i)$ we remove the (unique) cyclic edge $\vec{ij}$ directed away from $i$, we obtain an out arborescence at vertex $j$. This process is clearly reversible, and it follows readily that $|S(i)| = \sum_{\vec{ij} \in D} a_j$.

On the other hand, if from $G$ we remove the (unique) cyclic edge directed towards $i$, we obtain an out arborescence at vertex $i$. Each such arborescence is obtained $d_i$ times when this process is applied to all graphs $G$ in $S(i)$. Again, it readily follows that $|S(i)| = a_i d_i$.

Thus the two sides of equation (1) are equal because they are simply two different expressions for the same thing, namely $|S(i)|$. This proves the first part of Theorem 2.1.

To prove the second part, suppose $\mathbf{b} = (b_1, b_2, \cdots, b_n)$ is any fair payoff vector. Then, for every vertex $i$, $b_i > 0$, and

$$b_i d_i = \sum_{\vec{ij} \in D} b_j. \tag{2}$$

Since all the $a_i$'s are positive there exists a real number $r$ such that $c_i = ra_i - b_i \geqq 0$ for each $i$ and $c_k = 0$ for at least one vertex $k$. Combining (1) and (2), we obtain

$$\sum_{\vec{kj} \in D} c_j = c_k d_k = 0. \tag{3}$$

Since all the $c_i$'s are nonnegative it follows that $c_j = 0$ for each vertex $j$ such that $\vec{kj} \in D$. Since $D$ is strongly connected we have that $c_i = 0$ for every vertex $i$, i.e., $\mathbf{b} = r\mathbf{a}$.

This proves Theorem 2.1.

Theorem 2.1 can be easily generalized to *weighted* tournaments, that is, tournaments in which each game is assigned a positive real weight. A ranking is obtained in terms of the weighted arborescences where the weight of an arborescence is the product of the weights on its edges. Weighted tournaments can be employed to rank a set of players where the probability $p_{ij}$ that player $P_i$ will defeat $P_j$ is given for some (or all) pairs of players $P_i$, $P_j$. In the associated weighted tournament there corresponds a game in which $P_i$ defeats $P_j$ for each probability $p_{ij}$ and this game is given weight $p_{ij}$.

**3. Nowhere-zero $k$-flows.** Let $G$ be an undirected graph with vertex set $V$ and

edge set $E$. Assign an orientation to the edges of $G$. A *flow* $\phi$ in $G$ over a ring $R$ is a mapping from $E$ into $R$ such that the sum of the values of $\phi$ over the edges directed into each vertex equals the sum over the edges directed out of that vertex. (If a different orientation of $G$ is used we replace $\phi(e)$ by $-\phi(e)$ for each edge $e$ whose direction is reversed.) If no edge is mapped onto zero then $\phi$ is a *nowhere-zero* flow. A *k-flow* is a flow over the integers mod $k$.

Let $\alpha$ be the map which maps edge $e \in E$ onto the number of out arborescences at the head of $e$. By Theorem 2.1, $\alpha$ is a flow over the integers and $\alpha \pmod{k}$ is a nowhere-zero $k$-flow if the number of out arborescences at each vertex is not divisible by $k$. Hence we have the following theorem.

THEOREM 3.1. *A graph which can be oriented so that the number of out arborescences at each vertex is not divisible by k has a nowhere-zero k-flow.*

Now, a planar graph is face $k$-colorable iff it contains a nowhere-zero $k$-flow. (See [5].) Thus we have the following result.

COROLLARY 3.2. A planar graph $G$ is face $k$-colorable ($k = 2, 3, 4$) if $G$ can be oriented so that the number of out arborescences at each vertex is not divisible by $k$.

Tutte conjectured in [5] that every graph without bridges contains a nowhere-zero 5-flow. It can be easily shown that the existence of a nowhere-zero $k$-flow implies the existence of a nowhere-zero $m$-flow for all $m > k$. Jaeger has proved that every graph without bridges has a nowhere-zero 8-flow [2]. The following problem is proposed: Which graphs can be oriented so that the number of out arborescences at each vertex is not divisible by 5?

## REFERENCES

[1] H. E. DANIELS, *Round-robin tournament scores,* Biometrika 56 (1969), pp. 295–300.

[2] F. JAEGER, *On nowhere-zero flows in multigraphs, Proceedings of the Fifth British Combinatorial Conference* (Univ. Aberdeen, Aberdeen, 1975), Congressus Numerantium, No. XV, Utilitas Math., Winnepeg, Man., 1976, pp. 373–378.

[3] J. W. MOON AND N. J. PULLMAN, *Tournaments and handicaps,* Information Processing 68 (Proc. IFIP Congress, Edinburgh, 1968), Vol. 1: Mathematics, Software, North-Holland, Amsterdam, 1969, pp. 219–223.

[4] J. W. MOON AND N. J. PULLMAN, *On generalized tournament matrices,* SIAM Rev., 12 (1970), pp. 384–399.

[5] W. T. TUTTE, *A contribution to the theory of chromatic polynomials,* Canad. J. Math., 6 (1954), pp. 80–91.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104-0001 $01.00/0

# DYNAMIC-PROGRAMMING ALGORITHMS FOR RECOGNIZING SMALL-BANDWIDTH GRAPHS IN POLYNOMIAL TIME*

JAMES B. SAXE†

**Abstract.** In this paper we investigate the problem of testing the bandwidth of a graph: Given a graph, $G$, can the vertices of $G$ be mapped to distinct positive integers so that no edge of $G$ has its endpoints mapped to integers which differ by more than some fixed constant, $k$? We exhibit an algorithm to solve this problem in $O(f(k)N^{k+1})$ time, where $N$ is the number of vertices of $G$ and $f(k)$ depends only on $k$. This result implies that the "Bandwidth $\stackrel{?}{\leqq} k$" problem is not NP-complete (unless P = NP) for any fixed $k$, answering an open question of Garey, Graham, Johnson, and Knuth. We also show how the algorithm can be modified to solve some other problems closely related to the "Bandwidth $\stackrel{?}{\leqq} k$" problem.

**1. Introduction.** The subject of this paper is the computational complexity of a problem on graphs. To speak precisely of the problem, we will need the following notation and definitions.

NOTATION. Let $u$ and $v$ be vertices of a graph $G$. We will say "$u$—$v$ in $G$" to denote that $\{u, v\}$ is an edge of $G$. Where $G$ is clear from context, we will write simply "$u$—$v$".

DEFINITIONS. Let $G$ be a graph with vertex set $V$, and let $N = |V|$. A *layout* of $G$ is a one-to-one mapping, $f$, from $V$ onto $\{1, \cdots, N\}$. The *bandwidth* of $f$ is defined as the maximum distance between the images under $f$ of any two vertices that are connected by an edge of $G$. That is,

$$\text{bandwidth}(f) = \max \{f(u) - f(v) \mid u\text{—}v\}.$$

The *Bandwidth* of $G$ is defined as the least possible bandwidth for any layout of $G$. Thus,

$$\text{Bandwidth}(G) = \min \{\text{bandwidth}(f) \mid f \text{ is a layout of } G\}.$$

PROBLEM "(Bandwidth minimization)." Given an arbitrary graph, $G$, and a positive integer, $k$, determine whether Bandwidth($G$) $\leqq k$.

Note that the notion of graph bandwidth is equivalent to the more familiar notion of matrix bandwidth in that Bandwidth($G$) $\leqq k$ iff there exists a permutation matrix $P$ such that $(PCP^{-1})_{i,j} = 0$ whenever $|i - j| > k$, where $C$ is $G$'s connection matrix. For any particular positive integer $k$, we can define a restricted version of the bandwidth minimization problem as follows:

PROBLEM "(Bandwidth $\stackrel{?}{\leqq} k$)." Given a graph, $G$, determine whether Bandwidth($G$) $\leqq k$.

Papadimitriou [1976] has shown that the general bandwidth minimization problem, in which $k$ is specified in the input, is NP-complete. The problem was later studied by Garey, Graham, Johnson, and Knuth [1978], who found a linear-time algorithm for the problem "Bandwidth $\stackrel{?}{\leqq} 2$", and also improved on Papidimitriou's result by showing the problem for general $k$ to be NP-complete even when $G$ is restricted to be a tree with no vertex of degree greater than three. A number of questions are left open by their work, however. One such question is whether there exists a polynomial-time algorithm for the problem "Bandwidth $\stackrel{?}{\leqq} 3$". In this paper, we

* Received by the editors February 26, 1980, and in final form March 18, 1980. This work was supported by the U.S. Office of Naval Research under Contract N00014-76-C-0370.

† Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

will answer this question affirmatively by exhibiting an algorithm[1] which solves the problem "Bandwidth $\stackrel{?}{\leqq} k$" in polynomial time for any fixed $k$. Section 2 of this paper introduces the fundamental concepts and assumptions we will use in describing our algorithm. In § 3 the algorithm is described and its performance is analyzed. In § 4 we discuss some modifications of the algorithm to solve related problems. Finally, in § 5, we discuss some remaining open problems.

**2. Fundamental concepts and assumptions.** Throughout the following we will assume that $G$ denotes a graph with vertex set $V$ and edge set $E$, that $k$ denotes a particular positive integer,[2] and that we wish to determine whether $G$ has any layout of bandwidth $\leqq k$. We let $N$ denote the cardinality of $V$. Note that if $G$ is not connected then $G$ has a layout of bandwidth $\leqq k$ iff each of its components has such a layout. Also, it is clearly impossible for $G$ to have such a layout if $G$ has any vertex of degree greater than $2k$. We therefore assume, without loss of generality, that $G$ is a connected graph having no vertex of degree greater than $2k$. Note that an arbitrary graph can be partitioned into its connected components by depth-first search in $O(\max(n, e))$ time, where $n$ is the number of vertices and $e$ is the number of edges,[3] and that this is $O(n)$ if a fixed bound is given on the degree of any vertex. Moreover, an obvious modification to the depth-first search algorithm allows it to detect the presence of a vertex with degree greater than a fixed bound in time which is proportional only to the number of vertices and not to the number of edges.

We now introduce the key notion of a *partial* layout.

DEFINITIONS. A *partial* layout of $G$ is one-to-one function, $f$, from some subset of $V$ onto $\{1, \cdots, M\}$, for some $M$ such that $0 \leqq M \leqq N$. We say that $f$ is *feasible* if it can be extended to a (total) layout, $g$, such that bandwidth$(g) \leqq k$. The *bandwidth* of $f$ is the maximum distance between the images of any two edge-connected vertices of $G$ which are in the domain of $f$. If $u$—$v$ and $u$ is in the domain of $f$ and $v$ is not, then the edge $\{u, v\}$ is said to be *dangling* from $f$.

Consider a partial layout, $f$, of size $M$. Clearly, $f$ cannot possibly be feasible unless

1. bandwidth$(f) \leqq k$, and
2. whenever $u$ and $v$ are vertices of $G$ such that $f(u) < M - k$ and $u$—$v$, $v$ is also in the domain of $f$.

If $f$ satisfies both these conditions, then $f$ is said to be a *plausible* partial layout. The sequence $(f^{-1}(\max(M-k+1, 1)), \cdots, f^{-1}(M))$, taken together with the set of dangling edges of $f$, is called the *active region* of $f$. We now come to the theorem on which our principal algorithm depends.

THEOREM 2.1. *Let $f$ and $g$ be two plausible partial layouts of $G$ having identical active regions. Then,*

1. *$f$ and $g$ have identical domains, and*
2. *$f$ is feasible iff $g$ is feasible.*

*Proof.* Since $G$ is connected, the domains of $f$ and $g$ must each consist precisely of those vertices which are path-connected to vertices in the active region by paths not including any dangling edges. Thus, (1) holds. To see that (2) holds, we need only

[1] More correctly, a class of algorithms, one for each value of $k$.

[2] When using the "big-oh" notation, we will regard $k$ as fixed and therefore omit factors that depend only on $k$.

[3] See, for example, Aho, Hopcroft, and Ullman [1974, Chapt. 5].

note that any assignment of the remaining vertices which extends either $f$ or $g$ to a total layout of bandwidth $\leqq k$ must also extend the other to such a layout. □

Finally, we define the notion of a *successor* of a plausible partial layout (or active region), which will be necessary to explain our algorithms.

DEFINITION. Let $f$ be a plausible partial layout of $G$. Then a *successor* of $f$ is a plausible partial layout, $g$, which extends $f$ by precisely one element. In this case, the active region of $g$ is also said to be a *successor* of the active region of $f$. We also say that (the active region of) $f$ is a *predecessor* of (the active region of) $g$.

**3. The algorithm.** Theorem 2.1 allows us to say that two plausible partial layouts are *equivalent* if they have identical active regions. The algorithm we present is essentially a breadth-first search over the space of all the induced equivalence classes of plausible partial layouts, where each such equivalence class is uniquely characterized by the active region of its representatives. Alternatively, we may think of the algorithm as a dynamic-programming search over the plausible partial layouts. Each active region consists of at most $k$ vertices and each vertex has no more than $2k$ edges, each of which may or may not be dangling. Thus the number of equivalence classes is bounded above by[4]

$$\sum_{0 \leqq i \leqq k} \binom{N}{i} (i!)(2^{2k})^i = O(N^k).$$

Our algorithm uses the following two data structures:

1. A (fifo) queue, $Q$, whose elements are active regions.
2. An array, $A$, which contains one element for each possible active region. Each element, $A[r]$, of $A$ consists of a Boolean flag, $A[r]$.examined, telling whether the active region $r$ has already been considered in the search and a list, $A[r]$.unplaced, of vertices which is intended to list all vertices NOT in the domain of each plausible partial layout with active region $r$.

At the start of our algorithm, $Q$ is initialized to contain the single element representing the active region (henceforward denoted $\Phi$) of the empty partial layout, $\phi$. The flag $A[\Phi]$.examined is set to TRUE and $A[\Phi]$.unplaced is initialized to list all the elements of $V$. The remaining $A[r]$.examined are initially FALSE, and the remaining $A[r]$.unplaced are uninitialized. The algorithm now proceeds as follows:

ALGORITHM B (Bandwidth testing).

1. Extract an active region, $r$, from the head of $Q$.
2. From $A[r]$.unplaced, determine the successors of $r$.[5]
3. For each successor, $s$, of $r$ such that $A[s]$.examined is FALSE, perform the following steps:
   a. Set $A[s]$.examined to TRUE.
   b. Compute $A[s]$.unplaced by deleting the last vertex of $s$ from $A[r]$.unplaced.
   c. If $A[s]$.unplaced is the empty set, then halt asserting that Bandwidth$(G) \leqq k$.
   d. Insert $s$ at the end of $Q$.

---

[4] As we will mention in § 5, the coefficient on this bound is quite loose.

[5] For each vertex, $v$, in $A[r]$.unplaced, we must consider the possibility of using $v$ as the next vertex in any plausible partial layout which has $r$ as its active region. For any $v$ this will either (1) yield a new plausible partial layout whose active region is uniquely determined by $r$ and $v$ or (2) yield an implausible partial layout regardless of the particular plausible partial layout being extended by $v$ (provided that it has $r$ as its active region).

4. If $Q$ is empty, then halt asserting that Bandwidth($G$) $> k$. Otherwise, go to Step 1.

The space required by this algorithm is clearly $O(N^{k+1})$. To determine the running time, we note first that since there are $O(N^k)$ possible active regions, each of Steps 1 through 4 will be executed $O(N^k)$ times. The individual executions of Steps 1 and 4 each take only constant time, so the contribution of these steps to the total running time of the algorithm is $O(N^k)$. Since any active region, $r$, has at most $N$ successors (zero or one for each element of $A[r]$.unplaced), each execution of Step 2 takes $O(N)$ time. The contribution of Step 2 to the total execution time is therefore $O(N^{k+1})$. Determining the contribution of Step 3 is (a little) trickier. During a single execution of Step 3, Steps 3.a through 3.d may be executed as many as $N$ times, and the amount of computation in Step 3.b may be $\theta(N)$. Thus it appears possible that Step 3 may contribute $\theta(N^{k+2})$ to the total execution time. If we look more carefully, however, we see that 3.a through 3.d are executed at most once for each active region. Thus the total contribution of Step 3 is $O(N^{k+1})$. Adding the contributions of all the steps gives us the following result.

THEOREM 3.1. *Let $k$ be any positive integer. Then there is an algorithm which solves the problem "Bandwidth $\leqq k$" using $O(N^{k+1})$ time and $O(N^{k+1})$ space.*

*Proof.* To test the bandwidth of $G$, we first perform an $O(N)$-time depth-first search which either

(1) determines that $G$ has some vertex of degree greater than $2k$, or
(2) partitions $G$ into connected components none of which have any vertex of degree greater than $2k$.

In case (1), we know immediately that Bandwidth($G$) $> k$. In case (2), we apply Algorithm B to the connected components of $G$. □

While Algorithm B will tell us whether $G$ has a layout of bandwidth $\leqq k$, it does not actually produce such a layout. In order to allow such a layout to be recovered, we may associate with each active region, $s$, an additional field, $A[s]$.predecessor. When $s$ is appended to $Q$ in Step 3.d., we make $A[s]$.predecessor point to a predecessor of $s$ (namely, the $r$ we chose in Step 1).[6] If the algorithm finds an active region, $t$, such that $A[t]$.unplaced is empty, it is a simple matter to recover a layout by tracing back through the predecessor fields.

**4. Modifications for related problems.** Another question left open by Garey, Graham, Johnson, and Knuth [1978] is whether there exists a polynomial-time algorithm to *count* the layouts of a graph having bandwidth $\leqq k$, even for $k = 2$. We now give an affirmative answer to a closely related question by exhibiting a class of polynomial-time algorithms (one for each positive integer $k$) for determining the number of bandwidth $\leqq k$ layouts of any *connected* graph.[7]

Our algorithm for enumerating layouts of bandwidth $\leqq k$ is a slightly modified form of Algorithm B. The data structures are the same as those for Algorithm B, with the following additions:

1. Each entry, $A[r]$, of $A$ has a third field, $A[r]$.count, which will hold the number of (so far discovered) plausible partial layouts whose active region is $r$.

[6] Note that this pointer need only name the single vertex (if any) which is contained in $r$ but not in $s$.

[7] Note that the number of bandwidth $\leqq k$ layouts of an arbitrary graph is not uniquely determined by the numbers of bandwidth $\leqq k$ layouts of its connected components because the topologies of the components impose constraints on how the various layouts may overlap. The algorithms cannot be applied directly to nonconnected graphs because they depend on Theorem 2.1.

2. There is a variable, Total, which will hold the number of (so far discovered) layouts of bandwidth $\leqq k$.

At the start of the algorithm, Total and all the $A[r]$.count are initialized to zero, except for $A[\Phi]$.count, which is initialized to 1. The remaining variables are initialized as for Algorithm B. We then proceed as follows.

ALGORITHM E (Enumerate layouts).

1. Extract an active region, $r$, from the head of $Q$.
2. From $A[r]$.unplaced, determine the successors of $r$.
3. For each successor, $s$, of $r$, perform the following steps:
   a. If $A[s]$.examined is TRUE, go to $f$.
   b. Set $A[s]$.examined to TRUE.
   c. Compute $A[s]$.unplaced by deleting the last vertex of $s$ from $A[r]$.unplaced.
   d. If $A[s]$.unplaced is the empty set, then increase Total by $A[r]$.count and go to Step 4.
   e. Insert $s$ at the end of $Q$.
   $f$. Increase $A[s]$.count by $A[r]$.count.
4. If $Q$ is empty, then halt. Otherwise, go to Step 1.

Study of this algorithm gives us the following result.

THEOREM 4.1. *Let $k$ be any positive integer. Then there exists an $O(N^{k+1})$-time, $O(N^{k+1})$-space algorithm which, given any connected graph, $G$, computes the number of layouts of $G$ having bandwidth $\leqq k$.*

*Proof.* We claim that Algorithm E (preceded by a depth-first search to ensure that no vertex of $G$ has degree greater than $2k$) has the desired properties. By an analysis similar to that for Algorithm B, Algorithm E will run in $O(N^{k+1})$ time. We must now show that it correctly counts the layouts of bandwidth $\leqq k$. To do this, it suffices to show that by the time that any plausible partial layout, $r$, is selected in Step 1, $A[r]$.count contains the total number of plausible partial layouts whose active region is $r$. This in turn may be shown inductively if we can only show that no active region, $r$, is chosen in Step 1 until every predecessor of $r$ has been chosen. This last follows at once from the fact (which may be established by induction) that the active regions proceed through the queue in nondecreasing order of their lengths, where the *length* of an active region, $r$, is defined to be the number of vertices in the domain of any plausible partial layout whose active region is $r$. □

We may view bandwidth minimization as the problem of finding a layout with minimax edge length. We will now look at the corresponding minisum problem.

DEFINITION. Let $G$ be a graph with edge set $E$, and let $f$ be a layout of $G$. Then the *total edge length* of $f$ is given by the sum

$$\sum_{\{u,v\}\in E} |f(u) - f(v)|,$$

where each edge, $\{u, v\}$, contributes precisely once to the sum (rather than once as $u$—$v$ and once as $v$—$u$).

PROBLEM (Optimal Linear Arrangement.) Given a graph, $G$, and an integer, $t$, determine whether there is a layout of $G$ having total edge length less than or equal to $t$.

The Optimal Linear Arrangement (O.L.A.) problem was found to be NP-complete by Garey, Johnson, and Stockmeyer [1976]. However, Shamos [1979] has pointed out that the methods of the present work can be used to provide polynomial-time algorithms for a class of restricted versions of O.L.A. For every positive integer, $k$, we define a restriction of O.L.A. as follows.

PROBLEM. (O.L.A. for bandwidth $\leqq k$). Given a graph, $G$, determine the minimal total edge length of any layout of $G$ having bandwidth $\leqq k$ or determine that no such layout exists.

Applying the methods used above, we obtain the following result.

THEOREM 4.2. *Let $k$ be any positive integer. Then there exists an algorithm which solves* O.L.A. *for bandwidth $\leqq k$ in $O(N^{k+1})$ time and $O(N^{k+1})$ space.*

*Proof.* An algorithm having the desired properties when applied to connected graphs with no vertex having degree greater than $2k$ may be constructed by a slight modification of Algorithm E: instead of maintaining with each active region a count of the partial layouts having that active region, we maintain an indication of the minimum sum of the lengths of all edges whose endpoints are in the domains of all plausible partial layouts having that active region. The details are left to the reader. For arbitrary graphs we first perform a depth-first search which either detects the presence of a vertex with degree greater than $2k$ (implying that Bandwidth($G$) $\geqq k$) or partitions $G$ into its connected components, taking linear time in either case. We then compute the minimal total edge length for $G$ by finding and summing the minimal total edge lengths for the connected components. □

We note that the previous result remains valid if we consider edge-weighted graphs and the "total edge length" is taken as a weighted sum. For connected graphs, we can also use the method of Algorithm E to obtain a count of the layouts with minimal total edge length for bandwidth $\leqq k$.

Finally, all the previous results extend to "directed" versions of bandwidth minimization and O.L.A., in which $G$ is a directed graph and a layout, $f$, is acceptable only if $f(u) < f(v)$ whenever $(u, v)$ is an edge of $G$.[8]

**5. Open problems.** The most obvious problem left open by this work is that of improving the performance of Algorithm B. Although the expense of this algorithm is "only polynomial" in the size of the examined graph, it is still sufficiently expensive (particularly in terms of space) to render it impractical for all but the smallest cases (consider, for example, determining whether Bandwidth($G$) $\leqq 5$, where $G$ is a graph of forty vertices). The fact that Garey, Graham, Johnson, and Knuth [1978] have a linear-time algorithm for "Bandwidth $\overset{?}{\leqq} 2$", while Algorithm B takes cubic time for the same problem, offers some hope that the degree of the polynomial can be reduced for higher values of $k$ as well. Indeed, is conceivable (even if P $\neq$ NP) that there are linear algorithms for all values of $k$, with coefficients growing exponentially in $k$.

One approach to improving the performance is to attempt to reduce the number of active regions examined, and this can indeed be done to some extent. For example, we may prune the search by noting that, while a plausible partial layout may have $\theta(k^2)$ dangling edges, such a partial layout cannot actually be feasible if those edges lead to more than $k$ distinct vertices. Unfortunately, graphs of the form

$$v_1 — v_2 — \cdots — v_{N-1} — v_N$$

supply an existence proof that the number of equivalence classes of plausible partial layouts of bandwidth $\leqq k$ can in fact be $\theta(N^k)$.

In Algorithm B, we reduce the search space from the set of all plausible partial layouts to the much smaller set of equivalence classes of partial layouts. To look at it

[8] A good starting point for the reader who is interested in learning more about bandwidth minimization, O.L.A., and their variations is Appendix A1 of Garey and Johnson [1979].

another way, given two partial layouts, $f$ and $g$, if we recognize (by equality of active regions) that $f$ is feasible iff $g$ is feasible, then we feel free to search for completions of only one of the partial layouts. The algorithm of Garey, Graham, Johnson, and Knuth cuts down the search space by methods which are similar but more sophisticated. In particular, they can avoid searching for completions of a partial layout,[9] $f$, by choosing to search for completions of a layout, $g$, such that $g$ is feasible whenever $f$ is feasible, but not necessarily only when $f$ is feasible.

It is interesting to note that "worst-case" numbers of feasible active regions seem to arise precisely in circumstances where large pieces of the graph can be laid out in bandwidth much less than $k$. We define a *maximal* graph of size $N$ and bandwidth $k$ as a graph whose edge set is $\{\{v_i, v_j\} \mid |i - j| \leqq k\}$, where $\{v_i \mid 1 \leqq i \leqq N\}$ is the vertex set.[10] The algorithm of Garey, Graham, Johnson, and Knuth relies heavily on the fact that if all the even-numbered vertices or all the odd-numbered vertices are deleted from a maximal graph of bandwidth 2, the induced graph on the remaining vertices is a maximal graph of bandwidth 1. For testing higher bandwidths it is possible that similar use may be made of the fact that deleting every $k$th vertex from a maximal graph of bandwidth $k$ leaves a maximal graph of bandwidth $k - 1$.

Another potentially fruitful course of investigation would be to look for efficient algorithms for approximate bandwidth minimization. For example, given a graph, $G$, we may wish to produce a layout for $G$ whose bandwidth is no more than, say, twice the minimum possible. To the author's knowledge it has not yet been determined whether this problem (when phrased as a language recognition problem) is NP-complete.

**Acknowledgments.** The author gratefully acknowledges the helpful comments of Jon L. Bentley, Michael R. Garey, Christos H. Papadimitriou, Michael I. Shamos, and the referee.

## REFERENCES

A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley. Reading, MA, 1974.

M. R. Garey, R. L. Graham, D. S. Johnson, and D. E. Knuth, *Complexity results for bandwidth minimization*, SIAM J. Appl. Math., 34 (1978), pp. 477–495.

M. R. Garey and D. S. Johnson, *Computers and Intractibility: A Guide to the Theory of* NP-*Completeness*. W. H. Freeman and Company, San Francisco, 1979.

M. R. Garey, D. S. Johnson, and L. Stockmeyer, *Some simplified* NP-*complete graph problems*, Theor. Comput. Sci., 1 (1976), pp. 237–267.

C. H. Papadimitriou, *The* NP-*completeness of the bandwidth minimization problem*, Computing, 16 (1976), pp. 263–270.

M. I. Shamos, Private communication, 1979.

[9] In their terminology, a partial layout of $G$ is a map from a subset of the vertices of $G$ to an *arbitrary* set of integers.

[10] Note that a graph of $N$ vertices has bandwidth $\leqq k$ iff it is isomorphic to a subgraph of a maximal graph of size $N$ and bandwidth $k$.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104-0002$01.00/0

# AN $O((n \log p)^2)$ ALGORITHM FOR THE CONTINUOUS $p$-CENTER PROBLEM ON A TREE*

R. CHANDRASEKARAN† AND A. TAMIR‡

**Abstract.** This paper considers the problem of locating $p$ facilities on a tree network in order to minimize the maximum of the distances of the points on the network to their respective nearest facilities. An $O((n \log p)^2)$ algorithm for a tree network with $n$ nodes is presented.

**Introduction.** In this study we consider center location problems on undirected tree networks. Let $T = T(N, A)$ be an undirected tree, with $N$ and $A$ denoting the set of all nodes and the set of all arcs respectively. With each arc is associated a positive number called the length of the arc. We assume that $T$ is embedded in the Euclidean plane, so that the arcs are line segments whose endpoints are the nodes, and arcs intersect one another only at nodes. (Any tree with positive arc-lengths can be so embedded in $R^2$. See [6].) Using this embedding we can then talk about points, not necessarily nodes, on the arcs, and denote by $d(x, y)$ the distance, measured along the arcs of the tree, between any two points $x$, $y$ of the tree $T$.

In addition, a set, $D$, of points on $T$ is specified. $D$, which may be finite or infinite in cardinality, represents the set of demand points. Assume that supply centers can be located anywhere on the tree. Given a number, $p$, the objective is to find locations for $p$ supply points on $T$, such that the supremum of the distances of the demand points in $D$ to their respective nearest supply centers is minimized.

Two special cases of the above model have been treated in the literature. The first corresponds to the case where demand occurs only at the nodes of $T$, i.e., $D = N$. Whenever $|D| < \infty$, one can also associate weights with the demand points and consider minimizing the maximum of the weighted distances to the nearest supply centers. Efficient, polynomially bounded algorithms when $D = N$ are given in [13], [3] for general $p$, while further specializations when $p \leqq 2$ are discussed in [6], [8], [9], [10], [11], [14].

The second special case of the general model is the continuous case when $D = T$; i.e., each point of the tree is a demand point. This model is studied in [2], where it is solved in polynomial time.

The general model introduced above is related to the following $p$-center dispersion problem. A set, $S$, of points on the tree $T$ is specified. Given an integer $p$, the objective is to locate $p$ facilities at points in $S$ such that these $p$ facilities are as far from each other as possible.

In this study we focus on the case when the sets $D$ and $S$ in the center location and center dispersion problems, respectively, are identical and equal to the entire tree. Theorem 1 below, (due to Shier [14]), shows a duality result between the $p$-center location and $(p + 1)$- center dispersion problems, when $D = S = T$. It is convenient for the statement of the theorem to let $U_p = \{u_1, \cdots, u_p\}$ and $V_{p+1} = \{v_1, \cdots, v_{p+1}\}$ denote any finite subsets of $T$ of cardinalities $p$ and $p + 1$ respectively, and to define

$$f_D(U_p) = \max_{x \in D=T} \{\min_{u_i \in U_p} d(x, u_i)\}, \tag{1}$$

* Received by the editors March 19, 1979, and in final form March 28, 1980. This work was done while the authors were visiting Northwestern University.

† University of Texas at Dallas, Dallas, Texas.

‡ Tel-Aviv University, Faculty of Social Sciences, Department of Statistics, Tel-Aviv, Israel.

and

$$g(V_{p+1}) = \min \{d(v_i, v_j)/2 : 1 \leqq i < j \leqq p + 1\}. \tag{2}$$

THEOREM 1. [14]. *Let* $D = S = T$. *Then*

$$\min \{f_D(U_p) : U_p \subseteq T, |U_p| = p\} = \max \{g(V_{p+1}) : V_{p+1} \subseteq S = T, |V_{p+1}| = p + 1\}. \tag{3}$$

We mention that the proof in [14] can be modified to validate the above duality result for the general case when $D = S$ and $D$ is any subset of $T$ with $|D| > p$. (One would have to replace the min and max operations of (1) and (3) by inf and sup, respectively, and also omit the equality of $S$ and $D$ to $T$ in (1) and (3).) Another specialization of the general case, i.e., $D = S$ and $p < |D| < \infty$, is proved in [3], using the equality of the maximum anticlique and the minimum cardinality clique cover in perfect graphs.

Focusing on the subject of this paper, i.e., $D = S = T$, we show that for a given $p$ the minimum value of the objective function of the $p$-center location problem is equal to $d(i, j)/2k$, where $d(i, j)$ is the distance between some pair of nodes, $i$ and $j$, of $T$, and $k$ is an integer satisfying $1 \leqq k \leqq p$. This result is then used to improve the algorithm of [2], yielding the bound of $O((n \log p)^2)$ for the continuous $p$-center location problem, i.e., $D = T$, on a tree $T(N, A)$ with $n$ nodes. (Logarithms are taken to the base 2.) We also indicate how to improve the $O(n^2 \log n)$ bound of the algorithms of [13], [3] for the discrete $p$-center location problem, i.e., the case when $D = N$, to obtain an $O(n^2)$ time algorithm.

**The continuous $p$-center problem.** In this section we consider the problem of locating $p$ facilities on a tree network in order to minimize the maximum of the distances of the points on the network to their respective nearest facility. Using the notation presented above, we want to find $r(p)$ such that

$$r(p) = \min \{f_T(U_p) : U_p \subseteq T, |U_p| = p\}, \tag{4}$$

and also the locations for facilities that achieve this value.

Given a point $x$ on $T$ and $r > 0$, we define $N_r(x)$, the $r$-neighborhood of $x$, by $N_r(x) = \{y \in T : d(x, y) \leqq r\}$. The location problem is then to find the minimum $r$ such that $p$ $r$-neighborhoods will cover the entire $T$. Similarly, given $r > 0$, we consider the reverse problem of covering the tree with a minimum number of $r$-neighborhoods. This number is denoted by $M(r)$. It is clear that $M(r)$ is a monotone, nonincreasing, step function, which is continuous from the right. $r(p)$ is, therefore, the smallest $r$ such that $M(r) \leqq p$.

The algorithm of [2] for finding $r(p)$ is based on an $O(n)$ subroutine for finding $M(r)$ for an arbitrary $r > 0$. ($n$ is the number of nodes in $T$.)

In this section we show that $r(p) = d(i, j)/2k$, where $d(i, j)$ is the distance between some pair of tips, $i$ and $j$, of $T$, and $k$ is an integer satisfying $1 \leqq k \leqq p$. (A tip is a node of degree 1.) The latter property combined with the monotonicity of $M(r)$ will imply that the $O(n)$ routine for finding $M(r)$ is to be applied at most $O(n^2 p)$ times, before $r(p)$ is found.

To prove our claim on $r(p)$ we will need the algorithm of [2] for finding $M(r)$. Thus, for the sake of completeness we describe it here as well.

ALGORITHM 1. Suppose that the tree is rooted at some node and arranged in levels. Define the level of a node as the number of arcs in the unique path connecting the node with the root. Node $i$ is a son of node $j$ if $j$ is the immediate predecessor of $i$

on the path connecting $i$ with the root. We also say that $j$ is the father of $i$. Consider a maximal set of tips having the same father, say node $s$. If all sons of $s$ are tips we call such a set a cluster, and denote it by $C(s)$.

The algorithm will successively eliminate clusters from the tree, where at each iteration it will find the minimum number of supply centers, ($r$-neighborhoods), required to cover the cluster under consideration.

We start by motivating the first step of the Cluster Elimination Routine. If the length of any arc $(s, i)$—$i$ being a tip—is greater than $2r$, a facility must be located on $(s, i)$. Without loss of generality, that facility can be established at a point on $(s, i)$ whose distance from the tip $i$ is $r$. (Note that this facility covers only points on $(s, i)$.) One can then reduce the length of the arc by $2r$.

*The Cluster Elimination Routine.*

*Step 0.* Choose a cluster, $C(s)$, of the initial tree, (possibly one of the highest level).

*Step 1.* Let $\{(s, i)\}$, $i \in C(s)$, be the set of arcs connecting the tips to their predecessor $s$.

For each $i$ let $d(s, i) = k_i(2r) + b_i$, where $k_i$ is a nonnegative integer and $0 < b_i \leqq 2r$.

$$\text{Set } d(s, i) \leftarrow b_i \quad \text{for } i \in C(s).$$

(At this point $k_i$ facilities have already been established on arc $(s, i)$, with the distance between two adjacent facilities being $2r$. Also note that the trimmed arcs have positive lengths.)

*Step 2.* Let $$\alpha = \min_{i \in C(s)} \{d(s, i) : d(s, i) > r\} = d(s, i_1^*),$$

and

$$\beta = \max_{i \in C(s)} \{d(s, i) : d(s, i) \leqq r\} = d(s, i_2^*).$$

In case of a tie $i_1^*$ ($i_2^*$) can be chosen as the smallest index for which the minimum (maximum) is attained. Also, if $\alpha(\beta)$ is defined on an empty set it is set equal to $+\infty(-\infty)$. (Note that at least one of $\alpha$, $\beta$ is finite.)

(i) If $\alpha + \beta > 2r$, then for each $i$ such that $d(s, i) > r$, locate a facility on $(s, i)$ at a distance $r$ from the tip $i$ (of the reduced cluster obtained in Step 1). Remove each arc $(s, i)$ in $C(s)$ except $(s, i_2^*)$.

If $s$ is the root of the tree, locate a facility at $s$ and terminate. Otherwise remove node $s$ so that we have the case shown in Fig. 1, and go to Step 3.

(ii) If $\alpha + \beta \leqq 2r$, then for each $i \neq i_1^*$ with $d(s, i) > r$, locate a facility on $(s, i)$ at a distance $r$ from the tip $i$. Remove all the arcs $(s, i)$ except $(s, i_1^*)$.

If $s$ is the root of the tree, locate a facility on $(s, i_1^*)$ at a distance $r$ from $i_1^*$ and terminate. Otherwise, remove node $s$ as shown in Fig. 1, and go to Step 3.

*Step 3.* Choose a cluster of the remaining tree (possibly one of the highest level), and return to Step 1.

It is clear that the above algorithm takes $O(\max(n, M(r)))$ time, if the output is to be the $M(r)$ facility locations. However, the following method of recording the output reduces the time bound to $O(n)$. On an arc, if there are $k$ facilities to be located at a distance $2r$ from each other, the location of only the first one and their number may be output.

THEOREM 2. *Let $r(p)$ be the solution to the continuous p-center problem, i.e. $r(p)$ is defined by* (4). *Then $r(p) = d(\mathrm{i}, \mathrm{j})/2k$, where $d(i, j)$ is a distance between a pair of tips, $i$ and $j$, of the tree $T$, and $k$ is an integer, $1 \leqq k \leqq p$.*

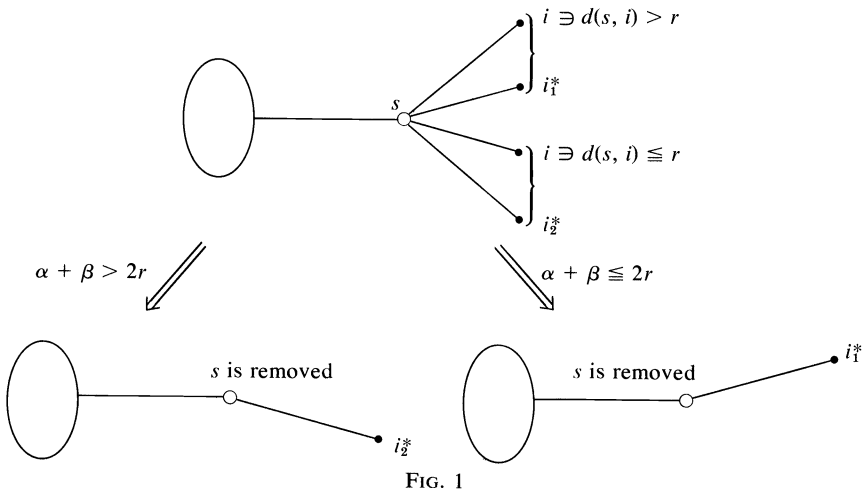



FIG. 1

*Proof.* Let $S = \{x_1, \cdots, x_p\}$ be the set of points on $T$ at which the $p$ optimal supply centers are located. Define $D = \{y : y \in T, \min_{1 \leqq i \leqq p} d(y, x_i) = r(p)\}$, and let $S'$ be the subset of supply points serving the members of $D$, i.e.,

$$S' = \{x : x \in S, d(x, y) = r(p) \text{ for some } y \in D\}.$$

First we claim that without loss of generality it can be assumed that each member of $S'$ is the midpoint of a simple path of length $2r(p)$, connecting two points of $D$. Suppose that $x \in S'$ does not have the above property. Then, the supply center there can be slightly perturbed to $x'$ such that the optimality is not affected; all points $y$ in $D$ served by $x$ satisfy $d(x', y) < r(p)$, and no additional points are added to $D$. Therefore, $x$ can be omitted from $S'$ and all points $y$ in $D$ served by $x$ can be omitted from $D$. Note that the minimality of $r(p)$ ensures that the set $S'$ remaining after this process is not empty.

To complete the proof of the theorem we show that each member of $D$ which is not a tip of $T$ must be the midpoint of a simple path of length $2r(p)$, connecting two points of $S'$.

Let $y \in D$. Then there exists $x_i \in S'$ with $d(y, x_i) = r(p)$. If $y$ is not a tip there exists $z \neq y$, $z \in T$, and $y$ is on the simple path between $z$ and $x_i$. Considering only the subpath connecting $z$ and $y$, we observe that all points on this subpath but $y$ are not served by $x_i$, since they are at a distance greater than $r(p)$ from $x_i$. So, let $x_k$ be the point in $S$, closest to $y$, and serving at least one point which is not $y$, on the above subpath. Clearly $d(y, x_k) = r(p)$, since $y$ is in $D$, and therefore $x_k$ is in $S'$.

Moreover, since $d(x_k, u) \leqq r(p)$ for some $u \neq y$ on that subpath, $y$ is the only intersection point of the path connecting $y$ and $x_k$ and the path connecting $y$ and $x_i$. Hence $y$ is on the simple path between $x_i$ and $x_k$ with $d(y, x_i) = d(y, x_k) = r(p)$.

Using the above properties satisfied by the members of $D$ and $S'$, we start with $x$ in $S'$ and consider the path of length $2r(p)$, which connects two points of $D$ and has $x$ as its midpoint. If at least one of these endpoints is not a tip, the path can be extended by $2r(p)$ such that the new path will still connect two members of D. Continuing this process, the no-cycle property of a tree ensures that we find a simple path of the tree connecting two tips and having total length of $2kr(p)$, $1 \leqq k \leqq p$. This completes the proof.

The above theorem implies that $r(p)$, the solution to the $p$-center problem, can be found by applying Algorithm 1 $O(n^2 p)$ times, thus yielding an $O(n^3 p)$ bound for

solving the continuous $p$-center problem. Next we show a reduction of this bound which is based on the nature of the $O(n^2p)$ possible values for $r(p)$.

Due to the monotonicity property of $M(r)$, found by Algorithm 1, it is clear that if $M(\bar{r}) \leqq p$ then $r(p) \leqq \bar{r}$, and one can ignore all values of $r$ greater than $\bar{r}$. Similarly, if $M(\bar{r}) > p$ we have $r(p) > \bar{r}$. Let $R$ be the set of possible values for $r(p)$ as specified by Theorem 2. We start by finding the median of $R$, say $r_1$, and then applying Algorithm 1 to find $M(r_1)$. Comparing $M(r_1)$ and $p$ we then eliminate half of the members of $R$ from further consideration, leaving the subset $R_1$. We then continue by finding the median of $R_1$, say $r_2$, computing $M(r_2)$, and so on. Let $r_i$ denote the median found at the $i$th iteration and let $R_i$ be the respective subset of $R$ that we are left with at this iteration. Next we show that the total effort of evaluating the sequence of medians $\{r_1, r_2, \cdots\}$ is $O(n^2 \log^2 p)$.

First, an effort of $O(n^2)$ yields the distances between all tips of $T$. For each such distance $d(i, j)$ the sequence $\{d(i, j)/2k\}$, $k = 1, \cdots, p$, is a monotone decreasing sequence. One can then apply the methods of [7], [12] to find $r_1$ in $O(n^2 \log p)$ time. Applying Algorithm 1 to $r_1$ (for $O(n)$ time), we can then use a binary search on each one of the sequences $\{d(i, j)/2k\}$, $k = 1, \cdots, p$, to find $R_1$. Since there are $n^2$ sequences, this effort amounts to $O(n^2 \log p)$. In general, at the $i$th iteration, two pointers are sufficient to limit that part of a sequence $\{d(i, j)/2k\}$, $k = 1, \cdots, p$ which is contained in $R_i$. Hence the storage requirement is of order $O(n^2)$. Successive applications of the methods of [7], [12] for $q = \log p$ times will yield $r_1, r_2, \cdots, r_q$. By that time the remaining set of possible values, $R_q$, will contain $O(n^2)$ elements. Therefore, the remaining medians in the sequence are found in total effort of $O(n^2)$ using the linear time algorithm of [1]. Thus, we have demonstrated that the total effort of our procedure to find $r(p)$ is of order $O(n^2 \log^2 p)$ with $O(n^2)$ storage.

Finally, using the duality result presented in the Introduction we observe that the optimal objective value of the $p$-center dispersion problem is also found in $O(n^2 \log^2 p)$ time. To find the locations of the $p$ centers achieving this optimal value, one can use the procedure given in [2]. As shown in [5] this procedure can be implemented in $O(n^2)$ time.

*Remarks.*

1) There are certain circumstances where the bound $O(n^2 \log^2 p)$ given above can be improved if a different method is used to find the sequence of medians. We mention two such procedures. The first one is based on the observation that the median of the set $R$ is also the median of the set $R^{-1}$, consisting of the reciprocals of $R$. But then the sequence $\{2k/d(i, j)\}$ $k = 1, \cdots, p$, is a linear sequence. It is shown in [4] how to find a median of set consisting of $n^2$ linear sequences in $O(n^2 \log n)$ time. Applying the latter procedure to compute $\{r_1, r_2, \cdots\}$ yields the bound $O(n^2 \log n \log p)$ for the algorithm to find $r(p)$.

For the second procedure we first sort the sequence of the $m = O(n^2)$ distances between the tips. Denoting this sorted sequence by $c_1 \geqq c_2 \cdots \geqq c_m$, we represent $R$ as the union of $p$ monotone sequences. For each $k = 1, \cdots, p$ we consider the sequence $\{c_i/2k\}$, $i = 1, \cdots, m$. Applying the methods of [7], [12] to this structure yields the bound $O(n^2 \log n + p \log n \log p)$ for the total effort to find $r(p)$.

2) The discrete $p$-center problem, i.e. the model where demand occurs only at the nodes of $T$, is solved in [13], [3] by an $O(n^2 \log n)$ algorithm. We indicate that this bound can be reduced to $O(n^2)$ for the method in [13]. The set $R$ of possible values for $r(p)$ for the discrete problem is known to contain $O(n^2)$ elements. All these elements are computed in $O(n^2)$ total effort. Then, for each given $r$, an $O(n)$ routine finding $M(r)$, the minimum number of $r$-neighborhoods covering all nodes, is given.

As was done above for the continuous $p$-center problem, one can generate the sequence of medians $\{r_1, r_2, \cdots\}$ and apply the procedure to find $M(r)$ a total of $O(\log(n^2)) = O(\log n)$ times. Since each time the cardinality of the remaining set $R_i$ is cut by half, the linear time algorithm of [1] will generate the entire sequence of medians in total effort of $O(n^2)$. This latter term is then the dominating term yielding the bound $O(n^2)$ for the effort to find $r(p)$.

## REFERENCES


[1] M. BLUM, R. W. FLOYD, V. R. PRATT, R. L. RIVEST, AND R. E. TARJAN, *Time bounds for selection,* J. Comput. System Sci., 7 (1973), pp. 448–461.

[2] R. CHANDRASEKARAN AND A. DAUGHETY, *Problems of location on trees,* Discussion Paper No. 357, Center for Mathematical Studies in Economics and Management, Northwestern University, 1978.

[3] R. CHANDRASEKARAN AND A. TAMIR, *Polynomially bounded algorithms for locating p-centers on a tree,* Discussion Paper No. 358, Center for Mathematical Studies in Economics and Management, Northwestern University, 1978.

[4] ———, *Optimizing over nested linear constraints,* in preparation.

[5] ———, *Locating obnoxious facilities,* Manuscript, Department of Statistics, Tel Aviv University, 1979.

[6] P. M. DEARING AND R. L., FRANCIS, *A minimax location problem on a network,* Transportation Sci., 8 (1974), pp. 333–343.

[7] Z. GALIL AND N. MEGIDDO, *A fast selection algorithm and the problem of optimum distribution of effort,* J. Assoc. Comput. Mach., 26 (1979), pp. 58–64.

[8] A. J. GOLDMAN, *Minimax location of a facility in a network,* Transportation Sci., 6 (1972), pp. 407–418.

[9] S. L. HAKIMI, E. F. SCHMEICHEL, AND J. G. PIERCE, *On p-centers in networks,* Transportation Sci., 12 (1978), pp. 1–15.

[10] G. Y. HANDLER, *Minimax location of a facility in an undirected tree graph,* Transportation Sci., 7 (1973), pp. 287–293.

[11] ———, *Finding two-centers of a tree: The continuous case,* Transportation Sci., 12 (1978), pp. 93–106.

[12] D. B. JOHNSON AND T. MIZOGUCHI, *Selecting the Kth element in $X + Y$ and $X_1 + X_2 + \cdots + X_m$,* SIAM J. Comput., 7 (1978), pp. 147–153.

[13] O. KARIV AND S. L. HAKIMI, *An algorithmic approach to network location problems. Part 1: the p-centers,* SIAM J. Appl. Math., 37 (1979), pp. 513–538.

[14] D. R. SHIER, *A min-max theorem for p-center problems on a tree,* Transportation Sci., 11 (1977), pp. 243–252.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980

0196-5212/80/0104-0003 $01.00/0

# THE ERDÖS-KO-RADO THEOREM FOR INTEGER SEQUENCES*

PETER FRANKL† AND ZOLTÁN FÜREDI‡

**Abstract.** For positive integers $n, k, t$ we investigate the problem how many integer sequences $(a_1, a_2, \cdots, a_n)$ we can take, such that $1 \leqq a_i \leqq k$ for $1 \leqq i \leqq n$, and any two sequences agree in at least $t$ positions. This problem was solved by Kleitman (J. Combin. Theory, 1 (1966), pp. 209–214) for $k = 2$, and by Berge (in *Hypergraph Seminar, Columbus, Ohio* (1972), Springer-Verlag, New York, 1974) for $t = 1$. We prove that for $t \geqq 15$ the maximum number of such sequences is $k^{n-t}$ if and only if $k \geqq t+1$.

**1. Introduction.** Let $t, k, n$ be positive integers with $k \geqq 2$, $n \geqq t$, and let $\mathcal{A}$ be a set of integer sequences $(a_1, a_2, \cdots, a_n)$, $1 \leqq a_i \leqq k$. We say that $(a_1, a_2, \cdots, a_n)$ and $(a_1', a_2', \cdots, a_n')$ intersect in at least $t$ positions if we can find $1 \leqq i_1 < i_2 < \cdots < i_t \leqq n$ such that $a_{i_j} = a_{i_j}'$ for $j = 1, \cdots, t$.

Let $f(n, k, t)$ denote the maximum cardinality $\mathcal{A}$ can have supposing that any two elements of $\mathcal{A}$ intersect in at least $t$ positions. Setting $\mathcal{A}_0 = \{(a_1, \cdots, a_n) | 1 \leqq a_i \leqq k, a_j = 1 \text{ for } j = 1, \cdots, t\}$, we obtain

$$f(n, k, t) \geqq k^{n-t}. \tag{1}$$

In the case $k = 2$ the problem reduces to the following: What is the maximum number of subsets of an $n$-set such that the symmetric difference of any two has cardinality at most $n - t$? This problem was posed by Erdös and solved by Kleitman [5], who proved that

$$f(n, 2, t) = \begin{cases} \sum_{i=0}^{[(n-t)/2]} \binom{n}{i} & \text{if } n-t \text{ is even,} \\ 2\sum_{i=0}^{[(n-t)/2]} \binom{n-1}{i} & \text{if } n-t \text{ is odd.} \end{cases} \tag{2}$$

The expression (2) is much greater than (1) except for $t = 1$, when we have equality. Berge [1] proved that

$$f(n, k, 1) = k^{n-1} \tag{3}$$

holds for $k \geqq 3$ as well. In fact he proved that if instead of $a_i \leqq k$ we suppose $a_i \leqq k_i$, $k_1 \leqq \cdots \leqq k_n$, then the corresponding bound is $k_2 k_3 \cdots k_n$. Livingston [7] proved that if equality holds in (3) then necessarily $\mathcal{A}$ is of the form $\mathcal{A}_0$ (up to isomorphism). In the present paper we are mainly concerned with the problem, for which triples $n, k, t$ is the bound (1) optimal. We have the following

CONJECTURE. *The bound* (1) *is optimal if and only if* $n \leqq t+1$ *or* $k \geqq t+1$.

*Remark.* It is easy to check that the conjecture holds for $n \leqq t+1$, i.e., $n = t$ and $n = t+1$. On the other hand, (2) and (3) settle it for $t = 1$.

THEOREM 1. *The conjecture holds for* $t \geqq 15$.

We give some results for the range $2 \leqq t \leqq 14$ as well.

**2. Preliminaries.** Our main tool in proving Theorem 1 will be the strongest form of the Erdös-Ko-Rado theorem (see [2]), proved in Frankl [3]. To state it we need some

* Received by the editors December 3, 1979, and in final form March 24, 1980.
† Centre National de la Recherche Scientifique, Paris, France.
‡ Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

definitions. Let $s$ be an integer $t \leqq s \leqq n$. Let $\mathcal{B}$ be a family of $s$-element subsets of $\{1, 2, \cdots, n\}$ satisfying $|B \cap B'| \geqq t$ for $B, B' \in \mathcal{B}$. Such a family is called $t$-intersecting. Let us define

$$\mathcal{B}_0 = \{B \subseteq \{1, 2, \cdots, n\} \mid \{1, 2, \cdots, t\} \subseteq B, |B| = s\}$$

$$\mathcal{B}_1 = \{B \subseteq \{1, 2, \cdots, n\} \mid |\{1, 2, \cdots, t+2\} \cap B| \geqq t+1, |B| = s\}.$$

Clearly both $\mathcal{B}_0$ and $\mathcal{B}_1$ are $t$-intersecting. Then we have

THEOREM 2[3]. *There exist positive constants $c_t$, depending on $t$ only and satisfying $c_t < 2$ for $t \geqq 2$, and $c_t < 1$ for $t \geqq 15$ such that for*

$$\frac{n-t}{s-t} > c_t(t+1), \tag{4}$$

*a $t$-intersecting family of maximum size is of the form $\mathcal{B}_0$ or $\mathcal{B}_1$, (up to isomorphism).*

As remarked in [3],

$$|\mathcal{B}_0| \gtrless |\mathcal{B}_1| \quad \text{iff} \quad n \gtrless (s-t+1)(t+1). \tag{5}$$

Let us now return to our set of sequences $\mathcal{A}$, which is $t$-intersecting; i.e., any two sequences in $\mathcal{A}$ intersect in at least $t$ positions. Let us define:

$$\mathcal{A}^+ = \{(a_1, a_2, \cdots, a_{n+1}) \mid (a_1, \cdots, a_n) \in \mathcal{A}, 1 \leqq a_{n+1} \leqq k\}.$$

It is evident that $\mathcal{A}^+$ is $t$-intersecting, yielding

$$f(n+1, k, t) \geqq k f(n, k, t). \tag{6}$$

Consequently the function $f(n, k, t)k^{-n}$ is nondecreasing in $n$ (and bounded by 1). Hence the following limit exists (and is at most 1):

$$g(k, t) = \lim_{n \to \infty} f(n, k)k^{-n}.$$

We will now bring $\mathcal{A}$ to a canonical form. Such a transformation was first used by Kleitman [6]. Let $1 \leqq j \leqq n$. Define the transformation,

$$T_{j,i}(a_1, \cdots, a_j \cdots, a_n) = \begin{cases} (a_1, a_2, \cdots, a_j', \cdots, a_n) & \text{if this sequence is not in } \mathcal{A}, \text{ and} \\ & \quad a_j = i,\ a_j' = 1; \\ (a_1, a_2, \cdots, a_j, \cdots, a_n) & \text{otherwise.} \end{cases}$$

It is easily seen that $T_{j,i}(\mathcal{A}) = \{T_{j,i}(A) \mid A \in \mathcal{A}\}$ is $t$-intersecting and has the same cardinality as $\mathcal{A}$. Repeated application of the transformation yields a system $\mathcal{A}'$ which is $t$-intersecting, $|\mathcal{A}'| = |\mathcal{A}|$, and for $1 \leqq j \leqq n$, $1 \leqq i \leqq k$,

$$T_{j,i}(\mathcal{A}') = \mathcal{A}'. \tag{7}$$

Without loss of generality we may assume $\mathcal{A} = \mathcal{A}'$. Let us associate with every $(a_1, \cdots, a_n) = A$, the set $B(A) = \{i \mid a_i = 1\}$.

PROPOSITION 1. *The family $\mathcal{B} = \{B(A) \mid A \in \mathcal{A}\}$ is $t$-intersecting.*

*Proof.* Let $A = (a_1, \cdots, a_n), A' = (a_1', \cdots, a_n') \in \mathcal{A}$. Let $\{i_1, i_2, \cdots, i_r\}$ be the set of $i$'s such that $a_i = a_i' \neq 1$. In view of (7), $A'' = (a_1'', \cdots, a_n'') \in \mathcal{A}$, where $a_i'' = a_i'$ for $i \notin \{i_1, \cdots, i_r\}$, $a_i'' = 1$ for $i \in \{i_1, \cdots, i_r\}$. As $(a_1, \cdots, a_n)$ and $(a_1'', \cdots, a_n'')$ agree in the $i$th position only for $i \in B(A) \cap B(A')$, the statement of the proposition follows. Now by the maximality of $\mathcal{A}$ we have

PROPOSITION 2. $\mathcal{A} = \{A = (a_1, \cdots, a_n) \mid 1 \leqq a_i \leqq k, B(A) \in \mathcal{B}\}$, *and consequently*

$$|\mathcal{A}| = \sum_{B \in \mathcal{B}} (k-1)^{n-|B|}. \tag{8}$$

Hence the problem of determining $f(n, k, t)$ reduces to finding the maximum of (8) over all $t$-intersecting families $\mathcal{B}$. We need an easy probabilistic result.

PROPOSITION 3. *For every positive $\varepsilon$ and $\delta$ the number of sequences $(a_1, \cdots, a_n)$ with $1 \leqq a_i \leqq k$ which contain more than $(1+\varepsilon)(n/k)$ 1's or less than $(1-\varepsilon)(n/k)$ 1's is less than $\delta k^n$ for $n > n_0(\delta, \varepsilon)$.*

Instead of a proof, just observe that $p(a_i = 1) = 1/k$; hence the mean value of 1's is $n/k$, and the events $a_i = 1$ are independent for $i = 1, \cdots, n$.

**3. The main results.** We prove Theorem 1 in a somewhat surprising way; namely we prove first that it holds asymptotically, i.e., $f(n, k, t) \leqq (1+o(1))k^{n-t}$ for $k, t$ fixed, $k > t \geqq 15$. Then we deduce $f(n, k, t) = k^{n-t}$ from it for every $n \geqq t$.

THEOREM 3. *For $k > t \geqq 15$ we have*

$$g(k, t) = \lim_{n\to\infty} f(n, k, t)k^{-n} = k^{-t}.$$

In view of Proposition 2,

$$f(n, k, t)k^{-n} = \left(\sum_{B\in\mathcal{B}} (k-1)^{n-|B|}\right)k^{-n},$$

for some $t$-intersecting family $\mathcal{B}$. Moreover, Proposition 3 gives that for any $\delta, \varepsilon > 0$, $n > n_0(\delta, \varepsilon)$, we have

$$f(n, k, t)k^{-n} < \left(\sum_{B} (k-1)^{n-|B|}\right)k^{-n} + \delta, \tag{9}$$

where $B$ runs over those elements of $\mathcal{B}$ which satisfy

$$(1-\varepsilon)(n/k) \leqq |B| \leqq (1+\varepsilon)(n/k).$$

Now for $(1-\varepsilon)(n/k) \leqq s \leqq (1+\varepsilon)(n/k)$, set

$$\mathcal{B}(s) = \{B \in \mathcal{B} \,|\, |B| = s\}.$$

As $k \geqq t+1$, for $n > n_0(\varepsilon)$ we have $(n-t)/(s-t) > c_t(t+1)$; i.e., (4) is satisfied and we may apply Theorem 2 to the $t$-intersecting family $\mathcal{B}(s)$. We deduce

$$|\mathcal{B}(s)| \leqq \max(|\mathcal{B}_0|, |\mathcal{B}_1|) = \max\left(\binom{n-t}{s-t}, (t+2)\left(\binom{n-t-2}{s-t-1} + \binom{n-t-2}{s-t-1}\right)\right). \tag{10}$$

By (5) for $k > t+1$ the value of (10) is $\binom{n-t}{s-t}$.

If $k = t+1$, then

$$\frac{1-\varepsilon}{t+1} < \frac{s}{n} < \frac{1+\varepsilon}{t+1}.$$

We can still obtain for $n > n_0(t, \varepsilon)$,

$$\binom{n-t-2}{s-t-1} \Big/ \binom{n-t}{s-t} = \frac{(s-t)\cdot(n-s)}{(n-t)\cdot(n-t+1)} < \frac{(t+\varepsilon)(1+\varepsilon)}{t^2+2t+1-\varepsilon},$$

and

$$\binom{n-t-s}{s-t-2} \Big/ \binom{n-t}{s-t} = \frac{s-t}{n-t}\cdot\frac{s-t-1}{n-t-1} < \frac{(1+\varepsilon)^2}{t^2+2t+1},$$

yielding

$$|\mathcal{B}(s)| \leqq |\mathcal{B}_0| \frac{(t+2)(t+\varepsilon)(1+\varepsilon)+(1+\varepsilon)^2}{t^2+2t+1-\varepsilon} < \binom{n-t}{s-t}(1+2\varepsilon), \tag{11}$$

whenever $\varepsilon$ is sufficiently small. Now from (9) and (11) we obtain

$$\begin{aligned} f(n,k,t)k^{-n} &< (1+2\varepsilon)\sum_{s=t}^{n}\binom{n-t}{s-t}(k-1)^{n-s}k^{-n}+\delta \\ &= (1+2\varepsilon)k^{-n}\sum_{j=0}^{n-t}\binom{n-t}{j}(k-1)^{j}+\delta \\ &= (1+2\varepsilon)k^{-n}k^{n-t}+\delta \\ &= (1+2\varepsilon)k^{-t}+\delta, \end{aligned}$$

which implies, since $\delta, \varepsilon > 0$ were arbitrary,

$$g(k,t) \leqq k^{-t}.$$

As $|\mathcal{A}_0| = k^{n-t}$ we have $g(k,t) \geqq k^{-t}$ as well, which concludes the proof of Theorem 3.

*Proof of Theorem* 1. Suppose that for some $t$-intersecting family $\mathcal{A}$ we have $|\mathcal{A}| \geqq k^{n-t}+1$. Then using (6) we deduce

$$f(n',k,t) \geqq k^{n'-n}f(n,k,t) \geqq k^{n'-n}|\mathcal{A}| \geqq k^{n'}(k^{-t}+k^{-n}),$$

whence $g(k,t) \geqq k^{-t}+k^{-n} > k^{-t}$, a contradiction (observe that now $n$ is fixed and we have $n' \to \infty$), which proves the *if* part of Theorem 1.

For the *only if* part, let us define

$$\mathcal{A}_1 = \{A = (a_1, \cdots, a_n) \mid 1 \leqq a_i \leqq k, |B(\mathcal{A}) \cap \{1, \cdots, t+2\}| \geqq t+1\}.$$

Obviously $\mathcal{A}_1$ is $t$-intersecting, and we have for $n \geqq t+2$, $k \leqq t$,

$$|\mathcal{A}_1| = k^{n-t-2}((t+2)(k-1)+1) = k^{n-t}(1+(t+1-k)k^{-2}) > k^{n-t}.$$

**4. Some remarks.** Using the same argument we could deduce

THEOREM 4. *If* $t \geqq k > c_t(t+1)$, *then*

$$g(k,t) = k^{-t-2}((t+1)(k-1)+1).$$

(By [3] we know that $c_t < 0.8$ for $t \geqq 15$.) Now Theorem 4 yields

THEOREM 5. *If* $t \geqq k > c_t(t+1)$, *then*

$$f(n,k,t) = k^{n-t-2}((t+1)(k-1)+1).$$

**5. Probabilistic arguments.** Now we want to apply the random walk method developed in [4] to obtain a general bound on $g(k,t)$, $k > 2$.

Let $\mathcal{B}$ be the $t$-intersecting family associated with the maximal set of $t$-intersecting sequences $\mathcal{A}$. With $\mathcal{B}$ we proceed as in [3]. For $1 \leqq i < j \leqq n$, the canonical transformation is the following.

$$K_{i,j}(B) = \begin{cases} B' = B - \{j\} \cup \{i\} & \text{if } i \notin B, j \in B, B' \notin \mathcal{B}, \\ B & \text{otherwise;} \end{cases}$$

$$K_{i,j}(\mathcal{B}) = \{K_{i,j}(B) \mid B \in \mathcal{B}\}.$$

Applying $K_{i,j}$ repeatedly we obtain a $t$-intersecting family $\mathcal{B}'$ which satisfies $K_{i,j}(\mathcal{B}') = \mathcal{B}'$ for all $1 \leqq i < j \leqq n$. We may suppose $\mathcal{B} = \mathcal{B}'$. The following propositions are taken essentially from [3].

PROPOSITION 4. *No subset $S$ of $T = \{1, 2, \cdots, t-1, t+1, \cdots, t+2l+1, \cdots\}$ belongs to $\mathcal{B}$.*

*Proof.* Otherwise an application of $K_{t+2l,t+2l+1}$ for all $l \geqq 0$ would yield $S' \in \mathcal{B}$ for $S' \subseteq T' = \{1, 2, \cdots, t, t+2, \cdots, t+2l, \cdots\}$. But $|S \cap S'| \leqq |T \cap T'| = t-1$, a contradiction.

Let us associate with a sequence $A = (a_1, \cdots, a_n)$ a random walk in the plane in the following way. We start from $(0, 0)$. Suppose that after $(i-1)$ moves we are in $(x, y)$. Then we move to $(x, y+1)$ or $(x+1, y)$ according to whether $a_i = 1$ or not. The random walks associated with different sets are different. Proposition 4 yields (see [3])

PROPOSITION 5. *The random walk associated with $A \in \mathcal{A}$ hits the line $y = x + t$.*

In probability language, considering the space of all possible sequences $(a_1, \cdots, a_n)$, $1 \leqq a_i \leqq k$, we move upward with probability $1/k$ and to the right with probability $(k-1)/k$. Now let us continue to walk indefinitely. Then for the probability of hitting the line $y = x + t$, $p(t)$ we obtain

$$p(0) = 1,$$

$$p(t) = (1/k)p(t-1) + ((k-1)/k)p(t+1) \quad \text{for } t \geqq 1,$$

and

$$p(t) \to 0 \quad \text{as } t \to \infty, \text{ because } k > 2.$$

Hence, we deduce

$$p(t) = (k-1)^{-t}.$$

Consequently we have

THEOREM 6. *For $k \geqq 3$ we have $k^{-n} f(n, k, t) \leqq (k-1)^{-t}$, and consequently*

$$g(k, t) \leqq (k-1)^{-t}. \tag{12}$$

From (2) it follows that $g(2, t) = 2^{-1}$ for every $t \geqq 1$, which is a great contrast to (12).

On the other hand, for $k$, $t$ fixed, let $(s, s+t)$ be the point of the line $y = x + t$ for which the probability that a random walk goes through it is the largest. Let $\mathcal{A}_s$ be the set of the corresponding sequences. Then obviously

$$\mathcal{A}_s = \{A = (a_1, \cdots, a_n) \,|\, |B(A) \cap \{1, 2, \cdots, t+2s\}| \geqq t+s\}. \tag{13}$$

Thus

$$g(k, t) \geqq \frac{|A_s|}{k^n}.$$

Then elementary computation shows that for some constant $d_k$ depending on $k$ only, we have

$$g(k, t) > \frac{(k-1)^{-t}}{d_k \sqrt{t}}.$$

Let us finish with a conjecture, setting the general case.

CONJECTURE. *Let $\mathcal{A}_s$ be defined by* (13). *Then*

$$f(n, k, t) = \max_{s \geqq 0} |\mathcal{A}_s|.$$

## REFERENCES

[1] C. BERGE, *Nombres de coloration de l'hypergraphe h-parti complet*, Hypergraph seminar, Columbus, Ohio, 1972, Springer-Verlag, New York, 1974, pp. 13–20.

[2] P. ERDÖS, C. KO AND R. RADO, *Intersection theorems for systems of finite sets*, Quart. J. Math., Oxford Ser., 12 (1961), pp. 313–320.

[3] P. FRANKL, *The Erdös-Ko-Rado theorem is true for $n = ckt$*, Proc. Fifth Hung. Comb. Coll. Keszthely, 1976, North-Holland, Amsterdam, 1978, pp. 365–375.

[4] ———, *Families of finite sets satisfying an intersection condition*, Bull. Austral. Math. Soc., 15 (1976), pp. 73–79.

[5] D. J. KLEITMAN, *On a combinatorial conjecture of Erdös*, J. Combin. Theory, 1 (1966), pp. 209–214.

[6] ———, *Families of non-disjoint subsets*, J. Combin. Theory, 1 (1966), pp. 153–155.

[7] M. L. LIVINGSTON, *An ordered version of the Erdös-Ko-Rado theorem*, J. Combin. Theory, Ser. A, 26 (1979), pp. 162–165.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104-0004 $01.00/0

# ON ADDITIVE BASES AND HARMONIOUS GRAPHS*

R. L. GRAHAM† AND N. J. A. SLOANE†

**Abstract.** This paper first considers several types of *additive bases.* A typical problem is to find $n_\gamma(k)$, the largest $n$ for which there exists a set $\{0 = a_1 < a_2 < \cdots < a_k\}$ of distinct integers modulo $n$ such that each $r$ in the range $0 \leqq r \leqq n-1$ can be written *at least* once as $r \equiv a_i + a_j$ (modulo $n$) with $i < j$. For example, $n_\gamma(8) = 24$, as illustrated by the set $\{0, 1, 2, 4, 8, 13, 18, 22\}$. The other problems arise if *at least* is changed to *at most*, or $i < j$ to $i \leqq j$, or if the words modulo $n$ are omitted. Tables and bounds are given for each of these problems. Then a closely related graph labeling problem is studied. A connected graph with $n$ edges is called *harmonious* if it is possible to label the vertices with distinct numbers (modulo $n$) in such a way that the edge sums are also distinct (modulo $n$). Some infinite families of graphs (odd cycles, ladders, wheels, $\cdots$) are shown to be harmonious while others (even cycles, most complete or complete bipartite graphs, $\cdots$) are not. In fact most graphs are not harmonious. The function $n_\gamma(k)$ is the size of the largest harmonious subgraph of the complete graph on $k$ vertices.

**1. Additive bases.** This paper is mostly concerned with *modular* versions of certain additive bases for the integers $\{1, 2, \cdots, n\}$, and with a closely related graph labeling problem, that of determining which graphs are harmonious.

Although our primary interest is in just two of these function ($n_\gamma$ and $v_\gamma$), it is most convenient to begin by defining eight closely related functions. Our notation is that $[1, n] := \{1, 2, \cdots, n\}$, $\mathbb{Z}_n$ denotes the integers modulo $n$, and $k \geqq 2$ is a natural number. The first four functions are concerned with *covering* $[1, n]$ or $\mathbb{Z}_n$ with sums.

- $n_\alpha(k)$ (resp. $n_\beta(k)$) is the *largest* number $n$ such that there exists a $k$-element set $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ of integers with the property that each $r \in [1, n]$ can be written in *at least* one way as $r = a_i + a_j$, with $i < j$ (resp $i \leqq j$).
- $n_\gamma(k)$ (resp. $n_\delta(k)$) is the *largest* number $n$ such that there exists a subset $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ of $\mathbb{Z}_n$ with the property that each $r \in \mathbb{Z}_n$ can be written in *at least* one way as $r = a_i + a_j$ with $i < j$ (resp. $i \leqq j$).

Since this does not assign a value to $n_\gamma(2)$ we define $n_\gamma(2) = 1$. The other four functions are concerned with *packing* $[0, v]$ or $\mathbb{Z}_v$ with sums.

- $v_\alpha(k)$ (resp. $v_\beta(k)$) is the *smallest* number $v$ such that there exists a $k$-element set $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ of integers with the property that the sums $a_i + a_j$ for $i < j$ (resp. $i \leqq j$) belong to $[0, v]$ and represent each element of $[0, v]$ *at most* once.
- $v_\gamma(k)$ (resp. $v_\delta(k)$) is the *smallest* number $v$ such that there exists a subset $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ of $\mathbb{Z}_v$ with the property that each $r \in \mathbb{Z}_v$ can be written in *at most* one way as $r = a_i + a_j$ with $i < j$ (resp. $i \leqq j$).

Although $n_\gamma$ and $v_\gamma$ do not seem to have been studied before, the other functions have an extensive literature. For example $n_\beta$ is the subject of a series of papers by Rohrbach, Moser, Hämmerer, Hofmeister, and others ([36], [37], [46], [59]–[61], [65], [79]) who refer to the set $A$ as an *interval basis* (*Abschnittsbasis*), or *2-basis*, and by Lunnon and others ([1], [1a], [43a], [56], [76]) under the name of the *postage stamp problem.* $n_\delta$ was briefly mentioned by Rohrbach in [66]. The functions $v_\beta$ and $v_\delta$ have been studied by Singer, Erdös, Turán, Bose, Chowla and others (see [11], [21], [40, Chapt. II]). The set $A$ associated with $v_\beta$ is often called a $B_2$-sequence. Other types of additive bases have been defined in [14], [19], [40], [45], [51]. (Since this paper impinges on many different parts of combinatorics we have attempted to include a fairly complete bibliography.)

* Received by the editors April 1, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

Our interest in $v_\gamma$ stems from its application to error-correcting codes. Let $A(k, 2d, w)$ denote the largest possible number of binary vectors, each containing $w$ 1's and $k-w$ 0's, such that any two vectors differ in at least $2d$ places ([6], [57]). It can be shown ([28], [29]) that

$$A(k, 6, w) \geqq \frac{1}{v_\gamma(k)} \binom{k}{w},$$

and there is a similar bound for $A(k, 2d, w)$ using sets in which all sums of $d-1$ distinct elements are distinct modulo $v$. When combined with Theorem 1, this implies

$$A(k, 6, w) \geqq \frac{k^{w-2}}{w!}(1+o(1)) \quad \text{as } k \to \infty,$$

which is stronger than any previously known bound (see [28]). We should also point out that the function $A(k, 2d, w)$ has been studied under another guise in extremal set theory by Erdös, Hanani, Schönheim, Kalbfleisch, Stanton and others (see [20], [73], [77]) in the following context. Let $D(t, k, s)$ denote the maximum number of $k$-element subsets of an $s$-element set $S$ such that every $t$-element subset of $S$ is contained in at most one of the $k$-element subsets. Then $D(t, k, s) = A(s, 2k-2t+2, k)$.

We shall justify our interest in $n_\gamma$ in § 3.

**2. Tables, bounds and properties.** Tables I–IV give values of these eight functions, and examples of the sets $A$ which attain them. Usually the (lexicographically) first

TABLE I
$n_\alpha(k)$ *and* $n_\beta(k)$.

| $k$ | $n_\alpha(k)$ | An example of the set $A$. |
|---|---|---|
| 2 | 1 | {0, 1} |
| 3 | 3 | {0, 1, 2} |
| 4 | 6 | {0, 1, 2, 4} |
| 5 | 9 | {0, 1, 2, 3, 6} |
| 6 | 13 | {0, 1, 2, 3, 6, 10} |
| 7 | 17 | {0, 1, 2, 3, 4, 8, 13} |
| 8 | 22 | {0, 1, 2, 3, 4, 8, 13, 18} |
| 9 | 27 | {0, 1, 2, 3, 4, 5, 10, 16, 22} |
| 10 | 33 | {0, 1, 2, 3, 4, 5, 10, 16, 22, 28} |
| 11 | 40 | {0, 1, 2, 4, 5, 6, 10, 13, 20, 27, 34} |
| 12 | 47 | {0, 1, 2, 3, 6, 10, 14, 18, 21, 22, 23, 24} |
| 13 | 56 | {0, 1, 2, 4, 6, 7, 12, 14, 17, 21, 30, 39, 48} |
| 14 | 65 | {0, 1, 2, 4, 6, 7, 12, 14, 17, 21, 30, 39, 48, 57} |

| $k$ | $n_\beta(k)$ | An example of the set $A$. |
|---|---|---|
| 2 | 2 | {0, 1} |
| 3 | 4 | {0, 1, 2} |
| 4 | 8 | {0, 1, 3, 4} |
| 5 | 12 | {0, 1, 3, 5, 6} |
| 6 | 16 | {0, 1, 3, 5, 7, 8} |
| 7 | 20 | {0, 1, 2, 5, 8, 9, 10} |
| 8 | 26 | {0, 1, 2, 5, 8, 11, 12, 13} |
| 9 | 32 | {0, 1, 2, 5, 8, 11, 14, 15, 16} |
| 10 | 40 | {0, 1, 3, 4, 9, 11, 16, 17, 19, 20} |
| 11 | 46 | {0, 1, 2, 3, 7, 11, 15, 19, 21, 22, 24} |
| 12 | 54 | {0, 1, 2, 3, 7, 11, 15, 19, 23, 25, 26, 28} |
| 13 | 64 | {0, 1, 3, 4, 9, 11, 16, 21, 23, 28, 29, 31, 32} |
| 14 | 72 | {0, 1, 3, 4, 9, 11, 16, 20, 25, 27, 32, 33, 35, 36} |

TABLE II
$n_\gamma(k)$ and $n_\delta(k)$.

| $k$ | $n_\gamma(k)$ | An example of the set $A$. | $k$ | $n_\delta(k)$ | An example of the set $A$. |
|---|---|---|---|---|---|
| 2 | 1 | — | 2 | 3 | {0, 1} |
| 3 | 3 | {0, 1, 2} | 3 | 5 | {0, 1, 2} |
| 4 | 6 | {0, 1, 2, 4} | 4 | 9 | {0, 1, 3, 4} |
| 5 | 9 | {0, 1, 2, 4, 7} | 5 | 13 | {0, 1, 2, 6, 9} |
| 6 | 13 | {0, 1, 2, 3, 6, 10} | 6 | 19 | {0, 1, 3, 12, 14, 15} |
| 7 | 17 | {0, 1, 2, 3, 4, 8, 13} | 7 | 21 | {0, 1, 2, 3, 4, 10, 15} |
| 8 | 24 | {0, 1, 2, 4, 8, 13, 22} | 8 | 30 | {0, 1, 3, 9, 11, 12, 16, 26} |
| 9 | 30 | {0, 1, 2, 4, 10, 15, 17, 22, 28} | 9 | 35 | {0, 1, 2, 7, 8, 11, 26, 29, 30} |
| 10 | 36 | {0, 1, 2, 3, 6, 12, 19, 20, 27, 33} | | | |

TABLE III
$v_\alpha(k)$ and $v_\beta(k)$.

| $k$ | $v_\alpha(k)$ | An example of the set $A$. | $k$ | $v_\beta(k)$ | An example of the set $A$. |
|---|---|---|---|---|---|
| 2 | 1 | {0, 1} | 2 | 2 | {0, 1} |
| 3 | 3 | {0, 1, 2} | 3 | 6 | {0, 1, 3} |
| 4 | 6 | {0, 1, 2, 4} | 4 | 12 | {0, 1, 4, 6} |
| 5 | 11 | {0, 1, 2, 4, 7} | 5 | 22 | {0, 1, 4, 9, 11} |
| 6 | 19 | {0, 1, 2, 4, 7, 12} | 6 | 34 | {0, 1, 4, 10, 12, 17} |
| 7 | 31 | {0, 1, 2, 4, 8, 13, 18} | 7 | 50 | {0, 1, 4, 10, 18, 23, 25} |
| 8 | 43 | {0, 1, 2, 4, 8, 14, 19, 24} | 8 | 68 | {0, 1, 4, 9, 15, 22, 32, 34} |
| 9 | 63 | {0, 1, 2, 4, 8, 15, 24, 29, 34} | 9 | 88 | {0, 1, 5, 12, 25, 27, 35, 41, 44} |
| 10 | 80 | {0, 1, 2, 4, 8, 15, 24, 29, 34, 46} | 10 | 110 | {0, 1, 6, 10, 23, 26, 34, 41, 53, 55} |

TABLE IV
$v_\gamma(k)$ and $v_\delta(k)$.

| $k$ | $v_\gamma(k)$ | An example of the set $A$. |
|---|---|---|
| 2 | 2 | {0, 1} |
| 3 | 3 | {0, 1, 2} |
| 4 | 6 | {0, 1, 2, 4} |
| 5 | 11 | {0, 1, 2, 4, 7} |
| 6 | 19 | {0, 1, 2, 4, 7, 12} |
| 7 | 28 | {0, 1, 2, 4, 8, 15, 20} |
| 8 | 40 | {0, 1, 5, 7, 9, 20, 23, 35} |
| 9 | 56 | {0, 1, 2, 4, 7, 13, 24, 32, 42} |
| 10 | 72 | {0, 1, 2, 4, 7, 13, 23, 31, 39, 59} |

| $k$ | $v_\delta(k)$ | An example of the set $A$. |
|---|---|---|
| 2 | 3 | {0, 1} |
| 3 | 7 | {0, 1, 3} |
| 4 | 13 | {0, 1, 3, 9} |
| 5 | 21 | {0, 1, 4, 14, 16} |
| 6 | 31 | {0, 1, 3, 8, 12, 18} |
| 7 | 48 | {0, 1, 3, 15, 20, 38, 42} |
| 8 | 57 | {0, 1, 3, 13, 32, 36, 43, 52} |
| 9 | 73 | {0, 1, 3, 7, 15, 31, 36, 54, 63} |
| 10 | 91 | {0, 1, 3, 9, 27, 49, 56, 61, 77, 81} |
| 11 | ? | |
| 12 | 133 | {0, 1, 3, 12, 20, 34, 38, 81, 88, 94, 104, 109} |

example of $A$ is given. The entries in the $n_\beta$ table are taken from [1], [56], and [76], and the entries in the $v_\delta$ table which come from difference sets (see (8)) are taken from [2a, Table 6.1]. The other entries are believed to be new.

The best bounds presently known for these functions are as follows.

THEOREM 1.

$$\tfrac{5}{18}(k-1)^2 < n_\alpha(k), n_\beta(k) < .4802k^2 + O(k), \tag{1}$$

$$\tfrac{5}{18}(k-1)^2 < n_\gamma(k), n_\delta(k) < \tfrac{1}{2}k^2 + O(k), \tag{2}$$

$$2k^2 - O(k^{3/2}) < v_\alpha(k), v_\beta(k) < 2k^2 + O(k^{36/23}), \tag{3}$$

$$k^2 - O(k) < v_\gamma(k) < k^2 + O(k^{36/23}), \tag{4}$$

$$k^2 - k + 1 \leqq v_\delta(k) < k^2 + O(k^{36/23}). \tag{5}$$

*Discussion of Proof.* Hämmerer and Hofmeister [36] showed that $n_\beta(k) > 5(k-1)^2/18$, and it is not difficult to modify their proof to give $n_\alpha(k) > 5(k-1)^2/18$. The lower bounds in (2) then follow from $n_\alpha(k) \leqq n_\gamma(k)$ and $n_\beta(k) \leqq n_\delta(k) - 1$ (see Lemma 2 below). The upper bound in (1) is due to Klotz [46]. Since there are $\binom{k}{2}$ sums $a_i + a_j (i < j)$ from a $k$-element set $A$, we have immediately that

$$n_\gamma(k) \leqq \binom{k}{2} \leqq v_\gamma(k), \tag{6}$$

and similarly

$$n_\delta(k) \leqq \binom{k+1}{2} \leqq v_\delta(k), \tag{7}$$

which imply the upper bound in (2). Notice that if equality holds on either side of (6) then it holds on both sides, and similarly in (7).

The lower bounds in (3) follow from a straightforward modification of the Erdös-Turán argument ([21], [40, Chapt. II, § 3, Theorem 4]); we omit the details. The lower bound in (4) will be proved at the end of this section. The lower bound in (5) holds because if the sums $a_i + a_j$ $(1 \leqq i \leqq j \leqq k)$ are distinct modulo $v$, then so are the $k(k-1)$ nonzero differences $a_i - a_j$; hence $v - 1 \geqq k^2 - k$. It follows that the equality signs can only hold in (7) when $k = 2$; thus

$$n_\delta(k) < \binom{k+1}{2} < v_\delta(k) \quad \text{for } k > 2.$$

We shall see in Theorem 5 that the equality signs can only hold in (6) when $k = 2, 3$ or 4. The upper bounds in (3)–(5) are all obtained by using Singer difference sets and the fact that (see [43]), whenever $x$ is sufficiently large, there is a prime $p$ with

$$x < p \leqq x + x^{13/23}$$

(compare [40, Chapt. II, § 3, Theorem 6]. In particular, difference sets attain the lower bound in (5), so

$$v_\delta(k) = k^2 - k + 1, \quad \text{whenever } k - 1 \text{ is a prime power.} \tag{8}$$

A projective plane of order 6 would have implied $v_\delta(7) = 43$, but since this plane does not exist we may regard the cyclic shifts of $A = \{0, 1, 3, 15, 20, 38, 42\}$ modulo 48 (corresponding to $v_\delta(7) = 48$) as giving, in a sense, the best approximation to such a plane. Other approximations are described in [40a] and [56a].

The following properties of these functions are easily established.

LEMMA 2.

$$\begin{aligned} &n_\alpha(k) \leqq v_\alpha(k), && n_\beta(k) \leqq v_\beta(k), \\ &n_\gamma(k) \leqq v_\gamma(k), && n_\delta(k) \leqq v_\delta(k), \\ &n_\alpha(k) \leqq n_\beta(k), && n_\gamma(k) \leqq n_\delta(k), \\ &v_\alpha(k) \leqq v_\beta(k) - 1, && v_\gamma(k) \leqq v_\delta(k), \\ &n_\alpha(k) \leqq n_\gamma(k), && n_\beta(k) \leqq n_\delta(k) - 1, \\ &v_\alpha(k) \geqq v_\gamma(k), && v_\beta(k) \geqq v_\delta(k) - 1. \end{aligned}$$

LEMMA 3.

(a) *If* $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ *attains* $n_\alpha(k)$, *then* $a_2 = 1$, $a_3 = 2$, $a_4 = 3$ *or* 4, *and* $a_k \leqq n_\alpha(k-1) + 1$. *Furthermore,*

$$n_\alpha(k) + 3 \leqq n_\alpha(k+1), \quad \text{for } k \geqq 3.$$

(b) *If* $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ *attains* $n_\beta(k)$, *then* $a_2 = 1$, $a_3 = 2$ *or* 3, $a_4 = 3, 4$ *or* 5, *and* $a_k \leqq n_\beta(k-1) + 1$. *Furthermore,*

$$n_\beta(k) + 2 \leqq n_\beta(k+1).$$

(c) *If* $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ *attains* $n_\gamma(k)$ (*or* $n_\delta(k)$, $v_\gamma(k)$ *or* $v_\delta(k)$) *and if* $n_\gamma(k)$ (*or* $n_\delta(k)$, $v_\gamma(k)$, $v_\delta(k)$) *is of the form* $p^r q^s$, *with* $p$, $q$ *prime*, $r, s \geqq 0$, *then we may assume that* $a_2 = 1$.

*Proof.* (a) and (b) are straightforward. (c) Suppose $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ is such that the sums $a_i + a_j$ $(i < j)$ cover $\mathbb{Z}_n$, where $n = n_\gamma(k) = p^r q^s$. If some $a_i$ is relatively prime to $n$ then $A' = a_i^{-1} A$ contains 0 and 1, and also attains $n_\gamma(k) = n$. If not, since not all the $a_i$ can be divisible by $p$, nor by $q$, we can find $a_t$ and $a_u$ such that $p \mid a_t$, $q \nmid a_t$, $p \nmid a_u$, $q \mid a_u$. Then $a_t - a_u$ is relatively prime to $n$, and $(a_t - a_u)^{-1}(A - a_u)$ contains 0 and 1 and attains $n_\gamma(k) = n$. Similarly for $n_\delta$, $v_\gamma$ and $v_\delta$. Q.E.D.

Parts (a) and (b) of this Lemma simplify the computation of $n_\alpha$ and $n_\beta$ (and the absence of similar results for the other six functions makes their calculation more difficult). The calculations are further simplified by the next lemma.

LEMMA 4. *If there is no* $k$*-element set* $A$ *such that the sums* $a_i + a_j$ $(i < j)$ *cover* $[1, m]$, *then* $n_\alpha(k) \leqq m - 1$; *and similarly for* $n_\beta(k)$. *If there is no* $k$*-element set* $A$ *such that the sums* $a_i + a_j$ $(i < j)$ *belong to* $[0, m]$ *and are distinct, then* $v_\alpha(k) \geqq m + 1$; *and similarly for* $v_\beta(k)$.

But these properties need not hold for the modular functions. Consider for example the problem of determining $n_\gamma(8)$. The set $A = \{0, 1, 2, 3, 4, 8, 13, 18\}$ covers $\mathbb{Z}_n$ for all $n$ in the range $8 \leqq n \leqq 22$, but no 8-element set covers $\mathbb{Z}_{23}$. Nevertheless $A = \{0, 1, 2, 4, 8, 13, 18, 22\}$ covers $\mathbb{Z}_{24}$, and $n_\gamma(8) = 24$. Similarly when determining $n_\delta(6)$ we find that $A = \{0, 1, 2, 5, 7, 11\}$ covers $\mathbb{Z}_n$ for $6 \leqq n \leqq 15$, $A = \{0, 1, 2, 4, 9, 14\}$ covers $\mathbb{Z}_{16}$, $A = \{0, 1, 2, 3, 8, 12\}$ covers $\mathbb{Z}_{17}$, no 6-element set covers $\mathbb{Z}_{18}$, $A = \{0, 1, 3, 12, 14, 15\}$ covers $\mathbb{Z}_{19}$, and $n_\delta(6) = 19$.

We conclude this section by determining when the equality signs can hold in (6).

THEOREM 5.

$$n_\gamma(k) = \binom{k}{2} = v_\gamma(k) \tag{9}$$

*if and only if* $k = 2$, 3 *or* 4; *otherwise* $n_\gamma(k) < \binom{k}{2} < v_\gamma(k)$.

*Proof.*

(i) If (9) holds and $k \equiv 0$ or 1 (modulo 4) then $k$ must be a perfect square (cf. [80]). For in this case $n = \binom{k}{2}$ is even, and so the parity of an element of $\mathbb{Z}_n$ is well defined. Let $A \subseteq \mathbb{Z}_n$ attain $n_\gamma(k) = v_\gamma(k) = \binom{k}{2}$, and suppose $\alpha$ of the $a_i$ are odd and $\beta$ are even, with $\alpha + \beta = k$. The number of odd sums $a_i + a_j$ $(i < j)$ is $\alpha\beta = \frac{1}{2}\binom{k}{2}$, hence $(\alpha - \beta)^2 = k^2 - k(k-1) = k$ is a perfect square.

(ii) Equation (9) holds for $n = 2, 3, 4$ (see Tables II, IV), part (i) eliminates $k = 5$, 8, 12 and 13, and a computer search elminated $k = 6$, 7, 9, 10 and 11.

(iii) The values of $k \geqq 14$ are eliminated by the following lemma. Q.E.D.

LEMMA 6. *Suppose* $A = \{0 = a_1 < a_2 < \cdots < a_k\}$ *is a subset of* $\mathbb{Z}_n$ *such that the sums* $a_i + a_j$ $(i < j)$ *represent each element of* $\mathbb{Z}_n$ *at most once. Let* $u = [n/3]$ *and assume*

$$k \leqq u - 1. \tag{10}$$

*Then*

$$k^2u^2 \leqq n\{u(u-1) + 3ku - k(k+1)\}. \tag{11}$$

*Proof.* The proof is a modification of the Erdös-Turán argument ([21], [40, p. 86]). Consider the $n$ subsets

$$\mathcal{I}_m = \{m, m+1, \cdots, m+u-1\}$$

of $\mathbb{Z}_n$, for $0 \leqq m \leqq n-1$, and let $A_m = |\mathcal{I}_m \cap A|$. Since each $a_i$ belongs to exactly $u$ subsets,

$$\sum_{m=0}^{n-1} A_m = ku. \tag{12}$$

Let $T$ be the number of triples $(a_i, a_j, m)$ with $1 \leqq i < j \leqq k$ and $a_i \in \mathcal{I}_m$, $a_j \in \mathcal{I}_m$. The number of pairs $(a_i, a_j)$ contained in $\mathcal{I}_m$ is $\frac{1}{2}A_m(A_m - 1)$, so

$$T = \tfrac{1}{2}\sum_{m=0}^{n-1} A_m(A_m - 1). \tag{13}$$

From (12), (13) and Cauchy's inequality,

$$T \geqq \frac{k^2u^2}{2n} - \frac{ku}{2}. \tag{14}$$

For $i \leqq j$ let

$$\rho(a_i, a_j) = \min\{a_j - a_i, n - a_j + a_i\}.$$

If $a_i$ and $a_j$, $i < j$, are contained in $\mathcal{I}_m$, $\rho(a_i, a_j)$ is an integer $d$ with $1 \leqq d \leqq u - 1$. Conversely, given $d \in [1, u-1]$, how many pairs $(a_i, a_j)$ satisfy $i < j$, $\rho(a_i, a_j) = d$? It is easly seen that the answer is 0, 1 or 2. If there is one solution we call $d$ *ordinary*, if two, *special*. A special $d$ is associated with a unique triple $a_i, a_j, a_k$ with

$$2a_j = a_i + a_k, \qquad \rho(a_i, a_j) = \rho(a_j, a_k).$$

Since there is at most one special $d$ associated with $a_j$, there are at most $k$ special $d$'s. An ordinary $d$ contributes $u - d$ to $T$ since the unique pair $(a_i, a_j)$ with $\rho(a_i, a_j) = d$ is contained in exactly $u - d$ of the sets $\mathcal{I}_m$. Similarly, a special $d$ contributes $2(u - d)$ to $T$,

and so

$$T = \sum_{d \text{ ordinary}} (u-d) + \sum_{d \text{ special}} 2(u-d)$$

$$\leqq \sum_{d=1}^{u-1} (u-d) + \sum_{\nu=1}^{s} (u-d_\nu),$$

where $d_1, \cdots, d_s$ are the special values of $d$, with $s \leqq k$. Using (10) we can bound this by

$$\begin{aligned} T &\leqq \tfrac{1}{2}u(u-1) + ku - (1+2+\cdots+k) \\ &= \tfrac{1}{2}u(u-1) + ku - \tfrac{1}{2}k(k+1), \end{aligned} \tag{15}$$

and (11) follows from (14) and (15). Q.E.D.

COROLLARIES.

(i) *If we set* $n = \binom{k}{2}$, *then for* $k \geqq 14$ (10) *is satisfied but* (11) *is not, which eliminates the cases* $k \geqq 14$ *of Theorem* 5.

(ii) *For n large,* (11) *implies*

$$k \leqq \sqrt{n} + O(1),$$

*which is equivalent to the lower bound in* (4).

**3. Harmonious graphs.** We call a connected graph with $v$ nodes and $e \geqq v$ edges *harmonious* if it is possible to label the nodes $x$ with distinct elements $\lambda(x)$ of $\mathbb{Z}_e$ in such a way that, when each edge $xy$ is labeled with $\lambda(x)+\lambda(y)$, the resulting edge labels are distinct. If the graph is a tree (with $v$ nodes and $e = v-1$ edges) we require exactly one node label to be repeated. Such a labeling of the nodes and edges is called a *harmonious labeling* of the graph. In a harmonious labeling the node labels are distinct (or contain exactly one duplicate, if the graph is a tree), and the induced edge labels are $0, 1, \cdots, e-1$. Fig. 1 shows some harmonious graphs with 5 nodes, and Fig. 2 gives harmonious labelings of all trees with 7 nodes.

Harmonious graphs arise naturally out of the problems considered in § 1. For if $n_\gamma(v) = v_\gamma(v) = \binom{v}{2}$ is attained by a set $A = \{a_1, \cdots, a_v\}$, for $v \geqq 3$, then $a_1, \cdots, a_v$ is a harmonious labeling of $K_v$, the complete graph on $v$ nodes. From Theorem 5 we obtain:

THEOREM 7. *The complete graph on v nodes is harmonious if and only if* $v \leqq 4$ (*see Fig.* 3).

For larger values of $v$ it is natural to ask how large a subgraph of $K_v$ can be harmonious. From the definition in § 1 we see that the answer is given by:

$n_\gamma(v)$ is the greatest number of edges in any harmonious graph on $v$ nodes.

For if $A = \{a_1, \cdots, a_v\}$ attains $n_\gamma(v)$, we label the nodes of $K_v$ with $a_1, \cdots, a_v$ and omit any edge whose label has already appeared. Since by definition the sums $a_i + a_j$ $(i<j)$ cover $\mathbb{Z}_e$, every edge label appears at least once. For example Fig. 1(n) shows the largest harmonious graph on 5 nodes, corresponding to the value $n_\gamma(5) = 9$, which is attained by $A = \{0, 1, 2, 4, 7\}$. One of the two edges labeled 2 has been omitted from $K_5$.

Although many other ways of labeling graphs have been studied in the literature ([8], [9], [25], [49], [54], [67]), this one appears to be new. However, there are many similarities between harmonious graphs and what are called graceful graphs. A connected graph with $v$ nodes and $e \geqq v-1$ edges is *graceful* if it is possible to label the nodes $x$ with distinct integers $\mu(x)$ from $\{0, 1, \cdots, e\}$ in such a way that, when each edge $xy$ is labeled with $|\mu(x)-\mu(y)|$, the resulting edge labels are distinct (and therefore

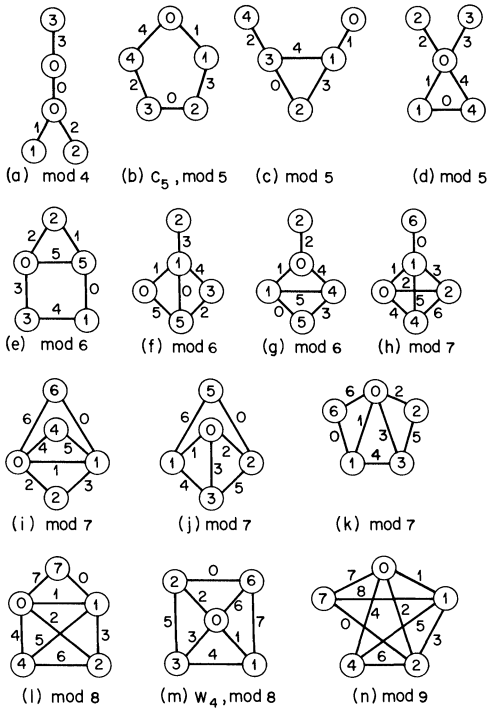



FIG. 1. *Some harmonious graphs with* 5 *nodes.*

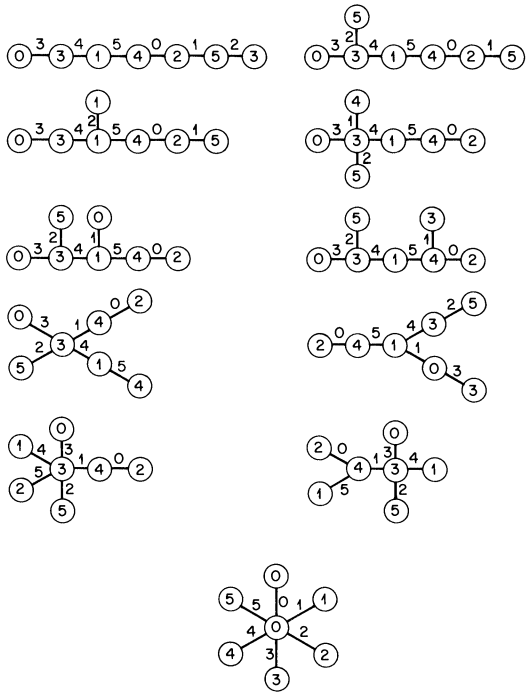


FIG. 2. *Harmonious labelings of the trees with* 7 *nodes* (*modulo* 6).

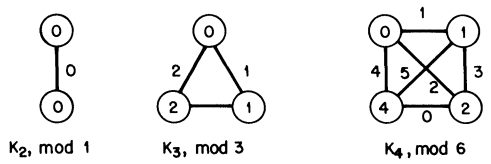



FIG. 3. *The complete graphs $K_2$, $K_3$ and $K_4$.*

all values in $\{1, 2, \cdots, e\}$ appear uniquely). A graceful labeling of a graph is also called a *β-valuation* or a *restricted difference basis.* These have an extensive literature ([3]–[5], [7]–[10], [15], [22]–[27], [30], [31], [33]–[35], [42], [44], [47], [48], [50], [53], [57a], [58], [64], [67], [69], [74], [78], [81]).

We are interested in determining which graphs are harmonious. The principal results we have obtained are summarized in Table V, which shows which graphs are harmonious ($H$) and for comparison which are graceful ($G$). The entries in the table are explained in the remaining sections.

TABLE V
*Comparison of harmonious and graceful graphs.*

| Graph | Harmonious? | Graceful? |
|---|---|---|
| Caterpillars | $H$ (§ 5) | $G$ [15], [67] |
| Trees | Conjectured to be $H$, true for $\leqq 9$ nodes | Conjectured to be $G$, true for $\leqq 16$ nodes [7], [67] |
| Cycle $C_{4m}$ | Not $H$ (§ 6) | $G$ [10], [64] |
| Cycle $C_{4m+1}$ | $H$ (§ 6) | Not $G$ [10], [64]. |
| Cycle $C_{4m+2}$ | Not $H$ (§ 6) | Not $G$ [10], [64] |
| Cycle $C_{4m+3}$ | $H$ (§ 6) | $G$ [10], [64] |
| Ladder $L_n$ | $H$ iff $n \geqq 3$ (§ 7) | $G$ [10, p. 121] |
| Friendship graph $F_n$ | $H$ iff $n \not\equiv 2$ (mod 4) (§ 8) | $G$ iff $n \equiv 0$ or 1 (mod 4) [4], [5] |
| Fan $f_n$ | $H$ (§ 9) | $G$ (§ 9) |
| Wheel $W_n$ | $H$ (§ 10) | $G$ [22], [44] |
| Complete graph $K_n$ | $H$ iff $n \leqq 4$ (Theorem 7) | $G$ iff $n \leqq 4$ [25], [74] |
| Complete bipartite $K_{m,n}$ | $H$ iff $m$ or $n = 1$ (Theorem 19) | $G$ [25], [67] |
| Small graphs | All with $\leqq 5$ nodes are $H$ except for 5 (Fig. 15) | All with $\leqq 5$ nodes are $G$ except for 3 [25] |
| Petersen | $H$ (Fig. 16) | $G$ [25] |
| Cube, octahedron | Not $H$ (Theorem 22) | $G$ [25] |
| Icosahedron | $H$ (Fig. 17) | $G$ [23] |
| Dodecahedron | $H$ (Fig. 18) | $G$ [23] |
| Most graphs | Not $H$ (Theorem 23) | Not $G$ (Theorem 24) |

Several of these families of graphs were suggested by the following application. Consider a network of transmitting stations, each of which must be able to communicate with certain others—those to which it is linked in the network. The total bandwidth available is divided into $e$ channels, where $e$ is the number of links in the network, and each station $x$ is assigned a number $\lambda(x)$ from $\mathbb{Z}_e$. When $x$ and $y$ communicate they use channel number $\lambda(x) + \lambda(y)$. If the numbering is harmonious, each channel is assigned to exactly one link.

Harmonious graphs are also interesting because they lead to modular versions of various combinatorial problems. For example, a harmonious labeling of the friendship graph $F_n$ (see § 8) may be regarded as a modular generalization of the Langford-Skolem

problem (see [2], [4], [17], [18], [32], [41], [52], [55], [62], [63], [68], [70], [75]), a version of that problem which does not seem to have been discussed before. Harmonious labelings of fans, wheels, complete bipartite graphs, etc. (see below) also have interesting combinatorial interpretations.

To conclude this section we mention that there is a curious geometric interpretation of the condition that a graph $G$ be harmonious. Let $P_e$ denote a fixed regular $e$-gon embedded in the plane. Then $G$ is harmonious if and only if the nodes of $G$ can be embedded into the nodes of $P_e$ so that no two edges of the embedded copy of $G$ are parallel. This follows from the observation that if the nodes of $P_e$ are labeled cyclically with $0, 1, \cdots, e-1$, then the direction of the chord joining $i$ and $j$ depends only on $i+j$ (modulo $e$). (The condition must be modified slightly if $G$ is a tree.) For example, Fig. 4 shows the graph of Fig. 1(f) embedded in a regular hexagon.

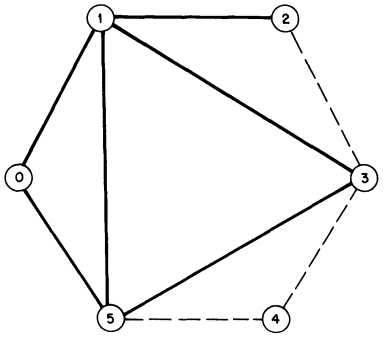



FIG. 4. *The harmonious labeling of Fig.* 1(*f*) *corresponds to an embedding of this graph in a regular hexagon in such a way that no two edges are parallel.*

**4. General properties of harmonious graphs.** The first property concerns equivalent labelings of the same graph.

THEOREM 8. *If $\lambda$ is a harmonious labeling of the nodes of a graph with e edges, then so is $a\lambda + b$, where a is an invertible element of $\mathbb{Z}_e$ and b is any element of $\mathbb{Z}_e$.*

*Proof.* The edge labels $\lambda(x)+\lambda(y)$ are changed to $a(\lambda(x)+\lambda(y))+2b$, but remain distinct. Q.E.D.

COROLLARY.

(i) *Any node in a harmonious graph can be assigned the label* 0.

(ii) *The repeated node label in a harmonious tree can be any element of $\mathbb{Z}_e$.*

On the other hand one harmonious graph may lead to others via the following constructions, which have the effect of moving an edge with a given label from one part of the graph to another.

THEOREM 9. *Let G be a harmoniously labeled graph containing* (i) *an edge wx with label $\lambda(w)+\lambda(x)$, and* (ii) *a pair of nodes y, z not joined by an edge but satisfying $\lambda(w)+\lambda(x)=\lambda(y)+\lambda(z)$. Then deleting the edge wx and inserting yz changes G to another harmonious graph.*

For example we can move the edge labeled 4 in Fig. 1(b) and obtain Fig. 1(c).

THEOREM 10. *Let G be a harmoniously labeled tree containing an edge wx labeled $\lambda(w)+\lambda(x)$, where x is an endpoint (of degree* 1), *and $\lambda(x)$ is the repeated node label. If y is any other node in G, we may delete edge wx and node x and replace them with a new node z and edge yz where z is labeled with $\lambda(z)=\lambda(w)+\lambda(x)-\lambda(y)$.*

For example the second and third trees in Fig. 2 are obtained from the first by moving the edge labeled 2.

The last theorem in this section gives a necessary condition for certain graphs to be harmonious.

THEOREM 11. *If a harmonious graph has an even number e of edges and the degree of every node is divisible by* $2^{\alpha}$ $(\alpha \geqq 1)$, *then e is divisible by* $2^{\alpha+1}$.

*Proof.* Let node $x$ have label $\lambda(x)$ and degree $\delta(x)$. The sum of the *edge* labels is $\sum_x \delta(x)\lambda(x) \equiv 0+1+\cdots+(e-1) \equiv \frac{1}{2}e(e-1) \equiv e/2$ (modulo $e$); hence $2^{\alpha}$ divides $e/2$ and so $2^{\alpha+1}$ divides $e$. Q.E.D.

For example the 1-skeleton of the octahedron has 12 edges and 6 nodes, each of degree 4, so is not harmonious.

**5. Are all trees harmonious?** It is easy to see that paths and stars are harmonious (see the first and last examples in Fig. 2). More generally, let a *caterpillar* be a tree with the property that the removal of its endpoints leaves a path.

THEOREM 12. *Any caterpillar is harmonious.*

*Proof.* Draw the caterpillar as a bipartite graph, as shown in Fig. 4a, with say $l$ nodes on the left and $r$ on the right. There are $e = l+r-1$ edges. If $e$ is odd, or if $e$ is even and $r$ is odd, choose $a \in \mathbb{Z}_e$ so that $2a = r-1$ (in $\mathbb{Z}_e$). If $e$ and $r$ are both even, then $l$ is odd and we choose $a$ so that $2a = 1-l$. We label the left-hand nodes $a, a+1, \cdots, a+l-1$ and the right-hand nodes $-a, 1-a, \cdots, r-1-a$, as in Fig. 4a. The full set of node labels is $\{0, 1, \cdots, e-1\}$ with either $a$ repeated (if $2a = r-1$) or $-a$ repeated (if $2a = 1-l$). The edge labels are $\{0, 1, \cdots, e-1\}$, and the graph is harmonious. Q.E.D.

We shall usually just specify the node labels and leave to the reader the straightforward verification that the labeling is harmonious.

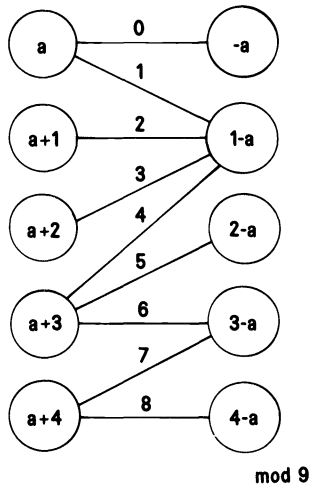



FIG. 4a. *A caterpillar with* $e = 9$ *edges, drawn as a bipartite graph with* $l = 5$ *nodes on the left and* $r = 5$ *nodes on the right. We obtain a harmonious labeling by choosing* $a = 2$, *so that* $2a = r-1$.

By repeatedly applying the constructions of Theorems 9 and 10 to caterpillars, it is easy to generate large numbers of harmonious trees. Those with 7 nodes are shown in Fig. 2, and in the same way we have established the following theorem, whose proof is omitted.

THEOREM 13. *All trees with* $\leqq 9$ *nodes are harmonious.*

We conjecture that all trees are harmonious (cf. [7]).

**6. Cycles.**

THEOREM 14. *The cycle $C_n$ with n nodes, $n \geqq 3$, is harmonious if and only if n is odd.*

*Proof.* If $n$ is odd we label the nodes $0, 1, \cdots, n-1$ (see Fig. 1(b)). If $n = 2m$ is even, suppose $a_0, a_1, \cdots, a_{2m-1}$ is a harmonious labeling of $C_{2m}$. The numbers $a_0+a_1, a_1+a_2, \cdots, a_{2m-1}+a_0$ are congruent (modulo $2m$) to some permutation of $0, 1, 2, \cdots, 2m-1$. Adding these numbers we obtain $2S \equiv S$ (modulo $2m$), where $S = 0+1+2+\cdots+2m-1 \equiv m$ (modulo $2m$). Hence $m \equiv 0$ (modulo $2m$), a contradiction. Q.E.D.

**7. Ladders are harmonious.** The *ladder* $L_n$ ($n \geqq 2$) is the product graph $P_2 \times P_n$, and contains $2n$ nodes and $3n-2$ edges (Figs. 5, 6).

THEOREM 15. *All ladders except $L_2$ are harmonious.*

*Proof.* $L_2 = C_4$ is not harmonious by the previous theorem. $L_{2a+1}$ ($a \geqq 1$) is harmonious: label one path $0, a+1, 1, a+2, 2, a+3, \cdots$ and the other $3a+1, 2a+1, 3a+2, 2a+2, 3a+3, 2a+3, \cdots$ (Fig. 5). $L_4$ is harmonious: label the paths 0, 5, 1, 9 and 2, 6, 3, 4. Finally Fig. 6 shows a harmonious labeling of $L_{2a}$ for $a \geqq 3$. Q.E.D.

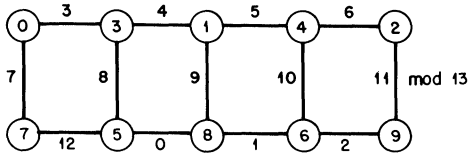



FIG. 5. *The ladder $L_5$.*

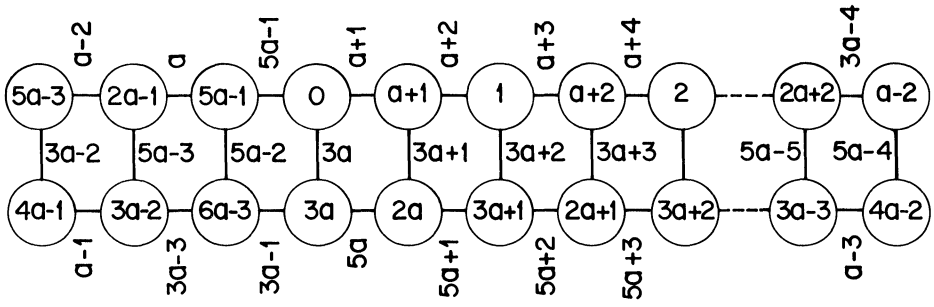



FIG. 6. *The ladder $L_{2a}$, $a \geqq 3$.*

The labeling of $L_{2a+1}$ is exceptionally pleasant since the edges are numbered consecutively. Furthermore by simply joining the ends of the ladder we obtain a harmonious labeling of the *prism* $P_2 \times C_{2a+1}$ (Fig. 7), and the pattern may be continued to produce a harmonious labeling of any $P_m \times C_{2a+1}$ (Fig. 8). The cube $P_2 \times C_4$ is not harmonious (Theorem 22 below), but $P_3 \times C_4$ is (Fig. 9).

**8. Friendship graphs.** The *friendship graph* $F_n$ ($n \geqq 1$) consists of $n$ triangles with a common vertex (see Fig. 10).

THEOREM 16. *$F_n$ is harmonious except when $n \equiv 2$ (modulo 4).*

*Proof.* If $n \equiv 2$ (modulo 4), $F_n$ is not harmonious by Theorem 11. If $n \equiv 0$ or 1 (modulo 4) it was shown by Skolem [75] that the numbers $\{1, 2, \cdots, 2n\}$ may be partitioned into $n$ pairs $(a_r, b_r)$ with $b_r - a_r = r$, for $r = 1, \cdots, n$. Then a harmonious

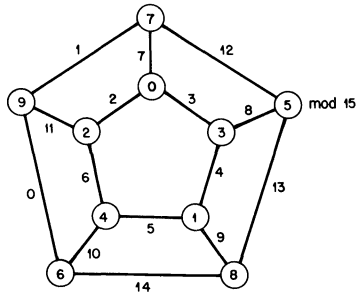



FIG. 7. *The prism* $P_2 \times C_5$.

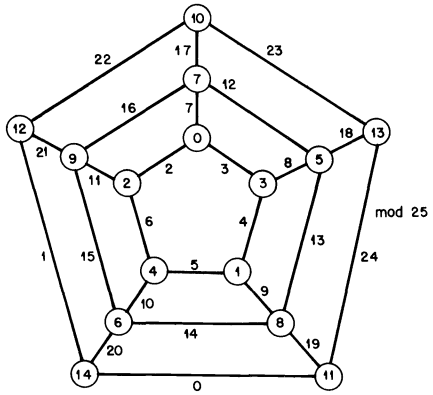



FIG. 8. *The prism* $P_3 \times C_5$.

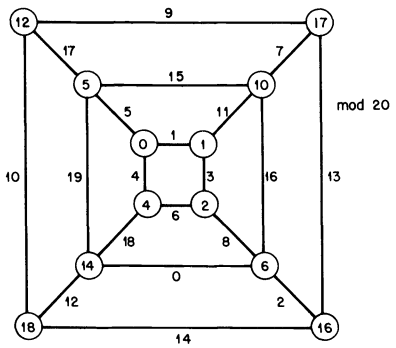



FIG. 9. *The prism* $P_3 \times C_4$.

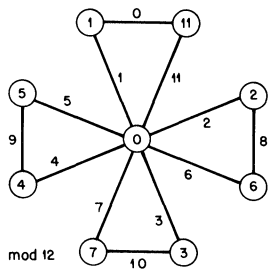



FIG. 10. *The friendship graph* $F_4$.

labeling of $F_n$ is obtained by labeling the vertices of the triangles with $(0, r, n+a_r)$, for $r=1, \cdots, n$ (see Fig. 10). If $n \equiv 3$ (modulo 4) it is known [4, Th. 1, the case $d=3$] that $\{1, 2, \cdots, 2n-6\}$ may be partitioned into $n-3$ pairs $(a_r, b_r)$ with $b_r - a_r = r+2$, for $r=1, \cdots, n-3$. We label the triangles of $F_n$ with $(0, 1, 3n-1)$, $(0, 2, 3n-6)$, $(0, 3n-2, 3n-3)$, and $(0, r+2, n+a_r)$ for $r=1, \cdots, n-3$. Q.E.D.

**9. Fans are harmonious.** The *fan* $f_n$ $(n \geqq 2)$ is obtained by joining all nodes of $P_n$ to a further node called the *center*, and contains $n+1$ nodes and $2n-1$ edges.

THEOREM 17. *$f_n$ is harmonious.*

*Proof.* Let $m = [n/2]$ and label the center with 0 and the nodes of the path with $m$, $n$, $m+1$, $n+1$, $m+2, \cdots$ (see Fig. 11).

*Remarks.*

(i) $f_{2m}$ may also be harmoniously labeled in such a way that the endpoints of the path are 1 and $-1$: label the nodes of the path with $1, 2, 5, 6, 9, 10, \cdots, 4m-3, 4m-2$.

(ii) $f_n$ is also graceful, although this fact does not seem to have been mentioned before: label the center with 0 and the nodes of the path with $2n-1$, 1, $2n-3$, 3, $2n-5, \cdots$.

(iii) Let $g_n(n \geqq 2)$ be the graph with $n+2$ nodes and $3n-1$ edges obtained by joining all nodes of $P_n$ to *two* additional nodes. A harmonious labeling of $g_{2m}$ is obtained by labeling the path with 2, 4, 8, 10, 14, 16, $\cdots, 6m-4, 6m-2$, and the two additional nodes with 0 and 1 (Fig. 12). But $g_{2m+1}$ does not seem to have such a simple labeling.

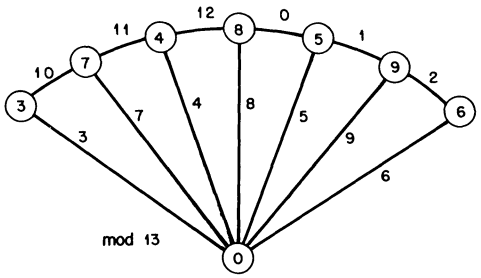



FIG. 11. *The fan $f_7$.*

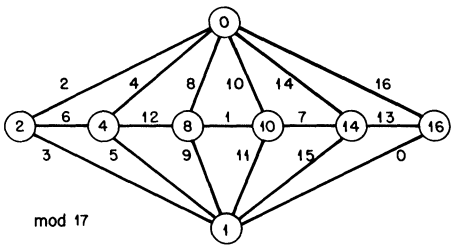



FIG. 12. *The graph $g_6$.*

**10. Wheels are harmonious.** The *wheel* $W_n$ $(n \geqq 3)$ is obtained by joining all nodes of $C_n$ to a further node called the center, and contains $n+1$ nodes and $2n$ edges (see Fig. 13). A harmonious labeling of $W_n$ is equivalent (by Theorem 8) to finding a subset $\{a_1, \cdots, a_n\}$ of $\mathbb{Z}_{2n}$ with the property that

$$a_1, \cdots, a_n, a_1+a_2, a_2+a_3, \cdots, a_n+a_1$$

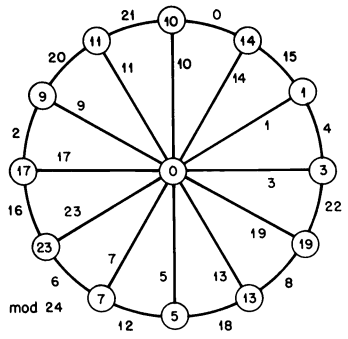



FIG. 13. *The wheel* $W_{12}$.

comprise all the elements of $\mathbb{Z}_{2n}$ (for then we may label the cycle with $a_1, \cdots, a_n$ and the center of the wheel with 0).

THEOREM 18. *$W_n$ is harmonious.*

*Proof.* The cases $W_{2m+1}$, $W_{4m}$, $W_{8m+2}$ and $W_{8m+6}$ will be handled separately. In each case the center is labeled with 0. For $W_{2m+1}$ the cycle is labeled $1, 3, 5, \cdots, 4m+1$. For $W_{4m}$ we divide the cycle into $2m$ pairs, $m$ of which will be labeled $(4i+1, 4i+3)$, $0 \leqq i \leqq m-1$; $m-1$ of which will be labeled $(4i+7, 4i+1)$, $m \leqq i \leqq 2m-2$; and one which will be labeled $(4m-2, 4m+2)$. The actual labeling of the cycle is

$$4m-2, 4m+2;$$

$$1, 3; 4m+7, 4m+1; 5, 7; 4m+11, 4m+5;$$

$$\cdots \qquad \cdots$$

$$4i-7, 4i-5; 4m+4i-1, 4m+4i-7;$$

$$4i-3, 4i-1; 4m+4i+3, 4m+4i-3;$$

$$4i+1, 4i+3; 4m+4i+7, 4m+4i+1;$$

$$\cdots \qquad \cdots$$

$$\cdots \qquad \cdots$$

$$4m-11, 4m-9; 8m-5, 8m-11;$$

$$4m-7, 4m-5; 8m-1, 8m-7;$$

$$4m-3, 4m-1.$$

Fig. 13 shows the labeling of $W_{12}$. To verify that this labeling is harmonious we observe that, out of the residues modulo $8m$, all the numbers congruent to 1 or 3 (modulo 4) except $4m+3$ and $8m-3$ appear as spoke labels, and $4m+3$ and $8m-3$, together with all numbers congruent to 0 (modulo 4) appear on the perimeter. Furthermore, the numbers congruent to 2 (modulo 4) appear on the perimeter in the order $\cdots 4m+8i-6, 4m+8i-10, 4m+8i+2, 4m+8i-2, \cdots$.

For $W_{8m+2}$ $(m \geqq 1)$ the cycle $C_{8m+2}$ will be labeled modulo 4 as follows:

$$\underbrace{2, 1, 2, 1, \cdots, 2, 1;}_{4m+2} \quad \underbrace{1, 1, \cdots, 1;}_{2m-1} \quad \underbrace{0, 0, \cdots, 0.}_{2m+1}$$

The actual labels for these three sets of nodes are

$$4m+2, 16m+1, 4m-2, 16m-3, \cdots, 12m+6, 8m+1;$$

$$4m-3, 8m-3, 4m-7, 8m-7, \cdots, 4m+5, 1;$$

and

$$4m, 8m+4, 4m+4, 8m+8, \cdots, 12m, 8m$$

(the last set being 4 times the labels of $f_{2m+1}$ given above). For example $W_{18}$ is labeled (modulo 36) as follows.

nodes: 10 33 6 29 2 25 34 21 30 17;

perimeter: 7 5 35 31 27 23 19 15 11 22

nodes: 5 13 1; 8 20 12 24 16

perimeter: 18 14 9 28 32 0 4 26.

For $W_{8m+6}$ $(m \geqq 0)$ the cycle is labeled

$$12m+2, 12m+5, 12m-2, 12m+1, \cdots, 4m-2, 4m+1;$$

$$16m+5, 4m-3, 16m+1, 4m-7, \cdots, 12m+9, 1;$$

$$16m+8, 16m+4, 16m-8, 16m-12, \cdots, 24, 20, 8, 4$$

(the last set being 4 times the second labeling of $f_{2m+2}$ given above). For example, $W_{14}$ is labeled (modulo 28) as follows.

nodes: 14 17 10 13 6 9 2 5; 21 1; 24 20 8 4

perimeter: 3 27 23 19 15 11 7 26 22 25 16 0 12 18

**11. Complete bipartite graphs.** Let $K_{m,n}$ denote the complete bipartite graph with $m+n$ nodes and $mn$ edges.

THEOREM 19. *$K_{m,n}$ is harmonious if and only if $m$ or $n = 1$.*

*Proof.* If $m$ or $n = 1$, the graph is a star and is harmonious (see § 5). Suppose $m > 1$ and $n > 1$. A harmonious labeling of $K_{m,n}$ is equivalent to a direct sum decomposition of $\mathbb{Z}_{mn} = A \oplus B$, where $A$ and $B$ are disjoint subsets of $\mathbb{Z}_{mn}$ with $|A| = m$, $|B| = n$. Since all the sums $a+b$ $(a \in A, b \in B)$ are distinct, so are all the differences $a-b$. But there are $mn$ differences, hence $0 = a-b$ must occur exactly once. Therefore $A$ and $B$ are not disjoint, and $K_{m,n}$ is not harmonious. Q.E.D.

The proof has an interesting corollary.

COROLLARY. *If $\mathbb{Z}_n = A \oplus B$ then $|A \cap B| = 1$.*

Although many papers have dealt with decompositions of this type ([12], [13], [16], [38], [71], [72]), this result does not seem to have been noticed before.

**12. The one-point union of two complete graphs.** The graph $K_n^{(2)}$ $(n \geqq 3)$ consists of two copies of $K_n$ sharing a common node, and contains $2n-1$ nodes and $n(n-1)$ edges (see Fig. 14). It is known that $K_n^{(2)}$ is never graceful [5].

THEOREM 20. *$K_n^{(2)}$ is harmonious if $n = 4$ but is not harmonious if $n$ is odd or $n = 6$.*

*Proof.* For $n = 4$ see Fig. 14, and for odd $n$ apply Theorem 11. The computer eliminated $n = 6$. Q.E.D.

We conjecture that $K_n^{(2)}$ is harmonious only when $n$ is 4.

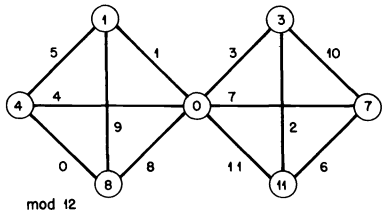



FIG. 14. *The graph $K_4^{(2)}$.*

**13. Some small graphs.**

THEOREM 21. *There are six connected graphs with $\leqq 5$ nodes that are not harmonious—see Fig.* 15.

*Proof.* It has already been shown that $C_4$, $F_2 = K_3^{(2)}$, $K_5$ and $K_{2,3}$ are not harmonious, and the other two graphs in Fig. 15 are easily eliminated by hand. Harmonious labelings of most of the other graphs with $\leqq 5$ vertices are given in Fig. 1, and the remainder are easily dealt with. Q.E.D.

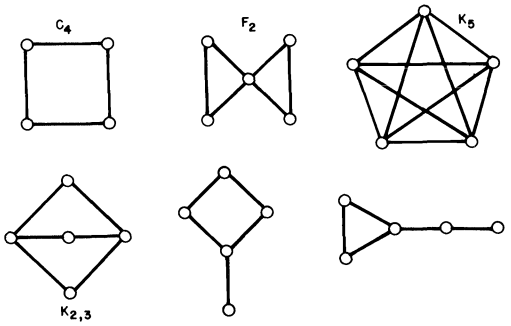



FIG. 15. *The six nonharmonious graphs with $\leqq 5$ nodes.*

For comparison we note that Golomb [25] showed there are three connected graphs with $\leqq 5$ nodes that are not graceful, namely $C_5$, $F_2$ and $K_5$; and Rao Hebbare [64] found that there are six nongraceful connected graphs with 6 nodes.

THEOREM 22. *The Petersen graph and the 1-skeletons of the tetrahedron, icosahedron and dodecahedron are harmonious, while the 1-skeletons of the cube and octahedron are not.*

*Proof.* For the first four see Figs. 16, 1(m), 17 and 18. The octahedron is not harmonious by Theorem 11, and the computer was used to check that the cube is not. Q.E.D.

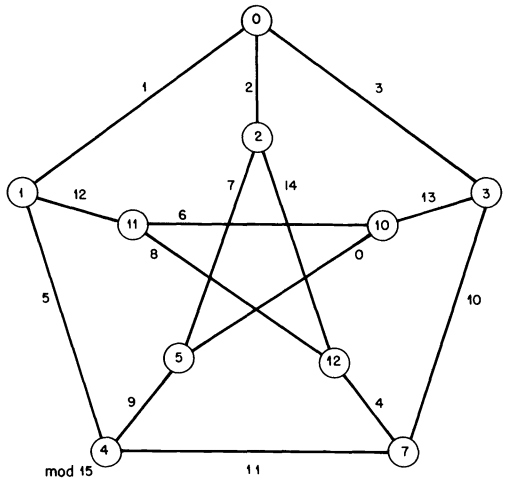



FIG. 16. *The Petersen graph.*

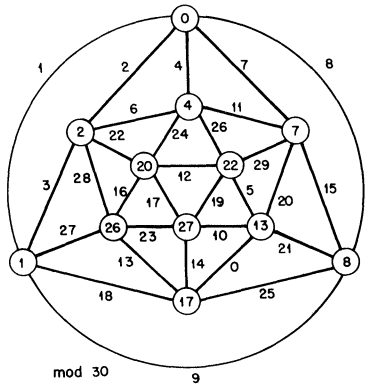



FIG. 17. *The icosahedron.*

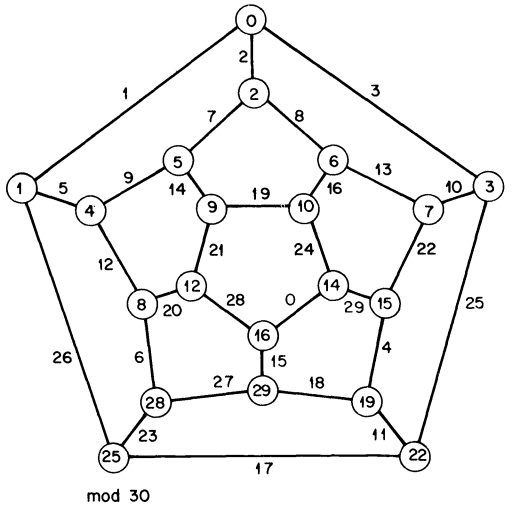



FIG. 18. *The dodecahedron.*

**14. Most graphs.**

THEOREM 23. *Almost all graphs are not harmonious.*

*Proof.* For our model of a random graph with $n$ nodes we assume that each of the $\binom{n}{2}$ possible edges independently exists or does not with probability $\frac{1}{2}$. Fix $\varepsilon \in (0, \frac{1}{2})$, and let $d$ be a fixed integer in the range $[(\frac{1}{2}-\varepsilon)\binom{n}{2}, (\frac{1}{2}+\varepsilon)\binom{n}{2}]$. We shall show that almost no graphs with $n$ nodes and $d$ edges are harmonious (as $n \to \infty$). Since almost all graphs with $n$ nodes have a number of edges in this range, the theorem follows.

There are $\binom{n(n-1)/2}{d}$ labeled graphs with $n$ nodes and $d$ edges, and so at least

$$\frac{1}{n!}\binom{n(n-1)/2}{d}$$

unlabeled graphs with $n$ nodes and $d$ edges.

Let $\lambda$ be a labeling of the $n$ nodes with distinct numbers from $\{0, 1, \cdots, d-1\}$. There are $d(d-1)\cdots(d-n+1) \leqq d^n$ such labelings. Let us consider how many graphs there are for which $\lambda$ is a harmonious labeling. Let $p_i$ be the number of pairs of nodes $\{v, v'\}$ with $\lambda(v)+\lambda(v') \equiv i$ (modulo $d$). Then

$$\sum_{i=0}^{d-1} p_i = \binom{n}{2}.$$

A graph is harmonious with this labeling if it consists of one edge taken from each of the classes counted by $p_i$. Thus there are

$$\prod_{i=0}^{d-1} p_i$$

labeled graphs for which $\lambda$ is a harmonious labeling. This product is maximized by taking the $p_i$'s as equal as possible; in particular

$$\prod_{i=0}^{d-1} p_i \leqq \left(\frac{n(n-1)}{2d}\right)^d.$$

Therefore there are at most

$$d^n\left(\frac{n(n-1)}{2d}\right)^d$$

harmonious labeled graphs. This is also an upper bound on the number of harmonious *unlabeled* graphs. To complete the proof we show that the ratio

$$\rho = \frac{d^n\left(\frac{n(n-1)}{2d}\right)^d}{\frac{1}{n!}\binom{n(n-1)/2}{d}}$$

approaches 0 when $n \to \infty$ and $d$ is in the required range. Write $d = (\frac{1}{2}-\mu)\binom{n}{2}$, with $\mu \in (-\frac{1}{2}, \frac{1}{2})$. Then

$$\rho < \frac{d^n n!\sqrt{8\binom{n}{2}(\frac{1}{2}-\mu)(\frac{1}{2}+\mu)}}{(\frac{1}{2}-\mu)^d 2^{\binom{n}{2} H_2[(1/2)-\mu]}},$$

where $H_2(x) = -x\log_2 x - (1-x)\log_2(1-x)$ (cf. [57, p. 309]). The denominator is

equal to

$$2^{-\binom{n}{2}[(1/2+\mu)]\log_2[(1/2)+\mu]}$$

and so $\rho \to 0$ as $n \to \infty$. Q.E.D.

The same argument establishes an unpublished result of Erdös (cf. [25]):

THEOREM 24. *Almost all graphs are not graceful.*

**15. Comparison of harmonious and graceful graphs.** A study of Table V suggests that the properties of being harmonious and graceful are roughly similar, although the entries for cycles show that the two properties are in general independent. Comparing the harmonious labelings of the previous sections with graceful labelings of the same graphs (see for example [10], [22], [44]) suggests that harmonious labelings are considerably more complicated. We know that $n_\gamma(v)$, the number of edges in the largest harmonious graph on $v$ nodes, is bounded by (2). On the other hand, if $g(v)$ denotes the number of edges in the largest graceful graph on $v$ nodes, it is known that $\lim_{v\to\infty} g(v)/v^2$ exists and satisfies

$$\tfrac{1}{3} \leqq \lim_{v\to\infty} \frac{g(v)}{v^2} \leqq 0.411, \tag{16}$$

(see [26], [42], [53], [58], [81]). Table VI compares the first few values of $n_\gamma(v)$ and $g(v)$: they are extremely close. We conclude therefore with an open problem: show that $\lim_{v\to\infty} n_\gamma(v)/v^2$ exists, and find improvements to (2) comparable with (16).

TABLE VI

*The size of the largest harmonious graph on v nodes ($n_\gamma(v)$) compared with the size of the largest graceful graph ($g(v)$). The values of $g(v)$ are taken from* [53] *and* [58].

| $v$ | $n_\gamma(v)$ | $g(v)$ |
|---|---|---|
| 2 | 1 | 1 |
| 3 | 3 | 3 |
| 4 | 6 | 6 |
| 5 | 9 | 9 |
| 6 | 13 | 13 |
| 7 | 17 | 17 |
| 8 | 24 | 23 |
| 9 | 30 | 29 |
| 10 | 36 | 36 |
| 11 | | 43 |
| 12 | | 50 |
| 13 | | 58 |
| 14 | | 68 |

**Acknowledgments.** We should like to thank P. Erdös and A. M. Odlyzko for helpful discussions, and F. R. K. Chung for proving Theorem 17 and parts of Theorem 18.

*Note added in proof.* Thom Grace (written communication, June 14, 1980) has shown that $g_{2m+1}$ is harmonious (see § 9): label the path $m, 0, m+1, 1, \cdots, m-1, 2m$, and the two additional nodes $3m$ and $5m+1$ (modulo $6m+2$). E. Levine (written communication, June 24, 1980) has shown that if $K_n^{(2)}$ is harmonious (see § 12) then $n$ is a sum of two squares.

## REFERENCES


[1] R. ALTER AND J. A. BARNETT, *Remarks on the postage stamp problem with applications to computers*, in Proc. 8th Southeastern Conference on Combinatorics, Graph Theory, Computing, Congress Num. XIX, Utilitas Math. Pub., Winnipeg, 1977, pp. 43–59.

[1a] ———, *A postage stamp problem*, Amer. Math. Monthly, 87 (1980), pp. 206–209.

[2] G. BARON, *Über Verallgemeinerungen des Langford'schen Problems*, in Combinatorial Theory and its Applications, P. Erdös et al., eds., Colloq. Math. Soc. János Bolyai, 4, North-Holland, Amsterdam, 3 vols., 1970, pp. 81–92.

[2a] L. D. BAUMERT, *Cyclic Difference Sets*, Lecture Notes in Math., 182, Springer-Verlag, Berlin, 1971.

[3] J.-C. BERMOND, *Problem*, in Combinatorics, A. Hajnal and V. T. Sós, eds., Colloq. Math. Soc. János Bolyai, 18, North-Holland, Amsterdam, 2 vols., 1978, p. 1190.

[4] J.-C. BERMOND, A. E. BROUWER AND A. GERMA, *Systèmes de triplets et différences associées*, Problèmes Combinatoires et Théorie des Graphs, Colloq. Intern. du Centre National de la Rech. Scient., 260, Editions du Centre Nationale de la Recherche Scientifique, Paris, 1978, pp. 35–38.

[5] J.-C. BERMOND, A. KOTZIG AND J. TURGEON, *On a combinatorial problem of antennas in radioastronomy*, in Combinatorics, A. Hajnal and V. T. Sós, eds., Colloq. Math. Soc. János Bolyai, 18, North-Holland, Amsterdam, 2 vols., 1978, pp. 135–149.

[6] M. R. BEST, A. E. BROUWER, F. J. MACWILLIAMS, A. M. ODLYZKO AND N. J. A. SLOANE, *Bounds for binary codes of length less than* 25, IEEE Trans. Information Theory, IT-24 (1978), pp. 81–93.

[7] G. S. BLOOM, *A chronology of the Ringel-Kotzig conjecture and the continuing quest to call all trees graceful*, Ann. N.Y. Acad. Sci., 326 (1979), pp. 32–51.

[8] G. S. BLOOM AND S. W. GOLOMB, *Numbered complete graphs, unusual rulers, and assorted applications*, in Theory and Applications of Graphs, Lecture Notes in Math., 642, Springer-Verlag, New York, 1978, pp. 53–65.

[9] ———, *Applications of numbered undirected graphs*, Proc. IEEE, 65 (1977), pp. 562–570.

[10] R. BODENDIEK, H. SCHUMACHER AND H. WEGNER, *Über graziöse Graphen*, Math.-Phys. Semesterberichte, 24 (1977), pp. 103–126.

[11] R. C. BOSE AND S. CHOWLA, *Theorems in the additive theory of numbers*, Comment. Math. Helvet., 37 (1962–63), pp. 141–147.

[12] N. G. DE BRUIJN, *On the factorization of finite Abelian groups*, Proc. Kon. Ned. Akad. Wetensch. Amsterdam, 56A (1953), pp. 258–264.

[13] ———, *On the factorization of cyclic groups*, ibid., pp. 370–377.

[14] ———, *On number systems*, Nieuw Archief voor Wiskunde, (3), IV (1956), pp. 15–17.

[15] J. CAHIT AND R. CAHIT, *On the graceful numbering of spanning trees*, Information Processing Letters, 3 (1975), pp. 115–118.

[16] L. CARLITZ AND L. MOSER, *On some special factorizations of* $(1-x^n)/(1-x)$, Canad. Math. Bull. 9 (1966), pp. 421–426.

[17] R. O. DAVIES, *On Langford's problem (II)*, Math. Gazette, 43 (1959), pp. 253–255.

[18] J. F. DILLON, *The generalized Langford-Skolem problem*, Proc. 4th Southeastern Conference on Combinatorics, Graph Theory, Computing, Utilitas Math. Pub., Winnipeg, 1973, pp. 237–247.

[19] P. ERDÖS AND R. L. GRAHAM, *Old and new problems and results in combinatorial number theory*, Monographies de l'Enseignement Math., to appear.

[20] P. ERDÖS AND H. HANANI, *On a limit theorem in combinatorical analysis*, Publ. Math. Debrecen, 10 (1963), pp. 10–13.

[21] P. ERDÖS AND P. TURÁN, *On a problem of Sidon in additive number theory, and some related problems*, J. London Math. Soc., 16 (1941), pp. 212–215 and 19 (1944), p. 208.

[22] R. FRUCHT, *Graceful numbering of wheels and related graphs*, Ann. N.Y. Acad. of Sci. 319 (1979), pp. 219–229.

[23] M. GARDNER, *Mathematical games: the graceful graphs of Solomon Golomb, or how to number a graph parsimoniously*, Scientific American, 226, 3 (1972), pp. 108–112; 226, 4 (1972), p. 104; 226, 6 (1972), p. 118.

[24] M. J. E. GOLAY, *Note on the representation of* $1, 2, \cdots, n$ *by differences*, J. London Math. Soc., 4 (1972), pp. 729–734.

[25] S. W. GOLOMB, *How to number a graph*, in Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1972, pp. 23–37.

[26] ———, *The largest graceful subgraph of the complete graph*, Amer. Math. Monthly, 81 (1974), 499–501.

[27] S. W. GOLOMB AND G. S. BLOOM, *Multifarious applications of numbered graphs*, in Proc. 2nd Caribbean Conference on Combinatorics and Computing, R. C. Read and C. C. Cadogan, eds., Univ. of West Indies, Cave Hill, Barbados, 1977, pp. 82–95.
[28] R. L. GRAHAM AND N. J. A. SLOANE, *Lower bounds for constant weight codes*, IEEE Trans. Information Theory, IT-26 (1980), pp. 37–43.
[29] ———, *On constant weight codes and harmonious graphs*, Utilitas Math., to appear.
[30] R. K. GUY, *Monthly research problems* 1969–73, Amer. Math. Monthly, 80 (1973), pp. 1120–1128.
[31] ———, *Monthly research problems* 1969–75, ibid., 82 (1975), pp. 995–1004.
[32] ———, *Packing* $[1, n]$ *with solutions of* $ax + by = cz$*: the unity of combinatorics*, in Teorie Combinatorie, Atti dei Convegni Lincei, 17, Vol. II, 1976, pp. 173–179.
[33] ———, *Monthly research problems* 1969–77, Amer. Math. Monthly, 84 (1977), pp. 807–815, and 85 (1978), p. 263.
[34] R. K. GUY AND V. L. KLEE, JR., *Monthly research problems* 1969–71, Amer. Math. Monthly, 78 (1971), pp. 1113–1122.
[35] A. GYÁRFÁS AND J. LEHEL, *A method to generate graceful trees*, Problèmes Combinatoires et Théorie des Graphs, Colloq. Intern. du Centre National de la Rech. Scient. 260, Éditions du Centre National de la Recherche Scientifique, Paris, 1978, pp. 207–209.
[36] N. HÄMMERER AND G. HOFMEISTER, *Zu einer Vermutung von Rohrbach*, J. Reine Angew. Math. 286/287 (1976), pp. 239–247.
[37] E. HÄRTTER, *Basen für Gitterpunktmengen*, J. Reine Angew. Math. 202 (1959), pp. 153–170.
[38] G. HAJÓS, *Sur la factorisation des groupes abéliens*, Časopis Pêst. Mat. Fys., 74 (1950), pp. 157–162.
[39] ———, *Sur le problème de factorisation des groupes cycliques*, Acta Math. Acad. Sci. Hungar., 1 (1950), pp. 189–195.
[40] H. HALBERSTAM AND K. F. ROTH, *Sequences*, Vol. 1, Oxford University Press, Oxford, 1966.
[40a] J. I. HALL, A. J. E. M. JANSEN, A. W. J. KOLEN AND J. H. VAN LINT, *Equidistant codes with distance* 12, Discrete Math., 17 (1977), pp. 71–83.
[41] H. HANANI, *A note on Steiner triple systems*, Math. Scand., 8 (1960), pp. 154–156.
[42] C. B. HASELGROVE AND J. LEECH, *Note on restricted difference bases*, J. London Math. Soc. 32 (1957), pp. 228–231.
[43] R. D. HEATH-BROWN AND H. IWANIEC, *On the difference between consecutive primes*, Bull Amer. Math. Soc., 1 (1979), pp. 758–760.
[43a] R. L. HEIMER AND H. LANGENBACH, *The stamp problem*, J. Recreational Math., 7 (1974), pp. 235–250.
[44] C. HOEDE AND H. KUIPER, *All wheels are graceful*, Utilitas Math., 14 (1978), p. 311.
[45] J. R. ISBELL, *Perfect addition sets*, Discrete Math., 24 (1978), pp. 13–18.
[46] W. KLOTZ, *Eine obere Schranke für die Reichweite einer Extremalbasis zweiter Ordnung*, J. Reine Angew. Math., 238 (1969), pp. 161–168.
[47] K. M. KOH, T. TAN AND D. G. ROGERS, *Two theorems on graceful trees*, Discrete Math., 25 (1979), pp. 141–148.
[48] A. KOTZIG, *On certain vertex-valuations of finite graphs*, Utilitas Math., 4 (1973), pp. 261–290.
[49] A. KOTZIG AND A. ROSA, *Magic valuations of finite graphs*, Canad. Math. Bull., 13 (1970), pp. 451–461.
[50] A. KOTZIG AND J. TURGEON, *β-Valuations of regular graphs with complete components*, in Combinatorics, A. Hajnal and V. T. Sós, eds., Colloq. Math. Soc. János Bolyai, 18, North-Holland, Amsterdam, 2 vols., 1978, pp. 697–703.
[51] C. W. H. LAM, *Nth power residue addition sets*, J. Combinatorial Theory, 20A (1976), pp. 20–33.
[52] C. D. LANGFORD, *Problem*, Math. Gazette, 42 (1958), p. 228.
[53] J. LEECH, *On the representation of* $1, 2, \cdots, n$ *by differences*, J. London Math. Soc., 31 (1956), pp. 160–169.
[54] ———, *Another tree labelling problem*, Amer. Math. Monthly, 82 (1975), pp. 923–925.
[55] E. LEVINE, *On the generalized Langford problem*, Fibonacci Quarterly, 6 (1968), pp. 135–138.
[56] W. F. LUNNON, *A postage stamp problem*, Computer J., 12 (1969), pp. 377–380.
[56a] D. MCCARTHY, R. C. MULLIN, P. J. SCHELLENBERG, R. G. STANTON, AND S. A. VANSTONE, *On approximations to a projective plane of order* 6, Ars Combinatoria 2 (1976), pp. 111–168.
[57] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.
[57a] M. MAHEO, *Strongly graceful graphs*, Discrete Math., 29 (1980), pp. 39–46.
[58] J. C. P. MILLER, *Difference bases: three problems in additive number theory*, in Computers in Number Theory, A. O. L. Atkin and B. J. Birch, eds., Academic Press, N.Y., 1971, pp. 299–322.
[59] L. MOSER, *On the representation of* $1, 2, \cdots, n$ *by sums*, Acta Arith., 6 (1960), pp. 11–13.

[60] L. MOSER, J. R. POUNDER AND J. RIDDELL, *On the cardinality of h-bases for n*, J. London Math. Soc., 44 (1969), pp. 397–407.
[61] M. B. NATHANSON, *Additive h-bases for lattice points*, Second International Conference on Combinatorial Mathematics, Ann. N.Y. Acad. Sci., 319 (1979), pp. 413–414.
[62] R. NOWAKOWSKI, *Generalizations of the Langford-Skolem problem*, M.Sc. Thesis, University of Calgary, 1975.
[63] C. J. PRIDAY, *On Langford's problem* (I), Math. Gazette, 43 (1959), pp. 250–253.
[64] S. P. RAO HEBBARE, *Graceful cycles*, Utilitas Math. 10 (1976), pp. 307–317.
[65] H. ROHRBACH, *Ein Beitrag zur additiven Zahlentheorie*, Math. Zeit., 42 (1937), pp. 1–30.
[66] ———, *Anwendung eines Satzes der additiven Zahlentheorie auf eine grappentheoretische Frage*, ibid., pp. 538–542.
[67] A. ROSA, *On certain valuations of the vertices of a graph*, Theory of Graphs (Internat. Symposium, Rome, July 1966), Gordon and Breach, N.Y. and Dunod, Paris, 1967, pp. 349–355.
[68] D. P. ROSELLE, *Distributions of integers into s-tuples with given differences*, Proc. Manitoba Conf. Number. Math., Dept. of Computer Science, University of Manitoba, Winnipeg, 1971, pp. 31–42.
[69] ———, *Comments and complements*, Amer. Math. Monthly, 81 (1974), pp. 1097–1099.
[70] D. P. ROSELLE AND T. C. THOMASSON, JR, *On generalized Langford sequences*, J. Combinatorial Theory, 11A (1971), pp. 196–199.
[71] A. D. SANDS, *On the factorisation of finite abelian groups*, Acta Math. Acad. Sci. Hungar., 8 (1957), pp. 65–86 and 13 (1962), pp. 153–169.
[72] ———, *The factorization of Abelian groups*, Quart. J. Math. Oxford (2), 10 (1959), pp. 81–91.
[73] J. SCHÖNHEIM, *On maximal systems of k-tuples*, Stud. Sci. Math. Hungar., 1 (1966), pp. 363–368.
[74] G. J. SIMMONS, *Synch-sets: a variant of difference sets*, Proc. 5th Southeastern Conference on Combinatorics, Graph Theory and Computing, Utilitas Math. Pub. Co., Winnipeg, 1974, pp. 625–645.
[75] T. SKOLEM, *On certain distributions of integers in pairs with given differences*, Math. Scand., 5 (1957), pp. 57–68.
[76] R. G. STANTON, J. A. BATE AND R. C. MULLIN, *Some tables for the postage stamp problem*, Proc. Fourth Manitoba Conference on Numerical Math., Utilitas Math. Pub., Winnipeg, 1974, pp. 351–356.
[77] R. G. STANTON, J. G. KALBFLEISCH AND R. C. MULLIN, *Covering and packing designs*, Proc 2nd Chapel Hill Conference on Combinatorical Mathematics and its Applications, R. C. Bose et al., eds., Chapel Hill, 1970, pp. 428–450.
[78] R. G. STANTON AND C. R. ZARNKE, *Labelling of balanced trees*, Proc. 4th Southeastern Conference on Combinatorics, Graph Theory, Computing, Utilitas Math. Pub., Winnipeg, 1973, pp. 479–495.
[79] A. STÖHR, *Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe*, J. Reine Angew. Math., 194 (1955), pp. 40–65 and 111–140.
[80] H. TAYLOR, *Odd path sums in an edge-labeled tree*, Math. Mag., 50 (1977), pp. 258–259.
[81] B. WICHMANN, *A note on restricted difference bases*, J. London Math. Soc., 38 (1962), pp. 465–466.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104-0005 $01.00/0

# ON UNIMODALITY FOR LINEAR EXTENSIONS OF PARTIAL ORDERS*

F. R. K. CHUNG†, P. C. FISHBURN† AND R. L. GRAHAM†

**Abstract.** R. Rivest has recently proposed the following intriguing conjecture: Let $x^*$ denote an arbitrary fixed element in an $n$-element partially ordered set $P$, and for each $k$ in $\{1, 2, \cdots, n\}$ let $N_k$ be the number of order-preserving maps from $P$ onto $\{1, 2, \cdots, n\}$ that map $x^*$ into $k$. Then the sequence $N_1, \cdots, N_n$ is unimodal. This note proves the conjecture for the special case in which $P$ can be covered by two linear orders. It also generalizes this result for $P$ that have disjoint components, one of which can be covered by two linear orders.

**1. Introduction.** Given a finite partially ordered set $(P, \prec)$, where $\prec$ is asymmetric, we say that an injection $\lambda$ from $P$ into the set $Z$ of integers is a *linear extension* of $P$ if, for all $x, y \in P$,

$$x \prec y \Rightarrow \lambda(x) < \lambda(y).$$

We shall presume that $P$ has $n$ elements and, in the main part of the paper, restrict ourselves to bijections $\lambda : P \to [n] \equiv \{1, 2, \cdots, n\}$. Generalizations are discussed later.

Let $x^*$ be an arbitrary fixed element in $P$. For each $k \in [n]$, define $N_k$ to be the number of linear extensions $\lambda : P \to [n]$ for which $\lambda(x^*) = k$. Rivest [2] has proposed the following tantalizing conjecture.

CONJECTURE. *The sequence $N_k$, $k \in [n]$, is unimodal.*

By unimodal we mean that, for all $1 \leqq i < j < k \leqq n$,

$$N_j \geqq \min\{N_i, N_k\}.$$

In this note we shall prove that the conjecture is valid for the important class of partially ordered sets that can be partitioned into two linearly ordered subsets, i.e., *chains*, with $\prec$-pairs allowed between the chains. In fact, we show that the $N_k$'s in this case satisfy the stronger property of logarithmic concavity, i.e.,

$$N_k^2 \geqq N_{k-1} N_{k+1} \quad \text{for } 1 < k < n.$$

A similar proof provides an interesting result involving the unimodality of certain sequences of integers.

**2. Lattice paths in $Z^2$.** We shall say that the partially ordered set $(P, \prec)$ *can be covered by two chains* if there is a partition $\{A, B\}$ of $P$ such that the restriction of $\prec$ on each of $A$ and $B$ is a linear order. To avoid the trivial case, we shall suppose that $\prec$ on $P$ is not linear, and that $(P, \prec)$ can be covered by two chains, denoted as $A = \{a_1 \prec \cdots \prec a_r\}$ and $B = \{b_1 \prec \cdots \prec b_s\}$, with $r \geqq 1$, $s \geqq 1$ and $r + s = n$. There can be "cross-relations" like $a_i \prec b_j$ or $b_j \prec a_i$ from $(P, \prec)$, but in any event $\prec$ must be asymmetric ($x \prec y \Rightarrow$ not $(y \prec x)$) and transitive.

Let $L$ denote the set of all ordered pairs of nonnegative integers. Each linear extension $\lambda : P \to [n]$ induces maps of $A$ and $B$ into $[n]$, with $\lambda(a_1) < \cdots < \lambda(a_r)$ and $\lambda(b_1) < \cdots < \lambda(b_s)$. To each such $\lambda$ we will associate a lattice path $\pi(\lambda)$ in $L$ as follows. The first point on $\pi(\lambda)$ is $(0, 0)$. If the $k$th point on $\pi(\lambda)$ is $(x_k, y_k)$ and if $\lambda(p) = k + 1$, then the $(k+1)$st point on $\pi(\lambda)$ is $(x_k + 1, y_k)$ if $p \in A$, and $(x_k, y_k + 1)$ if $p \in B$. The terminal point on $\pi(\lambda)$ is $(r, s)$. An example appears in Fig. 1.

* Received by the editors March 19, 1980.
† Bell Telephone Laboratories, Murray Hill, New Jersey 07974.

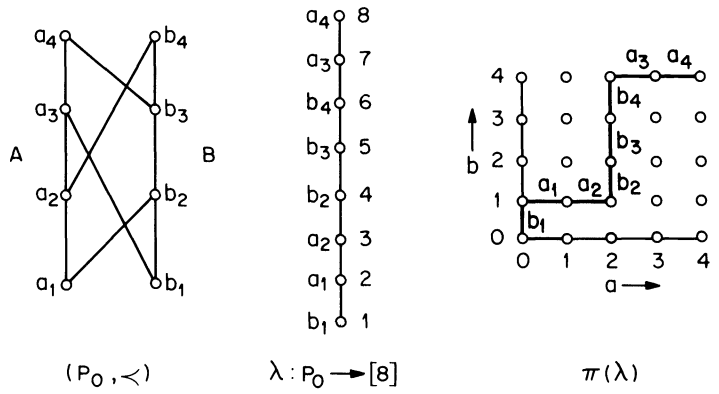



FIG. 1. *The correspondence between* $\lambda$ *and* $\pi(\lambda)$.

The fact that $\lambda$ preserves the linear orders on $A$ and $B$ is reflected in the fact that the indices of the $a_i$ and $b_j$ increase as we move along $\pi(\lambda)$ from $(0, 0)$ to $(r, s)$. But how do the *other* $<$-pairs show up in $\pi(\lambda)$? For Fig. 1, what constraint does $a_1 < b_2$ (which forces $\lambda(a_1) < \lambda(b_2)$) place on $\pi(\lambda)$? The answer is very simple. Each $a_i < b_j$ corresponds to a rectangular "barrier" which the path $\pi(\lambda)$ is not allowed to penetrate. This barrier is defined to be all lattice points $(x, y)$ in $L$ for which $x \leqq i$ and $y \geqq j - 1$, as illustrated in Fig. 2.

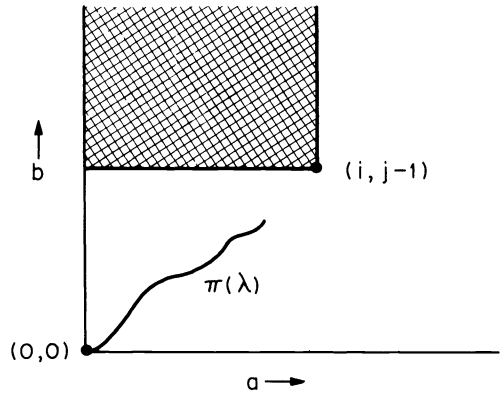



FIG. 2. *The barrier for* $a_i < b_j$.

The barrier for $a_i < b_j$ forces $\pi(\lambda)$ to reach a lattice point with $x$-coordinate $i$ *before* it reaches one with $y$-coordinate $j$, i.e., $a_i$ occurs before $b_j$ on $\pi(\lambda)$. This is precisely what is needed for $\lambda(a_i) < \lambda(b_j)$.

In a similar manner, $b_j < a_i$ corresponds to a rectangular barrier consisting of all $(x, y)$ in $L$ for which $x \geqq i - 1$ and $y \leqq j$. For $\lambda$ to be a linear extension of $P$, $\pi(\lambda)$ must not penetrate *any* of the barriers formed from the cross-relations in $(P, <)$. Fig. 3 shows the union of the barriers for $(P_0, <)$ from Fig. 1.

The next point we consider is how $\lambda(x^*) = k$ is reflected in $\pi(\lambda)$. Without loss of generality, we assume that $x^* = a_i$, so that $x^* \in A$. Then it is easy to see that $\lambda(a_i) = k$ iff $\pi(\lambda)$ contains the two points $(i - 1, k - i)$ and $(i, k - i)$. (Similarly, $\lambda(b_j) = k$ iff $\pi(\lambda)$ contains $(k - j, j - 1)$ and $(k - j, j)$.)

Suppose $N_{k-1}$ and $N_{k+1}$ are both positive, and let $\lambda^+$ and $\lambda^-$ be linear extensions of $P$ such that $\lambda^+(a_i) = k + 1$ and $\lambda^-(a_i) = k - 1$. Thus, $\pi(\lambda^+)$ contains points $(i - 1, k + 1 -$

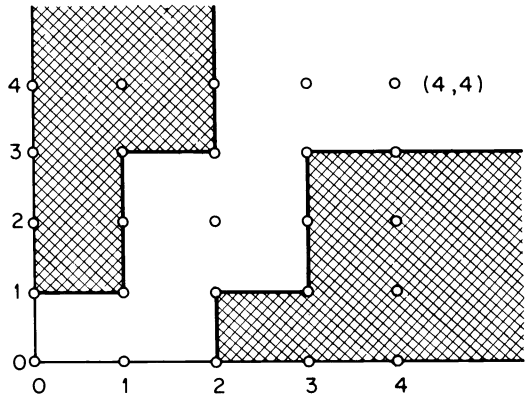



FIG. 3. *The union of barriers for* $(P_0, <)$.

$i)$ and $(i, k+1-i)$, and $\pi(\lambda^-)$ contains $(i-1, k-1-i)$ and $(i, k-1-i)$. Let $x_0$ be the largest integer that is $\leqq i-1$ such that, for some $y$, $(x_0, y+1)$ is on $\pi(\lambda^+)$ and $(x_0, y)$ is on $\pi(\lambda^-)$, and let $y_0$, which cannot exceed $k-1-i$, be the largest integer such that $(x_0, y_0+1)$ is on $\pi(\lambda^+)$ and $(x_0, y_0)$ is on $\pi(\lambda^-)$. Similarly, let $x_1$ be the smallest integer $\geqq i$ such that, for some $y$, $(x_1, y+1)$ is on $\pi(\lambda^+)$ and $(x_1, y)$ is on $\pi(\lambda^-)$, and let $y_1$, which cannot be less than $k-i$, be the smallest integer such that $(x_1, y_1+1)$ is on $\pi(\lambda^+)$ and $(x_1, y_1)$ is on $\pi(\lambda^-)$.

We now form two new lattice paths $\pi(\lambda_1)$ and $\pi(\lambda_2)$ as follows. Let $\pi(\lambda_1)$ consist of the points on $\pi(\lambda^-)$ from $(0, 0)$ to $(x_0, y_0)$, plus the points on $\pi(\lambda^+)$ from $(x_0, y_0+1)$ to $(x_1, y_1+1)$ translated by $-1$ in the $y$-direction, plus the points on $\pi(\lambda^-)$ from $(x_1, y_1)$ to $(r, s)$. Let $\pi(\lambda_2)$ consist of the points on $\pi(\lambda^+)$ from $(0, 0)$ to $(x_0, y_0+1)$, plus the points on $\pi(\lambda^-)$ from $(x_0, y_0)$ to $(x_1, y_1)$ translated by $+1$ in the $y$-direction, plus the points on $\pi(\lambda^+)$ from $(x_1, y_1+1)$ to $(r, s)$. It is of course possible to have $\pi(\lambda_1)=\pi(\lambda_2)$, or, equivalently, $\lambda_1=\lambda_2$, but this will not affect our conclusions. We observe that:

(i) $\pi(\lambda_1)$ and $\pi(\lambda_2)$ are lattice paths from $(0, 0)$ to $(r, s)$ which contain $(i, k-i)$ and $(i-1, k-i)$, and, therefore, $\lambda_1(a_i)=\lambda_2(a_i)=k$;

(ii) since $\pi(\lambda^+)$ lies strictly above $\pi(\lambda^-)$ in the region where the translations occur in the construction, neither $\pi(\lambda_1)$ nor $\pi(\lambda_2)$ penetrates any of the barriers formed by $(P, <)$. It follows that $\lambda_1$ and $\lambda_2$ are linear extensions of $P$;

(iii) if two ordered pairs of the form $(\lambda^+, \lambda^-)$ are distinct, then their associated $(\lambda_1, \lambda_2)$ pairs are distinct. This follows from the construction: if two $(\pi(\lambda^+), \pi(\lambda^-))$ differ prior to $i$ on the abscissa, then their associated $(\pi(\lambda_1), \pi(\lambda_2))$ will differ before $i$; if two $(\pi(\lambda^+), \pi(\lambda^-))$ differ after $i-1$, then their associated $(\pi(\lambda_1), \pi(\lambda_2))$ will differ after $i-1$.

Thus, our construction provides an injection from the ordered pairs $(\lambda^+, \lambda^-)$ into pairs $(\lambda_1, \lambda_2)$, where $\lambda^+$ and $\lambda^-$ are any linear extensions of $P$ for which $\lambda^+(a_i)=k+1$ and $\lambda^-(a_i)=k-1$, and $\lambda_1$ and $\lambda_2$ are linear extensions of $P$ that satisfy $\lambda_1(a_i)=\lambda_2(a_i)=k$. If $\alpha$, $\beta$ and $\gamma$ are the number of linear extensions of $P$ for which $\lambda(a_i)=k+1$, $\lambda(a_i)=k-1$, and $\lambda(a_i)=k$, respectively, then such an injection requires $\gamma^2 \geqq \alpha\beta$, for otherwise two $(\lambda_1, \lambda_2)$ pairs associated with distinct $(\lambda^+, \lambda^-)$ pairs would have to be identical.

The preceding argument applies analogously when $x^*=b_j$. Thus, we have proved the following result.

THEOREM 1. *Let $x^*$ be a fixed element in a partially ordered set $(P, <)$ on $n$ elements, and suppose $(P, <)$ can be covered by two chains. For $k \in \{1, 2, \cdots, n\}$, let $N_k$ be the*

*number of linear extensions* $\lambda : P \to \{1, 2, \cdots, n\}$ *for which* $\lambda(x^*) = k$. *Then*

$$N_k^2 \geqq N_{k-1}N_{k+1} \quad \text{for } k = 2, \cdots, n-1.$$

COROLLARY. *Given the hypotheses of Theorem* 1, *the sequence* $N_1, N_2, \cdots, N_n$ *is unimodal.*

The same basic argument for Theorem 1 can be used to prove the following result for sequences of integers. Let $A = (a_1 \geqq a_2 \geqq \cdots)$ be a nonincreasing sequence of nonnegative integers. Given $A$, let $S_n$ be the number of nonincreasing sequences $x = (x_1 \geqq x_2 \geqq \cdots \geqq x_n)$ of integers for which $0 \leqq x_k \leqq a_k$, for $k = 1, \cdots, n$.

THEOREM 2. *The sequence* $S_1, S_2, \cdots$ *is logarithmically concave, i.e.,*

$$S_n^2 \geqq S_{n-1}S_{n+1} \quad \textit{for all } n \geqq 2.$$

When $A$ is constant, say $A = (t, t, t, \cdots)$, Theorem 2 shows the (easily proved) logarithmic concavity of the binomial coefficients $\binom{t+k}{k}$ for $k = 1, 2, \cdots$.

**3. A generalization.** We now generalize our analysis of logarithmic concavity by considering disjoint partial orders along with linear extensions that map $P$ into $[m] \equiv \{1, \cdots, m\}$ when $m$ exceeds the cardinality of $P$. The following lemma provides a basis for the generalization.

LEMMA. *Let* $(P, \prec)$ *and* $(P \cup C, \prec)$ *be partially ordered sets on* $n$ *and* $n + \alpha$ *elements, respectively, that have the same ordered pairs in their partial orders with* $C \cap P = \varnothing$. *Let* $x^* \in P$ *be fixed, and let* $N_k$ *and* $N'_k$, *respectively, be the number of linear extensions* $\lambda : P \to [n]$ *and* $\lambda' : P \cup C \to [n + \alpha]$ *that have* $\lambda(x^*) = k$ *and* $\lambda'(x^*) = k$. *If* $N_1, \ldots, N_n$ *is logarithmically concave, then so is* $N'_1, \cdots, N'_{n+\alpha}$.

If $C$ is empty, there is nothing to prove; so suppose initially that $C = \{c\}$, with $\alpha = 1$. Since neither $c \prec x$ nor $x \prec c$ for each $x \in P$, each $\lambda$ for $P$ generates $n + 1$ $\lambda'$ for $P \cup \{c\}$ according to the $n + 1$ placements of $c$. With $N_0 = N_{n+1} = 0$,

$$N'_k = (k-1)N_{k-1} + (n-k+1)N_k \quad \text{for } k = 1, \cdots, n+1.$$

Using this relationship, $(N'_k)^2 - N'_{k-1}N'_{k+1}$, for $2 \leqq k \leqq n$, reduces to

$$\begin{aligned} k(k-2)[N_{k-1}^2 - N_{k-2}N_k] + (n-k)(n-k+2)[N_k^2 - N_{k-1}N_{k+1}] \\ + (k-2)(n-k)[N_{k-1}N_k - N_{k-2}N_{k+1}] + (N_{k-1} - N_k)^2, \end{aligned}$$

which must be nonnegative if $\{N_k\}$ is logarithmically concave.

This completes the proof of the lemma if $\alpha \leqq 1$, so suppose in this paragraph that $\alpha \geqq 2$ with $C = \{c_1, \cdots, c_\alpha\}$. The $\lambda' : P \cup C \to [n + \alpha]$ can be generated from the $\lambda : P \to [n]$ by adding one $c_i$ at a time. For a given $\lambda$, we first add $c_1$ to obtain $n + 1$ linear extensions from $P \cup \{c_1\}$ onto $[n + 1]$; for each of these $n + 1$, we then add $c_2$ to obtain $n + 2$ linear extensions from $P \cup \{c_1, c_2\}$ onto $[n + 2]$; and so forth. If $\{N_m\}$ is logarithmically concave, then successive applications of the result obtained in the preceding paragraph for each $c_i$ addition show that $\{N'_k\}$ must be logarithmically concave. The lemma is thus proved.

We now state our generalization, discuss its features, and then conclude this section with its proof.

THEOREM 3. *Suppose* $(P_1, \prec_1)$, $(P_2, \prec_2)$ *and* $(P, \prec)$ *are partially ordered sets on* $n_1$, $n_2$ *and* $n$ *elements respectively such that* $0 < n_1 \leqq n$, $P_1 \cup P_2 = P$, $P_1 \cap P_2 = \varnothing$ *and* $\prec_1 \cup \prec_2 = \prec$. *Let* $x^* \in P_1$ *be fixed, and let* $N_k$ $(k = 1, \cdots, n_1)$ *be the number of linear extensions* $\lambda : P_1 \to [n_1]$ *for which* $\lambda(x^*) = k$. *In addition, given* $m \geqq n$, *let* $M_k$ $(k =$

$1, \cdots, m)$ *be the number of linear extensions* $\lambda^*: P \to [m]$ *for which* $\lambda^*(x^*) = k$. *If* $N_1, \cdots, N_{n_1}$ *is logarithmically concave, then so is* $M_1, \cdots, M_m$.

When $n_2 = 0$ and $m > n$, this shows that logarithmic concavity for $\lambda: P \to [n]$ carries over to $\lambda^*: P \to [m]$. When $n_2 > 0$ and $m = n$, Theorem 3 says that logarithmic concavity for the elements within a part of $(P, <)$, namely $(P_1, <_1)$, carries over to all of $(P, <)$ for those same elements, provided that the rest of $(P, <)$ is not connected to the first part. The combination of these two cases provides the generalization stated in the theorem.

Theorems 1 and 3 together yield the following result.

THEOREM 4. *If an n-element partially ordered set* $(P, <)$ *can be partitioned into partially ordered sets* $(P_1, <_1)$ *and* $(P_2, <_2)$ *with no* $<$*-connection between* $P_1$ *and* $P_2$, *if* $(P_1, <_1)$ *can be covered by two chains, and if* $x^* \in P_1$, $m \geqq n$, *and* $M_k$ *is the number of linear extensions* $\lambda: P \to [m]$ *for which* $\lambda(x^*) = k$, *then* $M_1, \cdots, M_m$ *is logarithmically concave and unimodal.*

We now sketch the proof of Theorem 3 using the notation in its statement. In addition, let $T_k$ be the number of linear extensions $\lambda_0: P \to [n]$ for which $\lambda_0(x^*) = k$, and if $n_2 > 0$, let $\beta$ be the number of linear extensions $\lambda_2: P_2 \to [n_2]$, and let $N'_k$ be the number of linear extensions $\lambda': P_1 \cup C \to [n]$ that have $\lambda'(x^*) = k$ when $C$ is a completely unordered $n_2$-element set (see the lemma) that is disjoint from $P_1$.

If $n_2 = 0$ then $T_k = N_k$, so assume henceforth in this paragraph that $n_2 > 0$. We shall apply the lemma with $\alpha = n_2$. Consider a fixed $\lambda_2: P_2 \to [n_2]$ along with a generic $\lambda_1: P_1 \to [n]$. The $n_2$ numbers in $[n]$ that are not in $\lambda_1(P_1)$ can be bijectively assigned to the elements in $P_2$ in exactly one way that preserves the $\lambda_2$ order and yields a $\lambda_0: P \to [n]$—as compared to the $n_2!$ ways this could be done for the unordered set $C$. Since this is true for each such $\lambda_1$, it follows that the number of $\lambda_0: P \to [n]$ that have $\lambda_0(x^*) = k$ *and* have $P_2$ in its $\lambda_2$ order is $N'_k/n_2!$. Since there are $\beta$ such $\lambda_2$, $T_k = \beta N'_k/N_2!$. If $N_1, \cdots, N_{n_1}$ is logarithmically concave, then the lemma says that $T_1, \cdots, T_n$ is too.

This proves Theorem 3 if $m = n$. If $m > n$, we reapply the lemma with $\alpha = m - n$. In this case let $C'$ be a completely unordered $(m-n)$-element set disjoint from $P$ and, with respect to $(P \cup C', <)$, let $T'_k$ be the number of linear extensions $\lambda': P \cup C' \to [m]$ for which $\lambda'(x^*) = k$. By the lemma, if $\{T_k\}$ is logarithmically concave then so is $\{T'_k\}$. Since the $m-n$ numbers in $[m]$ that aren't in a $\lambda'(P)$ can be bijectively assigned to $C'$ in $(m-n)!$ ways, it follows that $M_k$ as defined in Theorem 3 equals $T'_k/(m-n)!$. When this is combined with preceding conclusions, we see that if $N_1, \cdots, N_{n_1}$ is logarithmically concave, then so is $M_1, \cdots, M_m$.

**4. Concluding remarks.** The preceding techniques can be used to prove other unimodality results for restricted lattice path problems. For example, consider lattice paths $\pi$ that are not allowed to penetrate barriers of the type shown in Fig. 3, so that $\pi$ is bounded between two increasing staircases. Let $D_{n,k}$ be the number of such paths that go through point $(k, n-k)$. Then, for each $n$, the sequence $D_{n,k}$, $0 \leqq k \leqq n$, is logarithmically concave and therefore unimodal. (Of course, here we are just looking at the intersections of lattice paths with the line $x + y = n$.) The reader is referred to the recent paper of Graham, Yao, and Yao [1] for similar applications of these ideas.

Finally, we note another open conjecture that is suggested by our analysis. Within the context used for the earlier conjecture, we propose:

CONJECTURE*. *The sequence* $N_k$, $k \in [n]$, *is logarithmically concave.*

Conjecture* is stronger than Rivest's Conjecture since unimodality follows from logarithmic concavity, but not conversely. Thus, a counterexample for Conjecture* need not disprove unimodality, while verification of Conjecture* would establish Rivest's Conjecture.

*Note added in proof.* R. Stanley has just proved Conjecture* using a very ingenious application of the Alexandroff-Fenchel theorem (which guarantees the logarithmic concavity of certain coefficients arising from the volume of weighted sums of $n$-dimensional polytopes).

## REFERENCES

[1] R. L. GRAHAM, A. C. YAO, AND F. F. YAO, *Some monotonicity properties of partial orders*, this Journal, 1 (1980), pp. 251–258.

[2] R. RIVEST (personal communication).

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104 0006$01.00/0

# OBTAINING SPECIFIED IRREDUCIBLE POLYNOMIALS OVER FINITE FIELDS*

SOLOMON W. GOLOMB†

**Abstract.** In numerous applications, it is necessary to find an irreducible polynomial $f(x)$ of degree $n$ over $GF(q)$ whose roots are primitive $d$th roots of unity. (Here $d$ must divide $q^n-1$.) Let $\alpha$ be one such root. A direct method is to write

$$f(x)=\prod_{i=0}^{n-1}(x-\alpha^{q^i})=\sum_{j=0}^{n}(-1)^j C_j x^{n-j},$$

where $C_0=1$ and all $C_j$ are in $GF(q)$. Explicitly, $C_j$ is the sum of all powers of $\alpha$ whose exponents, written as $n$-digit numbers in base $q$, look like binary numbers of weight $j$. Formulas for the number of such polynomials $f(x)$ are given, several computational shortcuts exploiting properties of cyclotomic polynomials are noted, and numerous illustrative examples are presented.

**1. Introduction.** There are numerous applications in which it is necessary to find an irreducible polynomial, $f(x)$, of degree $n$, over the finite field $GF(q)$, of $q$ elements, whose roots are primitive $d$th roots of unity. (A necessary condition for such an $f(x)$ to exist is that $d$ be a divisor of $q^n-1$.) Such applications include the generation of algebraic error-correcting codes with specified properties ([1]), the design of $n$-stage linear shift register sequences over $GF(q)$ with period $d$ ([2]), and the construction of fast transforms based on finite field properties ([3]).

Various tables of irreducible polynomials over finite fields, and especially over $GF(2)$, have been published ([1], [2], [4], [14]). A variety of methods are employed in the construction of such tables, including a close analog of the "sieve method" used in making tables of prime numbers. However, if one only wishes to obtain polynomials for a given set of values of $n$, $q$, and $d$, and no tables are available, then a direct algebraic method may be used. This article is an attempt to explain and illustrate this method, which is scarcely mentioned in the existing literature. (It is described briefly as the "synthetic method" in [2, pp. 66–69].) For a general background to the algebra involved, one may consult [1], [2], or [5].

**2. Methodology.** If $f(x)$ is an irreducible polynomial of degree $n>1$ over $GF(q)$, then the roots of $f(x)$ are primitive $d$th roots of unity for some divisor $d$ of $q^n-1$, where $d$ does not divide $q^m-1$ for any $m$, $1\leqq m<n$. The number of irreducible polynomials over $GF(q)$ of degree $n$ and "primitivity" $d$ is then $\phi(d)/n$, where $\phi$ is Euler's totient function, since there are $\phi(d)$ primitive $d$th roots of unity and these occur $n$ at a time as roots of irreducible polynomials of degree $n$.

If $\alpha$ is a root of $f(x)$, then the complete set of roots of $f(x)$ is $\{\alpha, \alpha^{q^1}, \alpha^{q^2}, \alpha^{q^3}, \cdots, \alpha^{q^{n-1}}\}$. Hence we may write:

$$f(x)=\prod_{i=0}^{n-1}(x-\alpha^{q^i})=x^n-C_1x^{n-1}+C_2x^{n-2}-C_3x^{n-3}+\cdots+(-1)^nC_n,$$

where the coefficients $C_j$ must be elements of the ground field, $GF(q)$. Clearly

$$C_1=\sum_{i=0}^{n-1}\alpha^{q^i}$$

* Received by the editors September 21, 1979, and in revised form February 5, 1980. This research was supported in part by the U.S. Air Force Office of Scientific Research under grant AFOSR 75-2798.

† Departments of Electrical Engineering and Mathematics, University of Southern California, University Park, Los Angeles, California 90007.

is the sum of the roots of $f(x)$, also called the "trace" of $f(x)$, or Tr $(\alpha)$. The powers of $\alpha$ which are summed have as exponents the integers $1, q, q^2, \cdots, q^{n-1}$, which are all the numbers which, when written as $n$-digit numbers in base $q$, consist of all cyclic permutations of $(000\cdots01)$.

Similarly, $C_j$ is the sum of all powers of $\alpha$ having exponents which, when written as $n$-digit numbers in base $q$, look like binary numbers of weight $j$.

If $d$ is a proper divisor of $q^n-1$, then the exponents on $\alpha$ are effectively reduced modulo $d$, since $\alpha^d = 1$. However there are computational advantages to determining the exponents first modulo $q^n-1$, and then reducing them modulo $d$.

Note that

$$C_n = \prod_{i=0}^{n-1} \alpha^{q^i} = \alpha^{(q^n-1)/(q-1)},$$

which is a nonzero element of $GF(q)$. Let $\sigma = (q^n-1)/(q-1)$. Then

$$C_{n-1} = \sum_{i=0}^{n-1} \alpha^{\sigma - q^i} = \alpha^{\sigma} \sum_{i=0}^{n-1} \left(\frac{1}{\alpha}\right)^{q^i} = C_n \cdot \text{Tr}\,(\alpha^{-1}).$$

This symmetry between $C_{n-1}(\alpha)$ and $C_1(1/\alpha)$ generalizes to a symmetry between $C_{n-j}$ and $C_j$ for all $j$.

The total number of irreducible polynomials of degree $n$ over $GF(q)$ is

$$\frac{1}{n}\sum_{t|n} \mu(t) q^{n/t},$$

where the summation is over all divisors $t$ of $n$, and $\mu$ is the Möbius function. This is also the number of "primitive" necklaces consisting of $n$ beads in $q$ colors, where a *primitive* necklace is one with no periodic substructure. The natural correspondence between polynomials and necklaces is the one established by the cyclic permutations of the base-$q$ digits of some power $\alpha^k$ of $\alpha$. These permutations of the digits correspond on the one hand to the necklaces, and on the other hand to the traces of the irreducible polynomials of degree $n$ over $GF(q)$.

The total number of necklaces of $n$ beads in $q$ colors (primitive or not) is

$$\frac{1}{n}\sum_{d|n} \phi(d) q^{n/d},$$

which is also the number of irreducible polynomials of all degrees $d$ which divide $n$ over $GF(q)$. Equivalently, this is the number of irreducible polynomial factors of $x^{q^n}-x$ over $GF(q)$, which is the number of different minimal polynomials over $GF(q)$ having all elements of $GF(q^n)$ as roots.

Let $\zeta = e^{2\pi i/m}$. Then the *cyclotomic polynomial*

$$\Phi_m(x) = \prod_{\substack{1 \le i \le m \\ (i,m)=1}} (x - \zeta^i),$$

is the polynomial whose roots are the primitive $m$th roots of unity over $Q$, the field of rationals. It is known ([6]) that $\Phi_m(x)$ is irreducible over $Q$ for all $m$, but as we have seen, $\Phi_m(x)$ frequently factors over $GF(q)$. Gauss [7] was the first to study sums of the type

$$\sum_{i=0}^{n-1} \zeta^{q^i},$$

but tended to think of these solely as complex numbers. He called these sums "periods" of the cyclotomic equation, and derived many of their properties. It was Kummer [8], [9] who studied the problem of interpreting these sums as elements of $GF(q)$, and a table of evaluations (containing a disturbing number of errors) was published by Reuschle [10].

To avoid confusion with all the other periodic phenomena which are present, these sums have also been called *cyclotomic cosets* ([2], [11]), and are the basis vectors of an algebra ([12]). (The term *cyclotomic coset* is used both for the set of exponents, and for the sum of $\alpha$ raised to each of these exponents. Fortunately, confusion rarely results from this imprecision.)

The assignment of values in $GF(q)$ to these cyclotomic cosets is less mysterious than it first appears. First, any sum of complex roots of unity equal to an integer can be carried over to $GF(q)$. This includes

$$\sum_{j=0}^{t-1} \alpha^j = 0,$$

(where $\alpha$ is a primitive $t$th root of unity), for all $t > 1$, and

$$\sum_{\substack{1 \leqq j \leqq t \\ (j,t)=1}} \alpha^j = \mu(t),$$

where $\mu$ is the Möbius function.

**3. Examples.**

1. $q = 2$, $n = 3$, $d = q^n - 1 = 7$. Let $\alpha^7 = 1$.

Then $f(x) = (x-\alpha)(x-\alpha^2)(x-\alpha^4) = x^3 + \eta_1 x^2 + \eta_2 x + 1$, where $\eta_1 = \alpha^1 + \alpha^2 + \alpha^4$ and $\eta_2 = \alpha^3 + \alpha^6 + \alpha^5$. ($\eta_1$ has exponents 001, 010, and 100, and $\eta_2$ has exponents 011, 110, 101.) Since $0 = 1 + \alpha + \alpha^2 + \cdots + \alpha^6 = 1 + \eta_1 + \eta_2$, either $\eta_1 = 1$, $\eta_2 = 0$ or $\eta_2 = 1$, $\eta_1 = 0$. Both assignments are valid, since there are $\phi(7)/3 = 2$ irreducible polynomials of degree 3 over $GF(2)$ for primitive 7th roots of unity. Thus

$$f_1(x) = x^3 + x^2 + 1,$$

$$f_2(x) = x^3 + x + 1.$$

2. $q = 2$, $n = 4$, $d = 2^4 - 1 = 15$. Let $\alpha^{15} = 1$.

Then $f(x) = (x-\alpha)(x-\alpha^2)(x-\alpha^4)(x-\alpha^8) = x^4 + C_1 x^3 + C_2 x^2 + C_3 x + 1$, where

$$C_1 = \alpha^1 + \alpha^2 + \alpha^4 + \alpha^8 = \eta_1,$$

exponents being the cycles of 0001;

$$C_2 = (\alpha^3 + \alpha^6 + \alpha^{12} + \alpha^9) + (\alpha^5 + \alpha^{10}) = \eta_a + \eta_b = 1 + 1 = 0,$$

since $\eta_a$ (exponents are the cycles of 0011) is the sum of the primitive 5th roots of unity, and $\eta_b$ (exponents are the cycles of 0101) is the sum of the primitive cube roots of unity;

$$C_3 = \alpha^7 + \alpha^{14} + \alpha^{13} + \alpha^{11} = \eta_2,$$

exponents being the cycles of 0111. Thus

$$f(x) = x^4 + \eta_1 x^3 + \eta_2 x + 1,$$

where $\eta_1 + \eta_2 = \mu(15) = 1$, so that either $\eta_1 = 1$, $\eta_2 = 0$, or $\eta_1 = 0$, $\eta_2 = 1$. Both assign-

ments are valid and the two polynomials are

$$f_1(x) = x^4 + x^3 + 1,$$
$$f_2(x) = x^4 + x + 1.$$

3. $q = 3$, $n = 2$, $d = 8$. Let $\alpha^8 = 1$. (Then $\alpha^4 = -1$).

Then $f(x) = (x-\alpha)(x-\alpha^3) = x^2 - \eta_1 x - 1$, where $\eta_1 = \alpha + \alpha^3 = -\eta_2 = -(\alpha^5 + \alpha^7)$. Also, $\eta_1 + \eta_2 = \mu(8) = 0$. Since *two* polynomials are sought, we may take either $\eta_1 = 1$, $\eta_2 = -1$, or $\eta_1 = -1$, $\eta_2 = 1$. (The choice $\eta_1 = \eta_2 = 0$ would create an odd number of solutions. Moreover, it is obvious by inspection that $x^2 - 1$ is reducible.) Thus the two solutions are

$$f_1(x) = x^2 - x - 1,$$
$$f_2(x) = x^2 + x - 1.$$

4. $q = 3$, $n = 4$, $d = 20$. Let $\alpha^{20} = 1$, so that $\alpha^{10} = -1$.

Then

$$f(x) = (x-\alpha)(x-\alpha^3)(x-\alpha^9)(x-\alpha^7)$$
$$= x^4 - \eta_1 x^3 + C_2 x^2 - \eta_2 x + 1,$$

where $\eta_1 = \alpha^1 + \alpha^3 + \alpha^7 + \alpha^9$, $\eta_2 = \alpha^{11} + \alpha^{13} + \alpha^{17} + \alpha^{19}$, and $\eta_1 + \eta_2 = \mu(20) = 0$. Also, $C_2 = (\alpha^4 + \alpha^8 + \alpha^{12} + \alpha^{16}) + 2\alpha^{10} = -1 - 2 = 0$, since $\alpha^4 + \alpha^8 + \alpha^{12} + \alpha^{16} = \mu(5) = -1$.

We seek $\phi(20)/4 = 2$ distinct irreducible polynomials, which we obtain from the assignments $\eta_1 = 1$, $\eta_2 = -1$ and $\eta_1 = -1$, $\eta_2 = 1$. The two resulting polynomials are

$$f_1(x) = x^4 - x^3 + x + 1,$$
$$f_2(x) = x^4 + x^3 - x + 1.$$

5. $q = 3$, $n = 4$, $d = 16$. Let $\alpha^{16} = 1$, so that $\alpha^8 = -1$.

Then

$$f(x) = (x-\alpha)(x-\alpha^3)(x-\alpha^9)(x-\alpha^{11})$$
$$= x^4 - \eta_1 x^3 + C_2 x^2 - \eta_2 x - 1,$$

where

$$\eta_1 = \alpha + \alpha^3 + \alpha^9 + \alpha^{11} = \alpha + \alpha^3 - \alpha - \alpha^3 = 0,$$
$$\eta_2 = \alpha^5 + \alpha^7 + \alpha^{13} + \alpha^{15} = \alpha^5 + \alpha^7 - \alpha^5 - \alpha^7 = 0,$$

and

$$C_2 = 2(\alpha^4 + \alpha^{12}) + (\alpha^{10} + \alpha^{14}) = 2(\alpha^4 - \alpha^4) + \alpha^8(\alpha^2 + \alpha^6)$$
$$= -(\alpha^2 + \alpha^6) = -\eta_a.$$

Thus,

$$f(x) = x^4 - \eta_a x^2 - 1,$$

and the two required polynomials result from the assignments $\eta_a = +1$ and $\eta_a = -1$, viz.,

$$f_1(x) = x^4 - x^2 - 1,$$
$$f_2(x) = x^4 + x^2 - 1.$$

Actually, this result can be obtained immediately from the solution to Example 3, since, over $GF(3)$, the polynomials for the primitive 16th roots are obtained from those for the primitive 8th roots by replacing $f(x)$ with $f(x^2)$.

A general rule ([13]) is that if $t$ divides $d$, and the irreducible polynomials $\{f_j(x)\}$ for primitive $d$th roots of unity over a field $F$ have degree $n$ while the irreducible polynomials $\{g_j(x)\}$ for primitive $(td)$th roots of unity over $F$ have degree $tn$, then we obtain the $g_j$'s from the $f_j$'s by the rule $g_j(x) = f_j(x^t)$. (Over the field $Q$ of rationals, $\Phi_{td}(x) = \Phi_d(x^t)$ whenever $t$ divides $d$. However, over $GF(3)$, the irreducible polynomials for primitive 8th roots have the same degree as the irreducible polynomials for primitive 4th roots, so that over $GF(q)$ it is also necessary to check the degrees of the polynomials involved.) The reader may wish to apply this general rule to the tables in [14].

6. $q = 2$, $n = 5$, $d = 31$. Let $\alpha^{31} = 1$.

Then

$$\begin{aligned} f(x) &= (x-\alpha)(x-\alpha^2)(x-\alpha^4)(x-\alpha^8)(x-\alpha^{16}) \\ &= x^5 + C_1x^4 + C_2x^3 + C_3x^2 + C_4x + 1 \\ &= x^5 + \eta_1x^4 + (\eta_2+\eta_3)x^3 + (\eta_5+\eta_6)x^2 + \eta_4x + 1, \end{aligned}$$

where

$$\begin{aligned} \eta_1 &= \alpha^1 + \alpha^2 + \alpha^4 + \alpha^8 + \alpha^{16} && \text{exponents are the cycles of (00001);} \\ \eta_2 &= \alpha^3 + \alpha^6 + \alpha^{12} + \alpha^{24} + \alpha^{17} && \text{exponents are the cycles of (00011);} \\ \eta_3 &= \alpha^9 + \alpha^{18} + \alpha^5 + \alpha^{10} + \alpha^{20} && \text{exponents are the cycles of (00101);} \\ \eta_4 &= \alpha^{27} + \alpha^{23} + \alpha^{15} + \alpha^{30} + \alpha^{29} && \text{exponents are the cycles of (01111);} \\ \eta_5 &= \alpha^{19} + \alpha^7 + \alpha^{14} + \alpha^{28} + \alpha^{25} && \text{exponents are the cycles of (00111);} \\ \eta_6 &= \alpha^{26} + \alpha^{21} + \alpha^{11} + \alpha^{22} + \alpha^{13} && \text{exponents are the cycles of (01011).} \end{aligned}$$

There are $\phi(31)/5 = 6$ irreducible polynomials over $GF(2)$ for the primitive 31st roots of unity. We know $\eta_1+\eta_2+\eta_3+\eta_4+\eta_5+\eta_6 = \mu(31) = -1$, and over $GF(2)$, three of the $\eta_i$'s must be assigned the value 1 while the other three are assigned the value 0. (Each $\eta_i$ is the trace of each of its summands. Since the elements of $GF(2^5)$ are partitioned by trace into two sets of equal size, the elements $\alpha^i$, $1 \leqq i \leqq 30$, must be also). The $\eta_i$'s have been arranged so that $\eta_{i+1}$ contains the cubes of the powers of $\alpha$ contained in $\eta_i$, where the subscripts on $\eta$ are treated modulo 6. A valid assignment of values in $GF(2)$ to the $\eta_i$'s will then yield the other valid assignments by cyclic permutation of the rows, as shown in Table 1.

TABLE 1.

| | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $\eta_5$ | $\eta_6$ | polynomial |
|---|---|---|---|---|---|---|---|
| $f_1$ | 1 | 1 | 0 | 1 | 0 | 0 | $x^5+x^4+x^3+x+1$ |
| $f_2$ | 0 | 1 | 1 | 0 | 1 | 0 | $x^5+x^2+1$ |
| $f_3$ | 0 | 0 | 1 | 1 | 0 | 1 | $x^5+x^3+x^2+x+1$ |
| $f_4$ | 1 | 0 | 0 | 1 | 1 | 0 | $x^5+x^4+x^2+x+1$ |
| $f_5$ | 0 | 1 | 0 | 0 | 1 | 1 | $x^5+x^3+1$ |
| $f_6$ | 1 | 0 | 1 | 0 | 0 | 1 | $x^5+x^4+x^3+x^2+1$ |

The important principle of "superposition of cosets" ([2, p. 56]) states that the shift register sequence corresponding to $f(x)$, which is also the sequence of coefficients in the power series expansion of $1/f(x)$, is obtained as the sequence $\{\mathrm{Tr}(\alpha^i)\}_{i=1}^{p}$. That is, the $i$th term, $a_i$, of the sequence is the value in $GF(q)$ of the coset $\eta_j$ to which $\alpha^i$ belongs.

The assignment of values to the $\eta$'s, as in Table 1, may be obtained from Reuschle's Table ([10]), which in turn is based on algebraic methods of Kummer ([9]). One may also use trial and error, or may work back from shift register sequences by the superposition of cosets principle.

7. $q=2$, $n=8$, $d=17$. Let $\alpha^{17}=1$.

Then

$$f(x)=(x-\alpha)(x-\alpha^2)(x-\alpha^4)(x-\alpha^8)(x-\alpha^{-1})(x-\alpha^{-2})(x-\alpha^{-4})(x-\alpha^{-8})$$
$$=x^8+C_1x^7+C_2x^6+C_3x^5+C_4x^4+C_3x^3+C_2x^2+C_1x+1.$$

(The identity $f(x)=x^nf(1/x)$ corresponds to a left-right symmetry of the coefficients of $f(x)$, and arises from the set of roots of $f(x)$ being closed under the operation $\alpha\to\alpha^{-1}$.)

Let

$$\eta_1=\alpha+\alpha^2+\alpha^4+\alpha^8+\alpha^{-1}+\alpha^{-2}+\alpha^{-4}+\alpha^{-8},$$

and

$$\eta_2=\alpha^3+\alpha^5+\alpha^6+\alpha^7+\alpha^{-3}+\alpha^{-5}+\alpha^{-6}+\alpha^{-7}.$$

Then

$$C_1=\eta_1;\qquad C_2=\eta_2+\eta_2+\eta_1+4=\eta_1;\qquad C_3=4\eta_1+3\eta_2=\eta_2;\quad\text{and}\quad C_4=1,$$

because $f(x)$ has coefficients which are left-right symmetric, and if $C_4=0$, then $f(1)=0$, meaning that $f(x)$ would be divisible by $x+1$, contradicting irreducibility. The computation of $C_3$, for example, comes from taking the cyclic patterns of weight 3, assigning the standard values to these as binary numbers, reducing these modulo 17, and then noting whether $\eta_1$ or $\eta_2$ contains that exponent on $\alpha$. Thus:

$$00000111\to\ 7\to\ 7\to\eta_2;$$
$$00001011\to 11\to -6\to\eta_2;$$
$$00010011\to 19\to\ 2\to\eta_1;$$
$$00100011\to 35\to\ 1\to\eta_1;$$
$$01000011\to 67\to -1\to\eta_1;$$
$$00010101\to 21\to\ 4\to\eta_1;$$
$$00100101\to 37\to\ 3\to\eta_2.$$

As a check, there must be $\binom{8}{3}=56$ powers of $\alpha$ added in $C_3$, and each of the seven $\eta$'s is a sum of eight powers of $\alpha$. We have $\eta_1+\eta_2=\mu(17)=-1$, so that over $GF(2)$ the two valid assignments are $\eta_1=1$, $\eta_2=0$, and $\eta_1=0$, $\eta_2=1$. Hence, from the general

$$f(x)=x^8+\eta_1x^7+\eta_1x^6+\eta_2x^5+x^4+\eta_2x^3+\eta_1x^2+\eta_1x+1,$$

we get

$$f_1(x)=x^8+x^7+x^6+x^4+x^2+x+1=111010111,$$
$$f_2(x)=x^8+x^5+x^4+x^3+1=100111001.$$

As verification, we have:

$$\begin{array}{rr} f_1(x) & 111010111 \\ \times f_2(x) & \times 100111001 \\ \hline f_1(x)f_2(x) = \Phi_{17}(x) = & 111010111 \\ \dfrac{(x^{17}-1)}{(x-1)} = & 111010111 \\ & 111010111 \\ & 111010111 \\ & 111010111 \\ \hline x^{16}+x^{15}+x^{14}+x^{13}+\cdots+x^2+x+1 & 1111111111111111 \end{array}$$

For further information about cyclotomic polynomials, and how they factor over finite fields, see [13].

8. Other cases which the reader may wish to try as exercises include:

$q=2, \quad n=6, \quad d=21.$ (two polynomials)

$q=2, \quad n=11, \quad d=23.$ (two polynomials)

$q=3, \quad n=3, \quad d=13.$ (four polynomials)

$q=3, \quad n=3, \quad d=26.$ (four polynomials, which can be obtained from the previous case by $x \to -x$)

$q=3, \quad n=4, \quad d=40.$ (four polynomials)

$q=2, \quad n=11, \quad d=89.$ (eight polynomials)
Caution: Reuschle's Table contains an error in this case! The correct polynomials are:

$$x^{11}+x^7+x^6+x+1$$
$$x^{11}+x^{10}+x^5+x^4+1$$
$$x^{11}+x^8+x^5+x^4+x^2+x+1$$
$$x^{11}+x^{10}+x^9+x^7+x^6+x^3+1$$
$$x^{11}+x^8+x^7+x^6+x^5+x^3+x^2+x+1$$
$$x^{11}+x^{10}+x^9+x^8+x^6+x^5+x^4+x^3+1$$
$$x^{11}+x^{10}+x^8+x^7+x^6+x^5+x^4+x^3+x^2+x+1$$
$$x^{11}+x^{10}+x^9+x^8+x^7+x^6+x^5+x^4+x^3+x+1$$

## REFERENCES


[1] W. W. PETERSON AND E. WELDON, *Error Correcting Codes*, second edition, M.I.T. Press, Cambridge MA, 1972.

[2] S. W. GOLOMB, *Shift Register Sequences*, Holden-Day, San Francisco, 1967; revised edition, Aegean Park Press, 1980.

[3] I. S. REED AND T. K. TRUONG, *The use of finite fields to compute convolutions*, IEEE Trans. Information Theory, IT-21 (1975), pp. 208–213.

[4] R. W. MARSH, *Table of Irreducible Polynomials over GF(2) through Degree* 19, Office of Technical Services, Commerce Dept., Washington DC, Oct. 24, 1957.

[5] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[6] B. L. VAN DER WAERDEN, *Modern Algebra*, vol. 1, Frederick Ungar, New York, 1949, pp. 160–163.

[7] C. F. GAUSS, *Disquisitiones Arithmeticae*, Leipzig, 1801.
[8] E. E. KUMMER, *Über die Divisoren gewisser Formen der Zahlen, welche aus der Theorie der Kreistheilung entstehen*, J. Reine Angew. Math., 30 (1846), pp. 107–116.
[9] ———, *Über die Zerlegung der aus Wurzeln der Einheit gebildeten complexen Zahlen in ihre Primfactoren*, J. Reine Angew. Math., 35 (1847), pp. 327–367.
[10] K. G. REUSCHLE, *Tafeln Complexer Primzahlen*, Akademie der Wissenschaften, Berlin, 1875.
[11] MARSHALL HALL, JR., *Combinatorial Theory*, Blaisdell, Waltham MA, 1967. (See especially Chapter 11.)
[12] S. W. GOLOMB, *Theory of transformation groups of polynomials over GF(2), with applications to linear shift register sequences*, Information Sciences, 1 (1968) pp. 87–109.
[13] ———, *Cyclotomic polynomials and factorization theorems*, Amer. Math. Monthly, 85 (1978), pp. 734–737.
[14] J. T. B. BEARD, JR. AND K. I. WEST, *Some primitive polynomials of the third kind*, Math. Comp., 28 (1974), pp. 1166–1167 (+microfiche).

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5204/80/0103-0007 $01.00/0

# TOTALLY NONNEGATIVE, *M*-, AND JACOBI MATRICES*

MORDECHAI LEWIN†

**Abstract.** It is shown among other results that a nonsingular *M*-matrix is a Jacobi matrix if and only if its inverse is totally nonnegative and it is a normal Jacobi matrix if and only if its inverse is oscillatory.

This is an extension of a previous result of Markham [Proc. Amer. Math. Soc., 161 (1972), pp. 326–330].

**1. Introduction.** A matrix is *totally nonnegative* (*totally positive*) if all its minors are nonnegative (positive). We shall use the definition of an *M-matrix* as it appears in [5]. The matrix $A$ is an *M-matrix* if it has only nonpositive offdiagonal elements and $A^{-1} \geqq 0$. In recent years the investigation of $M$-matrices has taken a trend towards what is now commonly known as the *inverse M-matrix problem*, the aim of which is to seek characterizing properties of nonsingular nonnegative matrices whose inverses are $M$-matrices. This problem has important applications in operations research as well as in physics. The interested reader may find a number of references in [1, p. 163]. See also [4].

Let $A$ be an $n \times n$ matrix. Let $A_{i,j}$ be the submatrix of $A$ of order $n-1$ obtained by deleting the $i$th row and the $j$th column.

In [5] Markham established the following result.

THEOREM M. *If $F$ is totally nonnegative and nonsingular, then $F^{-1}$ is an* M*-matrix if and only if* $\det(A_{i,j}) = 0$ *for* $i+j \equiv 0 \pmod 2$, $i \neq j$.

In this note we wish to investigate more closely the interrelation between totally nonnegative matrices, $M$-matrices and Jacobi matrices.

A totally nonnegative matrix is *oscillatory*, if some power of it is totally positive. Let $Z^{n,n}$ denote the set of $n \times n$ matrices with nonpositive offdiagonal entries. Let $A = (a_{ij})$. Define $A^* = (a_{ij}^*)$ with $a_{ij}^* = (-1)^{i+j} a_{ij}$. A matrix $A = (a_{ij})$ is a *Jacobi matrix* if $a_{ij} = 0$ for $|i-j| > 1$. A Jacobi matrix of order $n$ is a *normal* Jacobi matrix, if $a_{i,i+1}, a_{i+1,i} < 0$ for $1, 2, \ldots, n-1$. The matrix $A$ is called *sign-regular* if $A^*$ is totally nonnegative. Let $A$ be a square matrix of order $n$. let $\alpha, \beta$ be subsets of $\{1, \ldots, n\}$. Then $A[\alpha|\beta]$ is defined as the submatrix of $A$ obtained from $A$ by deleting the $i$th row whenever $i \notin \alpha$ and the $j$th column whenever $j \notin \beta$. See for example [6].

**2. The main result.** In [3] the following result appears.

THEOREM GK. *Let $A$ be a square matrix. Then $A^{-1}$ is totally nonnegative if and only if $A^*$ is totally nonnegative.*

We wish to establish the following result.

THEOREM 1. *Let $A$ be a matrix. Consider the following three conditions.*

(i) *$A^{-1}$ is totally nonnegative.*

(ii) *$A$ is an M-matrix.*

(iii) *$A$ is a Jacobi matrix.*

*Then any two of the three conditions imply the third.*

*Proof.*

*Case* 1. $A$ satisfies conditions (i) and (ii). By Theorem GK we have $A^*$ totally nonnegative. Consider $a_{ij} \neq 0$ for some $i, j$, $|i-j| > 1$. Then $a_{ij} < 0$. We may assume

*Received by the editors November 30, 1979, and in revised form February 13, 1980.

†Department of Mathematics, The Technion, Israel Institute of Technology, Haifa, Israel.

$i<j$. Then there exists an integer $k$ such that $i<k<j$. If $i$, $j$ have the same parity, then by Theorem M we have $a_{ij}^{*}=0$ and hence $a_{ij}=0$, so we assume $i$ and $j$ to be of different parities. Then $k$ has the parity of either $i$ or $j$. In either case we get $a_{ik}a_{jk}=0$. Then,

$$\det(A^{*}[\{i,k\}|\{k,j\}])=a_{kk}a_{ij}<0, \quad \text{since } a_{kk}>0.$$

This is a contradiction, so that $A$ is a Jacobi matrix.

*Case* 2. $A$ satisfies (i) and (iii). This case is part of [5, Theorem 2.1]. We prove it independently. As in Case 1 we have $A^{*}$ totally nonnegative following from condition (i). Then $A$ is sign-regular (see also [3, $3^{0}$, p. 87]). Since $A$ is a Jacobi matrix, this means that $A\in Z^{n,n}$, and hence $A$ is an $M$-matrix.

*Case* 3. Conditions (ii) and (iii) are satisfied. Then all the leading principal minors of $A$ are positive [2]. Therefore, all the leading principal minors of $A^{*}$ are positive. Applying [3, e, p. 94] it follows that $A^{*}$ is totally nonnegative, so that $A^{-1}$ is totally nonnegative. This proves Theorem 1.

Theorem 2 is analogous to Theorem 1 with *totally nonnegative* replaced by *oscillatory* and *Jacobi matrix* replaced by *normal Jacobi matrix.*

THEOREM 2. *Let $A$ be a matrix. Consider the following three conditions.*

(i) *$A^{-1}$ is oscillatory.*

(ii) *$A$ is an $M$-matrix.*

(iii) *$A$ is a normal Jacobi matrix.*

*Then any two of the three conditions imply the third.*

*Proof.*

*Case* 1. Conditions (i) and (ii) hold. Then, by [3, $8^{0}$, p. 88] we have $A^{*}$ oscillatory. By the characterization of oscillatory matrices in [3, p. 115], $A^{*}$ has a positive superdiagonal and a positive subdiagonal. By Theorem 1, Case 1, matrix $A$ is a Jacobi matrix, so that $A$ is a normal Jacobi matrix.

*Case* 2. $A$ satisfies (i) and (iii). Then by Theorem 1, Case 2, $A$ is an $M$-matrix.

*Case* 3. $A$ satisfies (ii) and (iii). Since $A$ is an $M$-matrix, we have by [2] that all the leading principal minors of $A$ are positive. This is also true for $A^{*}$, since $A$ is a Jacobi matrix. Applying [3, Theorem 11, p. 119] to $A^{*}$ we conclude that $A^{*}$, and hence $A^{-1}$, is oscillatory. This proves Theorem 2.

The following corollary is inherent in Theorems 1 and 2.

COROLLARY 1. *Let $A$ be an $M$-matrix. Then $A^{-1}$ is totally nonnegative if and only if $A$ is a Jacobi matrix, and $A^{-1}$ is oscillatory if and only if $A$ is a normal Jacobi matrix.*

Thus Corollary 1 has the implication that in reality if a totally nonnegative matrix $F$ is an inverse of an $M$-matrix, then not only $\det(F_{i,j})=0$ for all $i+j\equiv 0 \pmod 2$, $i\neq j$, $1\leqq i, j\leqq n$, as indicated by Theorem M, but actually $\det(F_{i,j})=0$ for all $i$ and $j$ for which $|i-j|>1$, $1\leqq i, j\leqq n$, so that Markham's condition in fact "cleans the table" as it were.

**Acknowledgment.** The author would like to thank Prof. Plemmons from the University of Tennessee for his constructive remarks and Dr. Neumann from the University of Nottingham (England), for having asked the right question.

## REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] M. FIEDLER AND V. PTÁK, *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.

[3] F. R. GANTMACHER AND M. G. KREIN, *Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme*, Akademie Verlag, Berlin, 1960.

[4] M. LEWIN AND M. NEUMANN, *The inverse M-matrix problem for* (0, 1)*-matrices*, Linear Algebra and Appl., to appear.

[5] T. L. MARKHAM, *Nonnegative matrices whose inverses are M-matrices*, Proc. Amer. Math. Soc., 16 (1972), pp. 326–330.

[6] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Prindle, Weber and Schmidt, Boston, 1964.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104-0008 $01.00/0

# THE ELIMINATION MATRIX: SOME LEMMAS AND APPLICATIONS*

JAN R. MAGNUS† AND H. NEUDECKER‡

**Abstract.** Two transformation matrices are introduced, $L$ and $D$, which contain zero and unit elements only. If $A$ is an arbitrary $(n, n)$ matrix, $L$ eliminates from vec$A$ the supradiagonal elements of $A$, while $D$ performs the inverse transformation for *symmetric* $A$. Many properties of $L$ and $D$ are derived, in particular in relation to Kronecker products. The usefulness of the two matrices is demonstrated in three areas of mathematical statistics and matrix algebra: maximum likelihood estimation of the multivariate normal distribution, the evaluation of Jacobians of transformations with symmetric or lower triangular matrix arguments, and the solution of matrix equations.

**1. Introduction.** If a matrix $A$ has a known structure (symmetric, skew symmetric, diagonal, triangular), some elements of $A$ are redundant in the sense that they can be deduced from this structure. Thus, if $A$ is a symmetric or lower triangular matrix of order $n$, its $\frac{1}{2}n(n-1)$ supradiagonal elements are redundant. If we eliminate these elements from vec$A$ (the column vector stacking the columns of $A$), this defines a new vector of order $\frac{1}{2}n(n+1)$ which we denote as $v(A)$. The matrix which, for arbitrary $A$, transforms vec$A$ into $v(A)$ is the elimination matrix $L$, first mentioned by Tracy and Singh (1972) and later by Vetter (1975) and Balestra (1976).

Of equal interest is the inverse transformation from $v(A)$ to vec$A$. For lower triangular $A$, we shall see that $L'v(A)=\text{vec}A$. We further introduce the duplication matrix $D$ such that, for symmetric $A$, $Dv(A)=\text{vec}A$. The matrix $D$ (or a matrix comparable to $D$) was previously defined by Tracy and Singh (1972), Browne (1974), Vetter (1975), Balestra (1976), and Nel (1978). $D^+$, the Moore-Penrose inverse of $D$, possesses the property, used by Browne (1974) and Nel (1978), $D^+\text{vec}A=v(A)$, for *symmetric* $A$.

The purpose of this paper is to study the matrices $L$ and $D$. Both matrices consist of zero and unit elements only. §2 gives the necessary definitions and basic tools. The next two sections contain the theoretical heart of the paper and establish a number of results on $L$ and $D$. §§5–7 are devoted to applications: maximum likelihood estimation of the multivariate normal distribution, the evaluation of Jacobians of transformations with symmetric or lower triangular matrix arguments, and, finally, the solution of matrix equations. An appendix presents the proofs of the lemmas in §4.

Not all results are new. Thus, Tracy and Singh (1972) established that $|L(A\otimes A)D|=|A|^{n+1}$. They obtained two other determinants as well (their examples 5.3 and 5.4), but these are both in error. Browne (1974) proved the important fact that $(D'(A\otimes A)D)^{-1}=D^+(A^{-1}\otimes A^{-1})D^{+\prime}$ for nonsingular $A$, while Nel (1978) evaluated the determinant of $D^+(A\otimes B)D$, when $AB=BA$, $A$ and $B$ symmetric. Concurrently with the present paper, Henderson and Searle (1979) wrote an article on the same topic. Inevitably there is some overlap between the two papers.

**2. Notation and preliminary results.** All matrices are real; capital letters represent matrices; lowercase letters denote vectors or scalars. An $(m, n)$ matrix is one having $m$

*Received by the editors September 26, 1978 and in final form February 7, 1980.

†Department of Economics, University of British Columbia, 2075 Wesbrook Mall, Vancouver, Canada V6T 1W5.

‡Institute of Actuarial Science and Econometrics, University of Amsterdam, 23 Jodenbreestraat, 1011 NH Amsterdam, Netherlands.

rows and $n$ columns; $A'$ denotes the transpose of $A$, $\mathrm{tr}A$ its trace, and $|A|$ its determinant. If $A$ is a square matrix, $\bar{A}$ denotes the lower triangular matrix derived from $A$ by setting all supradiagonal elements in $A$ equal to zero; $\mathrm{dg}(A)$ is the diagonal matrix derived from $A$ by setting all supra- *and* infradiagonal elements in $A$ equal to zero. If $x$ is an $n$-vector, and $f(x)=(f_1(x)\cdots f_m(x))'$ a differentiable vector function of $x$, then the matrix $\partial f/\partial x$ has order $(n, m)$ with typical element $(\partial f_j/\partial x_i)$.

The unit vector $e_i$, $i=1,\cdots,n$, is the $i$th column of the identity matrix $I_n$, i.e., it is an $n$-vector with one in its $i$th position and zeroes elsewhere. The $(n, n)$ matrix $E_{ij}$ has one in its $ij$th position and zeroes elsewhere, i.e., $E_{ij}=e_ie_j'$. We partition the identity matrix of order $\frac{1}{2}n(n+1)$ as follows:

$$I_{(1/2)n(n+1)}=(u_{11}u_{21}\cdots u_{n1}u_{22}\cdots u_{n2}u_{33}\cdots u_{nn}).$$

Formally, $u_{ij}$ is a unit vector of order $\frac{1}{2}n(n+1)$ with unity in its $[(j-1)n+i-\frac{1}{2}j(j-1)]$-th position and zeroes elsewhere $(1\leqq j\leqq i\leqq n)$.

If $A$ is an $(m, n)$ matrix and $A_{\cdot j}$ its $j$th column, then $\mathrm{vec}A$ is the $mn$-column vector

$$\mathrm{vec}A=\begin{bmatrix}A_{\cdot 1}\\ \vdots\\ A_{\cdot n}\end{bmatrix}.$$

If $A$ is square of order $n$, $v(A)$ denotes the $\frac{1}{2}n(n+1)$ vector that is obtained from $\mathrm{vec}A$ by eliminating all supradiagonal elements of $A$. For example, if $n=3$,

$$\mathrm{vec}A=(a_{11}a_{21}a_{31}a_{12}a_{22}a_{32}a_{13}a_{23}a_{33})',$$

and

$$v(A)=(a_{11}a_{21}a_{31}a_{22}a_{32}a_{33})'.$$

Finally, the Kronecker product of an $(m, n)$ matrix $A=(a_{ij})$ and an $(s, t)$ matrix $B$ is the $(ms, nt)$ matrix

$$A\otimes B=(a_{ij}B).$$

This settles the notation. Let us now state some preliminary results that will be used throughout. If $A=(a_{ij})$ is an $(n, n)$ matrix, then $A$, $\bar{A}$, and $\mathrm{dg}(A)$ can be expressed as

$$A=\sum_{ij}a_{ij}E_{ij};\qquad \bar{A}=\sum_{i\geqq j}a_{ij}E_{ij};\qquad \mathrm{dg}(A)=\sum_{i=1}^{n}a_{ii}E_{ii}. \tag{2.1}$$

A standard result on vecs is

$$\mathrm{vec}ABC=(C'\otimes A)\mathrm{vec}B, \tag{2.2}$$

if the matrix product $ABC$ exists. For vectors $x$ and $y$ of any order we then have

$$x\otimes y=\mathrm{vec}yx'\qquad\text{and}\qquad x\otimes y'=xy'=y'\otimes x. \tag{2.3}$$

The basic connection between the vec-function and the trace is

$$(\mathrm{vec}A)'\mathrm{vec}B=\mathrm{tr}A'B, \tag{2.4}$$

where $A$ and $B$ are $(m, n)$ matrices. From (2.2) and (2.4) follows

$$(\mathrm{vec}A)'(B\otimes C)\mathrm{vec}D=\mathrm{tr}A'CDB', \tag{2.5}$$

if the expression on the right-hand side exists.

We shall frequently use the commutation matrix $K$ defined implicitly as:

DEFINITION 2.1a (implicit definition of $K$). The $(n^2, n^2)$ *commutation matrix* $K$ performs for every $(n, n)$ matrix $A$ the transformation $K\text{vec}A = \text{vec}A'$.

In fact, $K$ is a special case of the $(mn, mn)$ matrix $K_{mn}$ which maps vec$A$ into vec$A'$ for an arbitrary $(m, n)$ matrix $A$. The matrix $K_{mn}$ was introduced by Tracy and Dwyer (1969). Many of its properties are derived in Magnus and Neudecker (1979), who also established the following explicit expression for $K$.

DEFINITION 2.1b (explicit definition of $K$).

$$K = \sum_{i=1}^{n} \sum_{j=1}^{n} (E_{ij} \otimes E_{ij}').$$

Closely related to the commutation matrix is the matrix $N$.

DEFINITION 2.2a (implicit definition of $N$). The $(n^2, n^2)$ matrix $N$ performs for every $(n, n)$ matrix $A$ the transformation $N\text{vec}A = \text{vec}\frac{1}{2}(A + A')$.

Its explicit expression is immediately derived.

DEFINITION 2.2b (explicit definition of $N$).

$$N = \tfrac{1}{2}(I + K).$$

Note that the implicit definitions of $K$ and $N$ are proper definitions in the sense that they uniquely determine $K$ and $N$. The following lemma gives some properties of $K$ and $N$.

LEMMA 2.1.

(i) $K = K' = K^{-1}$;

(ii) $K(A \otimes B) = (B \otimes A)K$, *for any* $(n, n)$ *matrices* $A$ *and* $B$;

(iii) $N = N' = N^2$;

(iv) $NK = N = KN$.

*For any* $(n, n)$ *matrix* $A$ *we have*

(v) $N(A \otimes A) = (A \otimes A)N = N(A \otimes A)N$;

(vi) $N(I \otimes A + A \otimes I) = (I \otimes A + A \otimes I)N = N(I \otimes A + A \otimes I)N$
$= 2N(I \otimes A)N = 2N(A \otimes I)N$.

*Proof.* The properties of $K$ follow from Magnus and Neudecker (1979). The properties of $N$ follow from those of $K$ since $N = \frac{1}{2}(I + K)$. □

Let us now give four results on the unit vector $u_{ij}$ of order $\frac{1}{2}n(n+1)$ and the $v(\cdot)$ operator.

$$\sum_{i \geqq j} u_{ij} u_{ij}' = I_{(1/2)n(n+1)}. \tag{2.6}$$

If $A$ is an $(n, n)$ matrix, then

$$v(A) = v(\bar{A}) = \sum_{i \geqq j} a_{ij} u_{ij} \quad \text{and} \quad v(\text{dg}(A)) = \sum_{i} a_{ii} u_{ii}; \tag{2.7}$$

$$u_{ij} = v(E_{ij}) \quad \text{and} \quad a_{ij} = u_{ij}' v(A), \qquad i \geqq j; \tag{2.8}$$

$$v(A) = v(\text{dg}(A)), \text{ if } A \text{ is upper triangular.} \tag{2.9}$$

Finally, we make use of the following standard facts in matrix differentiation. For every matrix $X$ and $Y$ of appropriate orders,

$$d(XY) = (dX)Y + X(dY), \tag{2.10}$$

$$d\,\text{tr}XY = \text{tr}(dX)Y + \text{tr}X\,dY. \tag{2.11}$$

For every nonsingular $X$,

$$d\log|X| = \operatorname{tr} X^{-1} dX, \tag{2.12}$$

$$dX^{-1} = -X^{-1}(dX)X^{-1}. \tag{2.13}$$

**3. Basic properties of $L$ and $D$.** Let us now introduce the elimination matrix $L$. As in the previous section, where we defined $K$ and $N$, the elimination matrix will be defined implicitly *and* explicitly.

DEFINITION 3.1a (implicit definition of $L$). The $(\frac{1}{2}n(n+1), n^2)$ *elimination matrix* $L$ performs for every $(n, n)$ matrix $A$ the transformation $L\operatorname{vec} A = v(A)$.

$L$, thus defined, eliminates from $\operatorname{vec} A$ the supradiagonal elements of $A$. We shall show that $L$ is uniquely determined by (3.1a). Let $A$ be an arbitrary $(n, n)$ matrix, and suppose that $\tilde{L}$ and $L$ both transform $\operatorname{vec} A$ into $v(A)$. Then $(\tilde{L} - L)\operatorname{vec} A = 0$ for every $A$. Hence, $\tilde{L} = L$. We can derive an explicit expression for $L$ as follows. Recall that $e_i$, $i = 1 \cdots n$, is the $i$th unit vector of order $n$. Then, using (2.7), (2.4), and (2.3), we find

$$\begin{aligned} v(A) &= \sum_{i \geqq j} a_{ij} u_{ij} = \sum_{i \geqq j} u_{ij}(e_i' A e_j) = \sum_{i \geqq j} u_{ij} \operatorname{tr}(e_j e_i' A) \\ &= \sum_{i \geqq j} u_{ij} \operatorname{tr}(E_{ij}' A) = \sum_{i \geqq j} u_{ij} (\operatorname{vec} E_{ij})' \operatorname{vec} A = \sum_{i \geqq j} (u_{ij} \otimes e_j' \otimes e_i') \operatorname{vec} A. \end{aligned}$$

This leads to the following explicit definition.

DEFINITION 3.1b (explicit definition of $L$).

$$L = \sum_{i \geqq j} u_{ij} (\operatorname{vec} E_{ij})' = \sum_{i \geqq j} (u_{ij} \otimes e_j' \otimes e_i').$$

An example, for $n = 3$, is

$$L = \left[\begin{array}{ccc:ccc:ccc} 1 & 0 & 0 & & & & & & \\ 0 & 1 & 0 & & 0 & & & 0 & \\ 0 & 0 & 1 & & & & & & \\ \hdashline & & & 0 & 1 & 0 & & & \\ & 0 & & 0 & 0 & 1 & & 0 & \\ \hdashline & & & & & & 0 & 0 & 1 \end{array}\right].$$

Most authors on $(0,1)$ matrices are interested only in transformations with *symmetric* matrices, and work with $LN$ rather than $L$. See, e.g., Browne (1974) and Nel (1978). The justification for this lies in the following lemma.

LEMMA 3.1. *For any $(n, n)$ matrix $A$ we have*

$$LN\operatorname{vec} A = \frac{1}{2} v(A + A'). \tag{i}$$

*In particular, when $A$ is symmetric,*

$$LN\operatorname{vec} A = v(A). \tag{ii}$$

*Proof.* Immediate from the implicit definitions of $N$ and $L$. □

Thus, if $A$ is symmetric, $L$ and $LN$ play the same role. In this paper we have chosen a more general approach, based on $L\operatorname{vec} A = v(A)$ for *arbitrary* $A$, largely because this allows us to study transformations with triangular matrices as well. The following lemma characterizes $L$ as a $(0,1)$ matrix with $\frac{1}{2}n(n+1)$ 1's, one in each row and not more than one in each column.

LEMMA 3.2.

(i) *$L$ has full row-rank $\frac{1}{2}n(n+1)$;*

(ii) *$LL' = I_{(1/2)n(n+1)}$;*

(iii) *$L^+ = L'$, where $L^+$ is the Moore-Penrose generalized inverse of $L$.*

*Proof.* We shall show that $LL' = I$. The other two results then follow directly.

$$LL' = \sum_{i \geqq j} (u_{ij} \otimes e_j' \otimes e_i') \sum_{h \geqq k} (u_{hk}' \otimes e_k \otimes e_h) = \sum_{i \geqq j} \sum_{h \geqq k} (u_{ij} u_{hk}' \otimes e_j' e_k \otimes e_i' e_h)$$

$$= \sum_{i \geqq j} u_{ij} u_{ij}' = I_{(1/2)n(n+1)}, \qquad \text{by (2.6).} \quad \square$$

Let us now determine three matrices that are useful for certain linear transformations.

LEMMA 3.3. *The matrices $L'L$, $LKL'$, and $L'LKL'L$ are diagonal and idempotent of rank $\frac{1}{2}n(n+1)$, $n$, and $n$ respectively. Let $A$ be an arbitrary $(n, n)$ matrix. Then,*

(i) $L'L\mathrm{vec}A = \mathrm{vec}\bar{A}$;

(ii) $L'L = \sum_{i \geqq j} (E_{jj} \otimes E_{ii})$;

(iii) $LKL'v(A) = v(\mathrm{dg}(A))$;

(iv) $LKL' = \sum_{i=1}^{n} u_{ii} u_{ii}'$;

(v) $L'LKL'L\mathrm{vec}A = \mathrm{vec}(\mathrm{dg}(A))$;

(vi) $L'LKL'L = \sum_{i=1}^{n} (E_{ii} \otimes E_{ii})$.

*Proof.* By the explicit definition of $L$ we have

$$L'L = \sum_{i \geqq j} (u_{ij}' \otimes e_j \otimes e_i) \sum_{h \geqq k} (u_{hk} \otimes e_k' \otimes e_h') = \sum_{i \geqq j} \sum_{h \geqq k} (u_{ij}' u_{hk} \otimes e_j e_k' \otimes e_i e_h')$$

$$= \sum_{i \geqq j} (e_j e_j' \otimes e_i e_i') = \sum_{i \geqq j} (E_{jj} \otimes E_{ii}),$$

so that, using (2.2) and (2.1),

$$L'L\mathrm{vec}A = \sum_{i \geqq j} (E_{jj} \otimes E_{ii}) \mathrm{vec}A = \sum_{i \geqq j} \mathrm{vec}(E_{ii} A E_{jj})$$

$$= \mathrm{vec} \sum_{i \geqq j} (e_i e_i' A e_j e_j') = \mathrm{vec} \sum_{i \geqq j} a_{ij} E_{ij} = \mathrm{vec}\bar{A}.$$

Further, for arbitrary $v(A)$,

$$LKL'v(A) = LK\mathrm{vec}\bar{A} = L\mathrm{vec}\bar{A}' = v(\bar{A}') = v(\mathrm{dg}(A))$$

$$= \sum_i a_{ii} u_{ii} = \sum_i u_{ii} u_{ii}' v(A),$$

by (i), the implicit definitions of $K$ and $L$, (2.9), (2.7) and (2.8). This proves (iii) and (iv). Similarly, for arbitrary $\mathrm{vec}A$,

$$L'LKL'L\mathrm{vec}A = L'v(\mathrm{dg}(A)) = \mathrm{vec}(\mathrm{dg}(A)) = \mathrm{vec} \sum_i (a_{ii} E_{ii})$$

$$= \mathrm{vec} \sum_i (e_i e_i' A e_i e_i') = \sum_i \mathrm{vec}(E_{ii} A E_{ii}) = \sum_i (E_{ii} \otimes E_{ii}) \mathrm{vec}A,$$

by (iii), (i), (2.1) and (2.2). It is easy to see that the three matrices are diagonal with only zeroes and ones on the diagonal. Hence they are idempotent. The rank of each of

the three matrices equals the number of ones on the diagonal, i.e., $\frac{1}{2}n(n+1)$, $n$, and $n$ respectively. □

Note that Lemma 3.3(i) implies that $L'L\text{vec}A=\text{vec}A$ if and only if $A$ is lower triangular. The matrix $LKL'$, as shown in the previous lemma, is diagonal with $n$ ones and $\frac{1}{2}n(n-1)$ zeroes. Hence, $I+LKL'$ is a nonsingular diagonal matrix with $n$ times 2 and $\frac{1}{2}n(n-1)$ times 1 on the diagonal. Because $LNL'=\frac{1}{2}L(I+K)L'=\frac{1}{2}(LL'+LKL')=\frac{1}{2}(I+LKL')$, it is diagonal too with $n$ times 1 and $\frac{1}{2}n(n-1)$ times $\frac{1}{2}$ on the diagonal. The following properties of $LNL'$ are of interest.

LEMMA 3.4. *The matrix $LNL'$ is diagonal with determinant*

(i) $|LNL'|=2^{-(1/2)n(n-1)}$.

*Its inverse is*

(ii) $(LNL')^{-1}=2I-LKL'$.

*Proof.* Since $LNL'$ is a diagonal matrix, its determinant is the product of its diagonal elements, i.e., $|LNL'|=2^{-(1/2)n(n-1)}$. Property (ii) is easily established using $LNL'=\frac{1}{2}(I+LKL')$ and the idempotency of $LKL'$. □

As we have seen, $L$ uniquely transforms vec$A$ into $v(A)$. The inverse transformation generally does not exist. We can, however, easily transform $v(A)$ into (the vecs of) a lower triangular matrix or a diagonal matrix, since

$$L'v(A)=\text{vec}\bar{A} \qquad \text{(Definition 3.1a and Lemma 3.3 (i))},$$

and

$$L'LKL'v(A)=\text{vec}\,\text{dg}(A) \qquad \text{(Definition 3.1a and Lemma 3.3(v))}.$$

Combining these two transformations one verifies that

$$(L'+KL'-L'LKL')v(A)=\text{vec}\big(\bar{A}+\bar{A}'-\text{dg}(A)\big).$$

We have thus found a matrix which transforms $v(A)$ into (the vec of) a *symmetric* matrix. Let us define this matrix implicitly.

DEFINITION 3.2a (implicit definition of $D$). The $(n^2,\frac{1}{2}n(n+1))$ *duplication matrix* $D$ performs for every $(n,n)$ matrix $A$ the transformation $Dv(A)=\text{vec}(\bar{A}+\bar{A}'-\text{dg}(A))$.

It is easy to see that $D$ is unique. Hence, $D=L'+KL'-L'LKL'=2NL'-L'LKL'$. Note that in particular, if $A$ is symmetric, $DL\text{vec}A=Dv(A)=\text{vec}A$. This is an important property that we will frequently use. The converse is also true; i.e., any $A$ satisfying $DL\text{vec}A=\text{vec}A$ is symmetric.

LEMMA 3.5.

(i) $LD=I_{(1/2)n(n+1)}$;

(ii) $DLN=N$;

(iii) $D=2NL'-L'LKL'=NL'(LNL')^{-1}$.

*Proof.* Let $A=A'$; then $LDv(A)=L\text{vec}A=v(A)$. Hence, $LD=I$, since the symmetry of $A$ does not restrict $v(A)$. Further, for arbitrary $A$,

$$DLN\text{vec}A=DL\text{vec}\tfrac{1}{2}(A+A')=Dv\big(\tfrac{1}{2}(A+A')\big)=\text{vec}\tfrac{1}{2}(A+A')=N\text{vec}A,$$

which proves (ii). It also implies that $DLNL'=NL'$, and because of the nonsingularity of $LNL'$, $D=NL'(LNL')^{-1}$. □

Note that $DLN=N$ is a defining property of $D$. In fact, it is just a reformulation of Definition 3.2a. The matrix $D$ can be explicitly expressed in terms of unit vectors of

order $\frac{1}{2}n(n+1)$ and $n$, i.e., in terms of $u_{ij}$, $e_i$, and $e_j$. From the explicit definition of $K$ and $L$, and the expression for $L'L$ (Lemma 3.3 (ii)), one verifies that

$$LK = \sum_{i \geqq j} u_{ij}(\text{vec}E_{ji})',$$

and

$$LKL'L = \sum_i u_{ii}(\text{vec}E_{ii})',$$

so that

$$\begin{aligned} D' &= L + LK - LKL'L \\ &= \sum_{i \geqq j} u_{ij}(\text{vec}E_{ij})' + \sum_{i \geqq j} u_{ij}(\text{vec}E_{ji})' - \sum_i u_{ii}(\text{vec}E_{ii})'. \end{aligned}$$

Hence, we may define $D$ as follows.

DEFINITION 3.2b (explicit definition of $D$). Let $T_{ij}$ be an $(n, n)$ matrix with 1 in its $ij$th and $ji$th position, and zeroes elsewhere. Then

$$D' = \sum_{i \geqq j} u_{ij}(\text{vec}T_{ij})'.$$

Note that $T_{ij} = E_{ij} + E_{ji}$ for $i \neq j$, and that $T_{ii} = E_{ii}$. An example, for $n = 3$, is

$$D = \left[\begin{array}{ccc:ccc:c} 1 & 0 & 0 & & & \\ 0 & 1 & 0 & & 0 & \\ 0 & 0 & 1 & & & \\ \hdashline 0 & 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \\ \hdashline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Further properties of $D$ are contained in the following two lemmas.

LEMMA 3.6.

(i) *$D$ has full column-rank* $\frac{1}{2}n(n+1)$;

(ii) $KD = D = ND$;

(iii) $D'D = (LNL')^{-1}$;

(iv) $D^+ = LN$.

*Proof.* Straightforward from the expression $D = NL'(LNL')^{-1}$ and the properties $DLN = N = N^2 = KN$, and $LD = I$. □

LEMMA 3.7. *Let $A$ be an arbitrary $(n, n)$ matrix. Then,*

(i) $D'\text{vec}A = v(A + A' - \text{dg}(A))$;

(ii) $DD'\text{vec}A = \text{vec}(A + A' - \text{dg}(A))$;

(iii) $DD' = 2N - L'LKL'L$.

*Proof.* From Lemmas 3.3(iii) and (v), and 3.5(iii), and Definitions 2.1a, 2.2b, 3.1a, and 3.2a, we have

$$D'\text{vec}A = (L + LK - LKL'L)\text{vec}A = v(A) + v(A') - v(\text{dg}(A)),$$

and

$$DD'\text{vec}A = Dv(A+A'-\text{dg}(A)) = \text{vec}(A+A'-\text{dg}(A))$$
$$= (I+K)\text{vec}A - L'LKL'L\text{vec}A = (2N - L'LKL'L)\text{vec}A,$$

for an arbitrary $(n, n)$ matrix $A$. Hence, $DD' = 2N - L'LKL'L$. □

The matrices $L$ and $D$, like the commutation matrix $K$, are useful in matrix differentiation. From Definition 2.1a, it follows that $\partial \text{vec}X/\partial \text{vec}X' = K$ where $X$ is an $(n, n)$ matrix. The corresponding results for $L$ and $D$ are contained in the following lemma.

LEMMA 3.8. *Let $X$ be an $(n, n)$ matrix. Then*

(i) $\partial \text{vec}X/\partial v(X) = \begin{cases} L, & \text{for lower triangular } X; \\ D', & \text{for symmetric } X; \end{cases}$

(ii) $\partial v(X)/\partial \text{vec}X = L'$.

*Proof.* Immediate from the relations $\text{vec}X = L'v(X)$ (lower triangular $X$), $\text{vec}X = Dv(X)$ (symmetric $X$), and $v(X) = L\text{vec}X$. □

*Comment.* More general results are easily obtained from Lemma 3.8, using the chain rule. In particular, let $Y = F(x)$ be an $(n, n)$ matrix whose elements are differentiable functions of a vector $x$. Then

(i) $\partial \text{vec}Y/\partial x = \begin{cases} (\partial v(Y)/\partial x)L & \text{if } Y \text{ is lower triangular for all } x; \\ (\partial v(Y)/\partial x)D' & \text{if } Y \text{ is symmetric for all } x; \end{cases}$

(ii) $\partial v(Y)/\partial x = (\partial \text{vec}Y/\partial \text{x})L'$ for all $Y$.

**4. Applications to Kronecker products.** From §2 we know that the commutation matrix $K$ possesses two major properties: a *transformation* property, $K\text{vec}A = \text{vec}A'$ (its definition), and a *Kronecker* property, $K(A \otimes B)K = B \otimes A$. The elimination matrix $L$ and the duplication matrix $D$ have likewise been defined by their transformation properties, viz. $L\text{vec}A = v(A)$ and, for symmetric $A$, $Dv(A) = \text{vec}A$. Let us now investigate their Kronecker properties. The applications in §§5–7 are based almost entirely on the lemmas in the present section. Proofs are postponed to the Appendix.

We shall first show that, if $A$ and $B$ have a certain structure (diagonal, triangular), Kronecker forms of the type $L(A \otimes B)L'$ and $L(A \otimes B)D$ often possess the same structure.

LEMMA 4.1. *Let $\Lambda$ and $M$ be diagonal $n$-matrices with diagonal elements $\lambda_i$ and $\mu_i$ $(i = 1 \cdots n)$. Let further $P$ and $Q$ be lower triangular $n$-matrices with diagonal elements $p_{ii}$ and $q_{ii}$, $i = 1 \cdots n$. Then,*

(i) *$L(\Lambda \otimes M)L' = L(\Lambda \otimes M)D$ is diagonal with elements $\mu_i \lambda_j (i \geqq j)$ and determinant $\prod_i \mu_i^i \lambda_i^{n-i+1}$;*

(ii) *$L(P \otimes Q)L'$ is lower triangular with diagonal elements $q_{ii}p_{jj} (i \geqq j)$ and determinant $\prod_i q_{ii}^i p_{ii}^{n-i+1}$;*

(iii) *$L(P \otimes Q)D$ is lower triangular and $L(P' \otimes Q')D$ is upper triangular. Both matrices have diagonal elements $q_{ii}p_{jj} (i \geqq j)$ and determinant $\prod_i q_{ii}^i p_{ii}^{n-i+1}$.*

Next we establish some properties of $L(P' \otimes Q)L'$, with lower triangular $P$ and $Q$. Notice that (i) is the Kronecker counterpart of the property $L'L\text{vec}P = \text{vec}P$ for lower triangular $P$.

LEMMA 4.2. *For lower triangular $n$-matrices $P = (p_{ij})$ and $Q = (q_{ij})$,*

(i) $L'L(P' \otimes Q)L' = (P' \otimes Q)L'$;

(ii)

$$[L(P'\otimes Q)L']^s = L[(P')^s\otimes Q^s]L', \quad \begin{cases} s=0,1,2,\cdots, \\ s=\cdots,-2,-1, & \text{if } P^{-1} \text{ and } Q^{-1} \text{ exist}, \\ s=\frac{1}{2}, & \text{if lower triangular } P^{1/2} \text{ and } Q^{1/2} \text{ exist}; \end{cases}$$

(iii) $L(P'\otimes Q)L' = D'(P'\otimes Q)L'$ *has eigenvalues* $q_{ii}p_{jj}$ $(i\geqq j)$ *and determinant* $\Pi_i q_{ii}^i p_{ii}^{n-i+1}$.

In Lemma 4.1 we have proved that, for lower triangular $P$ and $Q$, the matrices $L(P\otimes Q)L'$, $L(P\otimes Q)D$, $L(P'\otimes Q')L'$, and $L(P'\otimes Q')D$ are triangular as well, with diagonal elements $q_{ii}p_{jj}$, $i\geqq j$. Although the matrices $L(P'\otimes Q)L'$, $L(P\otimes Q')L'$, and $L(P\otimes Q')D$ are not triangular, they also possess eigenvalues $q_{ii}p_{jj}$, $i\geqq j$; see Lemma 4.2. The matrix $L(P'\otimes Q)D$ is more complicated and seems not to have such nice properties. In particular, its eigenvalues are in general different from $q_{ii}p_{jj}$, $i\geqq j$.

The results of Lemmas 4.1 and 4.2 enable us to find the following determinants which are of importance in the evaluation of Jacobians of transformations with lower triangular matrix arguments (see §6).

LEMMA 4.3. *For lower triangular n-matrices* $P$, $Q$, $R$, *and* $S$ *with diagonal elements* $p_{ii}$, $q_{ii}$, $r_{ii}$, *and* $s_{ii}$, $i=1\cdots n$, *we have*

(i) $|L(PQ'\otimes R'S)L'| = \Pi_i (r_{ii}s_{ii})^i (p_{ii}q_{ii})^{n-i+1}$;

(ii) $|L(P'\otimes Q + R'\otimes S)L'| = \Pi_{i\geqq j}(q_{ii}p_{jj} + s_{ii}r_{jj})$;

(iii) $|L(P\otimes Q'R)D| = \Pi_i (q_{ii}r_{ii})^i p_{ii}^{n-i+1}$;

(iv) $|L(PQ'\otimes R')D| = \Pi_i r_{ii}^i (p_{ii}q_{ii})^{n-i+1}$.

*If* $P, Q, R, S$ *are nonsingular,*

(v) $[L(PQ'\otimes R'S)L']^{-1} = L(Q'^{-1}\otimes S^{-1})L'L(P^{-1}\otimes R'^{-1})L'$.

*Finally,*

(vi) $$\left|L\sum_{h=1}^{H}[(P')^{H-h}\otimes P^{h-1}]L'\right| = H^n|P|^{H-1}\Pi_{i>j}\mu_{ij}, \qquad H=2,3,\cdots,$$

*where*

$$\mu_{ij} = \begin{cases} \dfrac{(p_{ii}^H - p_{jj}^H)}{(p_{ii}-p_{jj})}, & \text{if } p_{ii}\neq p_{jj}, \\ Hp_{ii}^{H-1}, & \text{if } p_{ii}=p_{jj}. \end{cases}$$

A variety of corollaries flow from Lemma 4.3 by putting one or more of the matrices $P$, $Q$, $R$, and $S$ equal to $I$. Also, the four matrices $L(PQ'\otimes Q'P)L'$, $L(PQ'\otimes P'Q)L'$, $L(PQ\otimes Q'P)D$, and $L(PQ'\otimes P'Q')D$ have the same determinant, namely $|P|^{n+1}|Q|^{n+1}$.

In Lemmas 4.1–4.3 we have studied triangular matrices only. The crucial properties for lower triangular matrices are $L'L\text{vec}P = \text{vec}P$ and its Kronecker counterpart $L'L(P'\otimes Q)L' = (P'\otimes Q)L'$, which enable us to discover further properties of the important matrix $L(P'\otimes Q)L'$. Let us now turn away from triangular matrices. An equally important property is $DL\text{vec}A = \text{vec}A$ for symmetric $A$. Its Kronecker counterpart is $DL(A\otimes A)D = (A\otimes A)D$ for *arbitrary* $A$, as we shall see shortly, and it enables us to study the matrix $L(A\otimes A)D$.

LEMMA 4.4. *For any* $(n,n)$ *matrix* $A$,

(i) $DL(A\otimes A)D = (A\otimes A)D$;

(ii) $$[L(A\otimes A)D]^s = L(A^s\otimes A^s)D, \quad \begin{cases} s=0,1,2,\cdots, \\ s=\cdots,-2,-1 & \text{if } A^{-1} \text{ exists}, \\ s=\frac{1}{2} & \text{if } A^{1/2} \text{ exists}; \end{cases}$$

(iii) *The eigenvalues of* $L(A\otimes A)D$ *are* $\lambda_i\lambda_j$, $i\geqq j$, *when* $A$ *has eigenvalues* $\lambda_i$ $(i=1\cdots n)$;

(iv) $|L(A\otimes A)D|=|A|^{n+1}$;

(v) $|D'(A\otimes A)D|=2^{(1/2)n(n-1)}|A|^{n+1}$;

*If* $A$ *is nonsingular,*

(vi) $[D'(A\otimes A)D]^{-1}=LN(A^{-1}\otimes A^{-1})NL'$.

*If* $AB=BA$, *and* $A$ *and* $B$ *have eigenvalues* $\lambda_i$ *and* $\mu_i$, $i=1\cdots n$,

(vii) $|D'(A\otimes B)D|=|A||B|\Pi_{i>j}(\lambda_i\mu_j+\lambda_j\mu_i)$.

Note that in (vii) we do not require $A$ and $B$ to be symmetric, in contrast to Nel (1978). In applying (vii) one must be careful to note that knowledge of $\lambda_i$ and $\mu_j$ is *not* sufficient in order to compute $\Pi_{i>j}(\lambda_i\mu_j+\lambda_j\mu_i)$. In general, it is necessary to carry out the simultaneous reduction of $A$ and $B$ to diagonal form, because the ordering of the eigenvalues is important. A case that can be solved without this reduction is

$$|D'(A^p\otimes A^q)D|=|A|^{p+nq}\prod_{i>j}(\lambda_i^{p-q}+\lambda_j^{p-q}),$$

where $p$ and $q$ are integers (positive, negative or zero).

Similar results hold for the Kronecker sum $I\otimes A+A\otimes I$:

LEMMA 4.5. *For any* $(n,n)$ *matrix* $A$ *with eigenvalues* $\lambda_i$ $(i=1\cdots n)$,

(i) $DL(I\otimes A+A\otimes I)D=(I\otimes A+A\otimes I)D=2N(I\otimes A)D=2N(A\otimes I)D$;

(ii) *the eigenvalues of* $L(I\otimes A+A\otimes I)D$ *are* $\lambda_i+\lambda_j$, $i\geqq j$;

(iii) $|L(I\otimes A+A\otimes I)D|=2^n|A|\Pi_{i>j}(\lambda_i+\lambda_j)$;

(iv) $[L(I\otimes A+A\otimes I)D]^{-1}=L(I\otimes A+A\otimes I)^{-1}D$;

*for nonsingular* $I\otimes A+A\otimes I$. *The results* (i) *and* (iii) *can be generalized to*

(v) $$DL\sum_{h=1}^{H}(A^{H-h}\otimes A^{h-1})D=\sum_{h=1}^{H}(A^{H-h}\otimes A^{h-1})D, \qquad H=2,3,\cdots,$$

*and*

(vi) $$\left|L\sum_{h=1}^{H}(A^{H-h}\otimes A^{h-1})D\right|=H^n|A|^{H-1}\Pi_{i>j}\mu_{ij}, \qquad H=2,3,\cdots,$$

*where*

$$\mu_{ij}=\begin{cases}\dfrac{(\lambda_i^H-\lambda_j^H)}{(\lambda_i-\lambda_j)}, & \text{if } \lambda_i\neq\lambda_j\\ H\lambda_i^{H-1}, & \text{if } \lambda_i=\lambda_j.\end{cases}$$

The next lemma concerns the determinant of the sum or the difference of the matrices $L(A\otimes A)D$ and $L(B\otimes B)D$.

LEMMA 4.6. *Let* $A$ *and* $B$ *be* $(n,n)$ *matrices. Then the determinant*

$$|L(A\otimes A\pm B\otimes B)D|$$

*equals*

(i) $|A|^{n+1}\Pi_{i\geqq j}(1\pm\lambda_i\lambda_j)$, *if* $A$ *is nonsingular and* $\lambda_i$, $i=1\cdots n$, *are the eigenvalues of* $BA^{-1}$;

(ii) $\Pi_{i\geqq j}(a_{ii}a_{jj}\pm b_{ii}b_{jj})$, *if* $A=(a_{ij})$ *and* $B=(b_{ij})$ *are lower triangular*;

(iii) $\Pi_{i\geqq j}(\mu_i\mu_j\pm\theta_i\theta_j)$, *if* $AB=BA$, *where* $\mu_i$ *and* $\theta_i$ $(i=1\cdots n)$ *denote the eigenvalues of* $A$ *and* $B$.

Again, knowledge of $\mu_i$ and $\theta_j$ is, in general, not sufficient to compute (iii). See the remarks under Lemma 4.4(vi).

A final lemma will prove useful in §§5 and 6.

LEMMA 4.7. *Let $P$ be a lower triangular and nonsingular $(n, n)$ matrix, and $\alpha$ a scalar. Let*

(i) $|L[P'\otimes P+\alpha\text{vec}P(\text{vec}P')']L'|=(1+\alpha n)|P|^{n+1}$;

(ii) $[L[P'\otimes P+\alpha\text{vec}P(\text{vec}P')']L')^{-1}=L[P'^{-1}\otimes P^{-1}-\beta\text{vec}P^{-1}(\text{vec}P'^{-1})']L'$,

*where $\beta=\alpha/(1+\alpha n)$. Let further $A$ be a symmetric and nonsingular $(n, n)$ matrix. Then,*

(iii) $|L[A\otimes A+\alpha\text{vec}A(\text{vec}A)']D|=(1+\alpha n)|A|^{n+1}$;

(iv) $[L[A\otimes A+\alpha\text{vec}A(\text{vec}A)']D]^{-1}=L[A^{-1}\otimes A^{-1}-\beta\text{vec}A^{-1}(\text{vec}A^{-1})']D$.

This ends the theoretical part of this paper.

**5. Maximum likelihood estimation of the multivariate normal distribution.** We shall now show the usefulness of $L$ and $D$ in a number of applications. Consider a sample of size $m$ from the $n$-dimensional normal distribution of $y$ with mean $\mu$ and positive definite covariance matrix $\Phi$. The maximum likelihood (ML) estimators of $\mu$ and $\Phi$ are well known, but the derivation of these estimators is often incorrect. The problem is to take properly into account the symmetry conditions on $\Phi$, as has recently been stressed by Richard (1975) and Balestra (1976). More precisely, we should not differentiate the likelihood function with respect to $\text{vec}\Phi$, but with respect to $v(\Phi)$. First we derive the ML estimators of $\mu$ and $\Phi$ (Lemma 5.1), then the information matrix and asymptotic covariance matrix (Lemma 5.2), and finally we investigate properties of the random vector $v(F)$, an unbiased estimator of $v(\Phi)$.

LEMMA 5.1. *Consider a sample of size $m$ from the $n$-dimensional normal distribution of $y$ with mean $\mu$ and positive definite covariance matrix $\Phi$. The maximum likelihood estimators of $\mu$ and $\Phi$ are*

$$\hat{\mu}=\left(\frac{1}{m}\right)\sum_i y_i\equiv\bar{y};$$

$$\hat{\Phi}=\left(\frac{1}{m}\right)\sum_i (y_i-\bar{y})(y_i-\bar{y})'.$$

*Proof.* The loglikelihood function for the sample is

$$\Lambda_m(y;\mu,v(\Phi))=-\frac{1}{2}nm\log 2\pi-\frac{1}{2}m\log|\Phi|-\frac{1}{2}\text{tr}\Phi^{-1}Z,$$

where

$$Z=\sum_{i=1}^{m}(y_i-\mu)(y_i-\mu)'.$$

Using well-known properties of matrix differentials (see (2.10)–(2.13)) and traces (2.4)–(2.5), the first differential of $\Lambda$ can be written as

$$\begin{aligned}
d\Lambda&=-\frac{1}{2}md\log|\Phi|-\frac{1}{2}\text{tr}(d\Phi^{-1})Z-\frac{1}{2}\text{tr}\Phi^{-1}dZ\\
&=-\frac{1}{2}m\text{tr}\Phi^{-1}d\Phi+\frac{1}{2}\text{tr}\Phi^{-1}(d\Phi)\Phi^{-1}Z\\
&\quad+\frac{1}{2}\text{tr}\Phi^{-1}\left[\sum_i(y_i-\mu)(d\mu)'+(d\mu)\sum_i(y_i-\mu)'\right]\\
&=\frac{1}{2}\text{tr}(d\Phi)\Phi^{-1}(Z-m\Phi)\Phi^{-1}+(d\mu)'\Phi^{-1}\sum_i(y_i-\mu)\\
&=\frac{1}{2}(\text{vec}d\Phi)'(\Phi^{-1}\otimes\Phi^{-1})\text{vec}(Z-m\Phi)+(d\mu)'\Phi^{-1}\sum_i(y_i-\mu)\\
&=\frac{1}{2}(dv(\Phi))'D'(\Phi^{-1}\otimes\Phi^{-1})\text{vec}(Z-m\Phi)+(d\mu)'\Phi^{-1}\sum_i(y_i-\mu).
\end{aligned}$$

Necessary for a maximum is that $d\Lambda=0$ for all $d\mu\neq 0$ and $dv(\Phi)\neq 0$. This gives

$$\Phi^{-1}\sum_i (y_i-\mu)=0,$$

and

$$D'(\Phi^{-1}\otimes\Phi^{-1})\mathrm{vec}(Z-m\Phi)=0.$$

The first condition implies $\hat{\mu}=(1/m)\Sigma y_i\equiv\bar{y}$. The second can be written as

$$D'(\Phi^{-1}\otimes\Phi^{-1})Dv(Z-m\Phi)=0,$$

that is

$$v(Z-m\Phi)=0,$$

since $D'(\Phi^{-1}\otimes\Phi^{-1})D$ is nonsingular. Thus,

$$\hat{\Phi}=(1/m)\hat{Z}=(1/m)\sum_i (y_i-\bar{y})(y_i-\bar{y})'. \quad \square$$

The precision and efficiency of an estimator is usually stated in terms of the information matrix defined by

$$\Psi_m\equiv -E\frac{\partial^2\Lambda_m}{\partial\theta\partial\theta'}, \qquad \theta'=(\mu',(v(\Phi))').$$

Its inverse is a lower bound for the covariance matrix of any unbiased estimator of $\mu$ and $v(\Phi)$. This is the Cramér-Rao inequality (see, e.g., Rao (1973)). The asymptotic information matrix is defined as

$$\Psi=\lim_{m\to\infty}\frac{1}{m}\Psi_m,$$

and its inverse is the asymptotic covariance matrix of the ML estimator.[1]

LEMMA 5.2. *The information matrix for $\mu$ and $v(\Phi)$ is the $(\frac{1}{2}n(n+3),\frac{1}{2}n(n+3))$ matrix*

$$\Psi_m=m\begin{pmatrix}\Phi^{-1} & 0\\ 0 & \frac{1}{2}D'(\Phi^{-1}\otimes\Phi^{-1})D\end{pmatrix};$$

*the asymptotic covariance matrix of the* ML *estimators $\hat{\mu}$ and $v(\hat{\Phi})$ is*

$$\Psi^{-1}=\begin{pmatrix}\Phi & 0\\ 0 & 2LN(\Phi\otimes\Phi)NL'\end{pmatrix},$$

*and the generalized asymptotic variance of $v(\hat{\Phi})$ is*

$$|2LN(\Phi\otimes\Phi)NL'|=2^n|\Phi|^{n+1}.$$

*Proof.* Recall that the first differential of $\Lambda$ is

$$d\Lambda=(d\mu)'\Phi^{-1}\sum_i (y_i-\mu)+\frac{1}{2}(dv(\Phi))'D'(\Phi^{-1}\otimes\Phi^{-1})\mathrm{vec}(Z-m\Phi).$$

[1]Some authors refer to $\Psi_m^{-1}$ (rather than $\Psi^{-1}$) as the asymptotic covariance matrix of $\hat{\theta}$.

The second differential is therefore

$$d^2\Lambda=(d\mu)'(d\Phi^{-1})\sum_i(y_i-\mu)-m(d\mu)'\Phi^{-1}(d\mu)$$
$$+\frac{1}{2}(dv(\Phi))'D'\big(d(\Phi^{-1}\otimes\Phi^{-1})\big)\mathrm{vec}(Z-m\Phi)$$
$$+\frac{1}{2}(dv(\Phi))'D'(\Phi^{-1}\otimes\Phi^{-1})\mathrm{vec}(dZ-md\Phi).$$

Taking expectations, and observing that $Ey_i=\mu$, $EZ=m\Phi$, and $EdZ=0$, we find

$$-Ed^2\Lambda=m(d\mu)'\Phi^{-1}(d\mu)+\left(\frac{m}{2}\right)(dv(\Phi))'D'(\Phi^{-1}\otimes\Phi^{-1})\mathrm{vec}\,d\Phi$$
$$=m(d\mu)'\Phi^{-1}(d\mu)+\left(\frac{m}{2}\right)(dv(\Phi))'D'(\Phi^{-1}\otimes\Phi^{-1})Ddv(\Phi).$$

The information matrix then follows. From Lemma 4.4 we know that

$$\big[D'(\Phi^{-1}\otimes\Phi^{-1})D\big]^{-1}=LN(\Phi\otimes\Phi)NL',$$

and

$$|D'(\Phi^{-1}\otimes\Phi^{-1})D|=2^{(1/2)n(n-1)}|\Phi|^{-(n+1)}.$$

Hence,

$$\Psi^{-1}=\left(\frac{1}{m}\Psi_m\right)^{-1}=\begin{pmatrix}\Phi & 0\\ 0 & 2LN(\Phi\otimes\Phi)NL'\end{pmatrix},$$

and

$$|2LN(\Phi\otimes\Phi)NL'|=2^{(1/2)n(n+1)}|D'(\Phi^{-1}\otimes\Phi^{-1})D|^{-1}$$
$$=2^{(1/2)n(n+1)}2^{-(1/2)n(n-1)}|\Phi|^{n+1}=2^n|\Phi|^{n+1}.\quad\square$$

The ML estimator $v(\hat{\Phi})$ is *not* an unbiased estimator of $v(\Phi)$. Let us therefore define

$$F\equiv\frac{1}{m-1}\sum_i(y_i-\bar{y})(y_i-\bar{y})'=\frac{m}{m-1}\hat{\Phi}.$$

The following properties of $v(F)$ can then be established.

LEMMA 5.3. *The random vector $v(F)$ is an unbiased estimator of $v(\Phi)$,*

(i) $Ev(F)=v(\Phi)$.

*Its covariance matrix is*

(ii) $\mathrm{cov}(v(F))=2(LN(\Phi\otimes\Phi)NL')/(m-1)$,

*and $v(F)$ is therefore a consistent estimator of $v(\Phi)$. In particular,*

(iii) $\mathrm{var}(f_{ij})=(\phi_{ij}^2+\phi_{ii}\phi_{jj})/(m-1)$, $\quad i\geqq j=1\cdots n$.

*Finally, the efficiency of $v(F)$ is*

(iv) $\mathrm{eff}(v(F))=[(m-1)/m]^{(1/2)n(n+1)}$.

*Proof.* We know that

$$m\hat{\Phi}=\sum_i(y_i-\bar{y})(y_i-\bar{y})'=\sum_i y_iy_i'-m\bar{y}\bar{y}'$$

is centrally Wishart distributed $W_n(m-1,\Phi)$, see Rao (1973, p. 537). Therefore, as derived in Magnus and Neudecker (1979, Corollary 4.2),

$$mE\hat{\Phi}=(m-1)\Phi,$$

and

$$\operatorname{cov}(m\operatorname{vec}\hat{\Phi}) = (m-1)(I+K)(\Phi\otimes\Phi).$$

Thus, $v(F)=(m/(m-1))v(\hat{\Phi})$ is an unbiased estimator of $v(\Phi)$ and its covariance matrix is

$$\begin{aligned}\operatorname{cov}(v(F)) &= \frac{1}{(m-1)^2}\operatorname{cov}(mv(\hat{\Phi})) = \frac{1}{(m-1)^2}\operatorname{cov}(L\operatorname{vec} m\hat{\Phi}) \\ &= \frac{1}{(m-1)^2}\cdot(m-1)L(I+K)(\Phi\otimes\Phi)L' \\ &= \frac{2}{(m-1)}\cdot LN(\Phi\otimes\Phi)L' = \frac{2}{(m-1)}\cdot LN(\Phi\otimes\Phi)NL'.\end{aligned}$$

We see that $\operatorname{cov}(v(F))\to 0$ as $m\to\infty$. This shows that $v(F)$ is a consistent estimator of $v(\Phi)$, given (i). The diagonal elements of $LN(\Phi\otimes\Phi)NL'$ can be derived as follows. Let $i\geqq j$; then, by (2.8), Lemma 3.3 (i), and (2.5),

$$\begin{aligned}u_{ij}'LN(\Phi\otimes\Phi)NL'u_{ij} &= \big(v(E_{ij})\big)'LN(\Phi\otimes\Phi)NL'v(E_{ij}) \\ &= \tfrac{1}{4}\big(\operatorname{vec}(E_{ij}+E_{ji})\big)'(\Phi\otimes\Phi)\operatorname{vec}(E_{ij}+E_{ji}) \\ &= \tfrac{1}{2}\big(\operatorname{tr}E_{ij}\Phi E_{ij}\Phi+\operatorname{tr}E_{ij}\Phi E_{ji}\Phi\big) = \tfrac{1}{2}\big(\phi_{ij}^2+\phi_{ii}\phi_{jj}\big).\end{aligned}$$

Thus,

$$\operatorname{var}(f_{ij}) = \left(\frac{1}{(m-1)}\right)\big(\phi_{ij}^2+\phi_{ii}\phi_{jj}\big).$$

Finally, the efficiency of $v(F)$ is [see Anderson (1958, p. 57)]

$$\begin{aligned}\operatorname{eff}(v(F)) &= \frac{\left|\dfrac{-E\partial^2\Lambda}{\partial v(\Phi)\partial v(\Phi)'}\right|^{-1}}{|\operatorname{cov}(v(F))|} \\ &= \left|\frac{m}{2}D'(\Phi^{-1}\otimes\Phi^{-1})D\right|^{-1}\left|\frac{2}{(m-1)}LN(\Phi\otimes\Phi)NL'\right|^{-1} \\ &= \left[\frac{(m-1)}{m}\right]^{(1/2)n(n+1)}. \quad \square\end{aligned}$$

Lemmas 5.1 and 5.2 can be straightforwardly generalized by allowing the $y_i$ to have different expectations $\mu_i$. Clearly, it is not possible to estimate all $\mu_i$ $(i=1\cdots m)$ and $v(\Phi)$, i.e., $nm+\frac{1}{2}n(n+1)$ parameters, from $nm$ observations. If, however, we assume that the $\mu_i$ depend upon a fixed number of parameters $(\theta_1\cdots\theta_K)\equiv\theta'$, and $\hat{\theta}$ is the ML estimator of $\theta$, then the ML estimator of $\Phi$ is

$$\hat{\Phi} = \left(\frac{1}{m}\right)\sum_i\big(y_i-\mu_i(\hat{\theta})\big)\big(y_i-\mu_i(\hat{\theta})\big)',$$

and the asymptotic covariance matrix of $v(\hat{\Phi})$ is again

$$\text{as.cov}\big(v(\hat{\Phi})\big) = 2LN(\Phi\otimes\Phi)NL'.$$

**6. Jacobians.** Let the matrix $Y$ be a one-to-one function of a matrix $X$. The matrix $J=J(Y, X)=(\partial \operatorname{vec} Y/\partial \operatorname{vec} X)'$ is called the Jacobian matrix and its determinant the Jacobian of the transformation of $X$ to $Y$.

Because the ordering of the variables is arbitrary, the value of the Jacobian can vary in sign, but since only the absolute value matters, this should not worry us. Note that our definition of a Jacobian differs from some textbooks', where $J(Y, X)$ is defined as $|\partial \operatorname{vec} Y/\partial \operatorname{vec} X|^{-1}$.

Consider for example the linear transformation

$$Y=AX,$$

where $X$ and $Y$ are $(m, n)$ matrices, and $A$ is a nonsingular $(m, m)$ matrix. Taking differentials and vecs we have

$$dY=A\,dX,$$

and

$$d\operatorname{vec} Y=(I\otimes A)d\operatorname{vec} X,$$

so that

$$|J(Y, X)|=\left|\frac{\partial \operatorname{vec} Y}{\partial \operatorname{vec} X}\right|=|I\otimes A|=|A|^n.$$

The evaluation of Jacobians of transformations involving symmetric or lower triangular matrix arguments is not straightforward, since in this case $X$ contains only $\frac{1}{2}n(n+1)$ "essential" variables. To account for this, a variety of methods have been used, notably differential techniques (Deemer and Olkin (1951) and Olkin (1953)), induction (Jack (1966)), and functional equations induced on the relevant spaces (Olkin and Sampson (1972)). Our approach finds its root in Tracy and Singh (1972) who used modified matrix differentiation results to obtain Jacobians in a simple fashion.

Consider the relation between the $\frac{1}{2}n(n+1)$ variables $y_{ij}$ and the $\frac{1}{2}n(n+1)$ variables $x_{ij}$ given by

$$Y=AXA',$$

where $X$ (and hence $Y$) is symmetric. Taking differentials and vecs, we have

$$d\operatorname{vec} Y=(A\otimes A)d\operatorname{vec} X,$$

and, using the definitions of $L$ and $D$,

$$dv(Y)=L(A\otimes A)D\,dv(X),$$

so that by Lemma 4.4 (iv)

$$|J(Y, X)|=\left|\frac{\partial v(Y)}{\partial v(X)}\right|=|L(A\otimes A)D|=|A|^{n+1}.$$

See also Deemer and Olkin (1951), Anderson (1958, pp. 156 and 162), Jack (1968), Tracy and Singh (1972), and Olkin and Sampson (1972). Anderson unnecessarily assumes that $X$ has a Wishart distribution or that $A$ is triangular. A more general transformation is

$$Y=AXA'\pm BXB',$$

where $X$ again is symmetric. This yields

$$dv(Y)=L(A\otimes A\pm B\otimes B)D\,dv(X),$$

and the Jacobian matrix is

$$J(Y, X)=L(A\otimes A\pm B\otimes B)D,$$

of which we know the determinant from Lemma 4.6. See Tracy and Singh (1972) for an earlier (wrong) solution in the case $AB=BA$.

We now turn to *nonlinear* transformations involving symmetric matrix arguments. Consider

$$Y=XAX,$$

where $A$ and $X$ are symmetric. Differentiating,

$$dY=(dX)AX+XA(dX),$$

so that

$$d\text{vec}Y=(XA\otimes I+I\otimes XA)d\text{vec}X,$$

and

$$dv(Y)=L(XA\otimes I+I\otimes XA)Ddv(X).$$

Thus, from Lemma 4.5 (iii), the Jacobian is

$$|J(Y,X)|=|L(XA\otimes I+I\otimes XA)D|=2^n|A||X|\prod_{i>j}(\lambda_i+\lambda_j),$$

where $\lambda_i$, $i=1\cdots n$, are the eigenvalues of $XA$. This problem has been studied by Tracy and Singh (1972), but not solved satisfactorily. See also Olkin and Sampson (1972).

The inverse transformation

$$Y=X^{-1},$$

for symmetric $X$ gives

$$dv(Y)=-L(X^{-1}\otimes X^{-1})Ddv(X).$$

Disregarding the minus sign, the Jacobian is (Lemma 4.4 (iv))

$$|J(Y,X)|=|L(X^{-1}\otimes X^{-1})D|=|X|^{-(n+1)}.$$

See Jack (1968), Zellner (1971, pp. 226 and 395), and Olkin and Sampson (1972). Zellner assumes (unnecessarily) that $X$ is positive definite.

More interesting is the transformation, again for $X=X'$,

$$Y=|X|X^{-1}.$$

Totally differentiating yields

$$\begin{aligned}dY&=(d|X|)X^{-1}+|X|dX^{-1}\\&=|X|(\text{tr}X^{-1}dX)X^{-1}-|X|X^{-1}(dX)X^{-1},\end{aligned}$$

so that

$$\begin{aligned}d\text{vec}Y&=|X|\left[(\text{vec}X^{-1})(\text{vec}X^{-1})'d\text{vec}X-(X^{-1}\otimes X^{-1})d\text{vec}X\right]\\&=-|X|\left[X^{-1}\otimes X^{-1}-(\text{vec}X^{-1})(\text{vec}X^{-1})'\right]d\text{vec}X,\end{aligned}$$

and

$$dv(Y)=-|X|L\left[X^{-1}\otimes X^{-1}-(\text{vec}X^{-1})(\text{vec}X^{-1})'\right]Ddv(X).$$

The Jacobian is

$$\begin{aligned}|J(Y,X)|&=|X|^{(1/2)n(n+1)}\left|L\left[X^{-1}\otimes X^{-1}-(\text{vec}X^{-1})(\text{vec}X^{-1})'\right]D\right|\\&=|X|^{(1/2)n(n+1)}(1-n)|X|^{-(n+1)}\qquad\text{(by Lemma 4.7 (iii))}\\&=-(n-1)|X|^{(1/2)(n+1)(n-2)}.\end{aligned}$$

See Deemer and Olkin (1951) for a solution along completely different lines, assuming $X$ to be positive definite rather than only symmetric.

As a final example of the usefulness of $L$ and $D$ in evaluating Jacobians of transformations with symmetric matrix arguments, consider

$$Y=X^p, \qquad p=2,3,\cdots.$$

Upon differentiating we find

$$\begin{aligned} dY &= (dX)X^{p-1}+X(dX)X^{p-2}+\cdots+X^{p-1}(dX) \\ &= \sum_{h=1}^{p} X^{h-1}(dX)X^{p-h}, \end{aligned}$$

which gives

$$d\operatorname{vec}Y=\sum_{h=1}^{p}(X^{p-h}\otimes X^{h-1})d\operatorname{vec}X,$$

and

$$dv(Y)=L\sum_{h=1}^{p}(X^{p-h}\otimes X^{h-1})Ddv(X),$$

so that the Jacobian matrix is

$$J(Y,X)=L\sum_{h=1}^{p}(X^{p-h}\otimes X^{h-1})D,$$

the determinant of which is given in Lemma 4.5 (vi).

Summarizing, we have considered six relations between the $\frac{1}{2}n(n+1)$ variables of a symmetric matrix $Y$ and the $\frac{1}{2}n(n+1)$ variables of a symmetric matrix $X$. The results are given in Table 6.1.

Let us now investigate transformations involving *lower triangular* matrix arguments. Consider the relation between lower triangular $Y$ and lower triangular $X$ given by

$$Y=PXQ,$$

where $P$ and $Q$ are also lower triangular. We find

$$d\operatorname{vec}Y=(Q'\otimes P)d\operatorname{vec}X,$$

and thus

$$dv(Y)=L(Q'\otimes P)L'dv(X).$$

Hence, the Jacobian is

$$|J(Y,X)|=|L(Q'\otimes P)L'|=\prod_i p_{ii}^i q_{ii}^{n-i+1} \qquad \text{(Lemma 4.3 (i))}.$$

This problem has been solved by Olkin and Sampson (1972), Deemer and Olkin (1951) for $Q=I$, and Olkin (1953) for $P=I$.

More general is the transformation

$$Y=PXQ+RXS,$$

with lower triangular $P, Q, R, S$. This leads to

$$dv(Y)=L(Q'\otimes P+S'\otimes R)L'dv(X),$$

and the Jacobian follows from Lemma 4.3 (ii).

TABLE 6.1
*Jacobians of transformations with symmetric matrix arguments*

| | Transformation | Jacobian $\lvert J(Y,X)\rvert$ | Conditions, particularities |
|---|---|---|---|
| (i) | $Y=AXA'$ | $\lvert A\rvert^{n+1}$ | $\lvert A\rvert\neq 0$ |
| (ii) | $Y=AXA'+BXB'$ | $\lvert A\rvert^{n+1}\prod_{i\geq j}(1\pm\lambda_i\lambda_j)$ | $\lvert A\rvert\neq 0, \lambda_i\ (i=1\cdots n)$ eigenvalues of $BA^{-1}$ |
| | | $\prod_{i\geq j}(a_{ii}a_{jj}\pm b_{ii}b_{jj})$ | $A=(a_{ij})$ and $B=(b_{ij})$ lower triangular |
| | | $\prod_{i\geq j}(\mu_i\mu_j\pm\theta_i\theta_j)$* | $AB=BA$, $\mu_i$ and $\theta_i\ (i=1\cdots n)$ eigenvalues of $A$ and $B$ |
| (iii) | $Y=XAX$ | $2^n\lvert A\rvert\lvert X\rvert\prod_{i>j}(\lambda_i+\lambda_j)$ | $A=A', \lambda_i\ (i=1\cdots n)$ eigenvalues of $XA$ |
| (iv) | $Y=X^{-1}$ | $\lvert X\rvert^{-(n+1)}$ | $\lvert X\rvert\neq 0$ |
| (v) | $Y=\lvert X\rvert X^{-1}$ | $(n-1)\lvert X\rvert^{(1/2)(n+1)(n-2)}$ | $\lvert X\rvert\neq 0$ |
| (vi) | $Y=X^p\ (p=2,3,\cdots)$ | $p^n\lvert X\rvert^{p-1}\prod_{i>j}\mu_{ij}$ | $\mu_{ij}=\begin{cases}(\lambda_i^p-\lambda_j^p)/(\lambda_i-\lambda_j), & \text{if } \lambda_i\neq\lambda_j,\\ p\lambda_i^{p-1}, & \text{if } \lambda_i=\lambda_j,\end{cases}$ where $\lambda_i\ (i=1\cdots n)$ are eigenvalues of $X$ |

*See the remarks under Lemmas 4.4 (vi) and 4.6 (iii).

Next, we consider the relation between symmetric $Y$ and lower triangular $X$ given by

$$Y=B'XA+A'X'B.$$

Using the same technique, we have

$$\begin{aligned} d\operatorname{vec}Y&=(A'\otimes B')d\operatorname{vec}X+(B'\otimes A')d\operatorname{vec}X'\\ &=[A'\otimes B'+(B'\otimes A')K]d\operatorname{vec}X \qquad \text{(Definition 2.1a)}, \end{aligned}$$

and

$$dv(Y)=L[A'\otimes B'+(B'\otimes A')K]L'dv(X).$$

The Jacobian is

$$\begin{aligned} \lvert J(Y,X)\rvert&=\lvert L[A'\otimes B'+(B'\otimes A')K]L'\rvert=\lvert L[(A'\otimes B')+K(A'\otimes B')]L'\rvert\\ &=\lvert 2LN(A'\otimes B')L'\rvert=2^{(1/2)n(n+1)}\lvert L(A\otimes B)NL'\rvert\\ &=2^{(1/2)n(n+1)}\lvert L(A\otimes B)DLNL'\rvert=2^{(1/2)n(n+1)}2^{-(1/2)n(n-1)}\lvert L(A\otimes B)D\rvert\\ &=2^n\lvert L(A\otimes B)D\rvert, \end{aligned}$$

by the definition of $N$ and Lemmas 2.1 (ii), 3.5 (ii), and 3.4 (i). The determinant $\lvert L(A\otimes B)D\rvert$ can of course be evaluated for each specific $A$ and $B$. In particular, if $A=P$ and $B=Q'R$, or $A=PQ'$ and $B=R'$, where $P$, $Q$, and $R$ are lower triangular, we can express this determinant in terms of the diagonal elements of $P$, $Q$, and $R$, by Lemma 4.3 (iii)–(iv). Special cases have been solved by Deemer and Olkin (1951) ($A=P'$ and $B=I$) and by Olkin (1953) ($A=I$ and $B=P$).

Turning now to *nonlinear* transformations involving lower triangular matrix arguments, we first consider the relation

$$Y=XPX,$$

with lower triangular $P$. We find

$$dY=(dX)PX+XP(dX),$$

and

$$dv(Y)=L(X'P'\otimes I+I\otimes XP)L'dv(X).$$

By Lemma 4.3 (ii) the Jacobian is

$$|J(Y,X)|=|L(X'P'\otimes I+I\otimes XP)L'|=\prod_{i\geqq j}(p_{jj}x_{jj}+p_{ii}x_{ii})$$

$$=2^n|P||X|\prod_{i>j}(p_{ii}x_{ii}+p_{jj}x_{jj}).$$

The next transformation is between a symmetric $Y$ and lower triangular $X$,

$$Y=X'AX+XBX',$$

where $A=A'$ and $B=B'$. Proceeding as before we find

$$d\operatorname{vec}Y=(I\otimes X'A+XB\otimes I)d\operatorname{vec}X+(X'A\otimes I+I\otimes XB)d\operatorname{vec}X'$$

$$=2N(I\otimes X'A+XB\otimes I)d\operatorname{vec}X,$$

so that

$$dv(Y)=2LN(I\otimes X'A+XB\otimes I)L'dv(X).$$

The Jacobian is thus

$$|J(Y,X)|=2^{(1/2)n(n+1)}|L(I\otimes AX+BX'\otimes I)NL'|$$

$$=2^n|L(I\otimes AX+BX'\otimes I)D|.$$

Special cases are the transformations $Y=XX'$ ($A=0$, $B=I$) and $Y=X'X$ ($A=I$, $B=0$), for which the Jacobians can be expressed in terms of the diagonal elements of $X$ by Lemma 4.3 (iii)–(iv). See Deemer and Olkin (1951), Olkin (1953), Jack (1966), Olkin and Sampson (1972), and Zellner (1971, p. 392).

The Jacobians of the transformations $Y=X^{-1}$, $Y=|X|X^{-1}$, and $Y=X^p$, $p=2,3,\cdots$, for lower triangular $X$ and $Y$ can be determined in a fashion very similar to their symmetric counterparts. For $Y=X^{-1}$ we find

$$dv(Y)=-L(X'^{-1}\otimes X^{-1})L'dv(X).$$

For $Y=|X|X^{-1}$,

$$dv(Y)=-|X|L\left[X'^{-1}\otimes X^{-1}-\operatorname{vec}X^{-1}(\operatorname{vec}X'^{-1})'\right]L'dv(X),$$

and for $Y=X^p$,

$$dv(Y)=L\sum_{h=1}^{p}\left[(X')^{p-h}\otimes X^{h-1}\right]L'dv(X).$$

The Jacobians of the three transformations are easily recognized as determinants which have been studied in §4 (Lemmas 4.3 (i), 4.7 (i) and 4.3 (vi)).

The above discussion about Jacobians of transformations with lower triangular matrix arguments is summarized in Table 6.2.

TABLE 6.2
*Jacobians of transformations with lower triangular matrix arguments*

| | Transformation | Jacobian $\lvert J(Y,X)\rvert$ | Conditions, particularities |
|---|---|---|---|
| (i) | $Y=PXQ$ | $\Pi_i p_{ii}^i q_{ii}^{n-i+1}$ | $P, Q$ lower triangular |
| (ii) | $Y=PXQ+RXS$ | $\Pi_{i\geq j}(p_{ii}q_{jj}+r_{ii}s_{jj})$ | $P, Q, R, S$ lower triangular |
| (iii) | $Y=B'XA+A'X'B$ | $2^n\lvert L(A\otimes B)D\rvert$ | |
| (iiia) | $Y=R'QXP+P'X'Q'R$ | $2^n\Pi_i(q_{ii}r_{ii})^i p_{ii}^{n-i+1}$ | $P, Q, R$ lower triangular |
| (iiib) | $Y=RXPQ'+QP'X'R'$ | $2^n\Pi_i(p_{ii}q_{ii})^{n-i+1}r_{ii}^i$ | $P, Q, R$ lower triangular |
| (iv) | $Y=XPX$ | $2^n\lvert P\rvert\lvert X\rvert\Pi_{i>j}(p_{ii}x_{ii}+p_{jj}x_{jj})$ | $P$ lower triangular |
| (v) | $Y=X'AX+XBX'$ | $2^n\lvert L(I\otimes AX+BX'\otimes I)D\rvert$ | $A=A', B=B'$ |
| (va) | $Y=XX'$ | $2^n\Pi_i x_{ii}^{n-i+1}$ | |
| (vb) | $Y=X'X$ | $2^n\Pi_i x_{ii}^i$ | |
| (vi) | $Y=X^{-1}$ | $\lvert X\rvert^{-(n+1)}$ | $\lvert X\rvert\neq 0$ |
| (vii) | $Y=\lvert X\rvert X^{-1}$ | $(n-1)\lvert X\rvert^{(1/2)(n+1)(n-2)}$ | $\lvert X\rvert\neq 0$ |
| (viii) | $Y=X^p (p=2,3,\cdots)$ | $p^n\lvert X\rvert^{p-1}\Pi_{i>j}\mu_{ij}$ | $\mu_{ij}=\begin{cases}(x_{ii}^p-x_{jj}^p)/(x_{ii}-x_{jj}) & \text{if } x_{ii}\neq x_{jj},\\ px_{ii}^{p-1} & \text{if } x_{ii}=x_{jj},\end{cases}$ where $x_{ii}\ (i=1\cdots n)$ are the diagonal elements of $X$ |

**7. Matrix equations.** A third area where we can demonstrate the usefulness of the matrices $L$ and $D$ is the solution of matrix equations. Suppose we are given a matrix equation $Y=F(X)$, where we know a priori that $X$ is symmetric (or triangular). We wish to solve $X$ in terms of $Y$. If $Y$ is a one-to-one function of $X$, as in the preceding section on Jacobians, then $X=F^{-1}(Y)$ is the unique solution. If, however, $Y$ is not in one-to-one correspondence with $X$, we have to restrict the solution space of $X$ to symmetric (or triangular) matrices. In other words, we should not solve for $X$, but for $v(X)$. An example may clarify this approach.

LEMMA 7.1. *The vector equation*

$$Q\operatorname{vec}X=\operatorname{vec}A,$$

*where $Q$ and $A$ are $(n^2,n^2)$ and $(n,n)$ matrices respectively, and $X$ is known to be symmetric, has a solution for $X$ if and only if*

$$QD(QD)^+\operatorname{vec}A=\operatorname{vec}A,$$

*in which case the general solution is*

$$\operatorname{vec}X=D(QD)^+\operatorname{vec}A+D\big[I-(QD)^+QD\big]\operatorname{vec}P,$$

*where* $\operatorname{vec}P$ *is an arbitrary* $\frac{1}{2}n(n+1)$*-vector and* $(QD)^+$ *denotes the Moore-Penrose inverse of* $QD$.

*Proof.* Since $X$ is symmetric, we have $\operatorname{vec}X=Dv(X)$ and thus $QDv(X)=\operatorname{vec}A$. The consistency and solution of this system follow from Penrose (1955, p. 409). Thus,

if a solution exists, it has the form

$$v(X)=(QD)^{+}\text{vec}A+\left[I-(QD)^{+}QD\right]\text{vec}P,$$

for arbitrary $P$. Premultiplication with $D$ gives the desired result. □

This problem has also been studied by Vetter (1975, p. 187), but not solved satisfactorily. If $Q$ is nonsingular, a solution exists if and only if $(I-K)Q^{-1}\text{vec}A=0$. The (unique) solution then takes the form $\text{vec}X=Q^{-1}\text{vec}A$. If $LQD$ is nonsingular, we may write the solution, if it exists, as $\text{vec}X=D(LQD)^{-1}L\text{vec}A$. This is Vetter's solution. He assumes nonsingularity of $Q$ and of $LQD$ (neither of which implies the other!) and also tacitly the existence of a solution for $\text{vec}X$. Note that if we know $X$ to be lower triangular rather than symmetric, the solution is obtained by replacing $D$ with $L'$.

As a final example, let us consider a problem which arises in dynamic econometric models. It concerns the equilibrium covariance matrix. We want to find the matrix of partial derivatives of $S$ with respect to $A$ for $S=ASA'+V$, with symmetric $V$ and $S$, when all eigenvalues of $A$ are less than 1 in absolute value. This problem was first studied by Conlisk (1969) who derived $\partial\text{vec}S/\partial a_{ij}$ for each element of $A$ separately. Neudecker (1969) gave a compact expression for $\partial\text{vec}S/\partial\text{vec}A$. His derivation is wrong, but the result is correct.

LEMMA 7.2. *Consider the matrix equation*

$$S=ASA'+V,$$

*when $S$ and $V$ are symmetric $(n,n)$ matrices, and all eigenvalues of $A$ are smaller than* 1 *in absolute value. The partial derivatives of $S$ with respect to $A$ can be expressed as*

$$\left(\frac{\partial\text{vec}S}{\partial\text{vec}A}\right)'=2N(I\otimes I-A\otimes A)^{-1}(AS\otimes I).$$

*The partial derivatives of the distinct elements of $S$ with respect to $A$ can be expressed as*

$$\left(\frac{\partial v(S)}{\partial\text{vec}A}\right)'=2LN(I\otimes I-A\otimes A)^{-1}(AS\otimes I).$$

*Proof.* We take differentials and vecs:

$$dS=A(dS)A'+(dA)SA'+AS(dA)'+dV,$$

$$d\text{vec}S=(A\otimes A)d\text{vec}S+(AS\otimes I)d\text{vec}A+(I\otimes AS)d\text{vec}A'+d\text{vec}V.$$

Using Lemma 2.1 (ii), and the definitions of $K$ and $N$, we have

$$\begin{aligned}(I\otimes I-A\otimes A)d\text{vec}S&=\left[AS\otimes I+(I\otimes AS)K\right]d\text{vec}A+d\text{vec}V\\&=\left[AS\otimes I+K(AS\otimes I)\right]d\text{vec}A+d\text{vec}V\\&=2N(AS\otimes I)d\text{vec}A+d\text{vec}V.\end{aligned}$$

Since the eigenvalues of $A$ are smaller than 1 in absolute value, the matrix $I\otimes I-A\otimes A$ is nonsingular, and

$$(I\otimes I-A\otimes A)^{-1}=\sum_{h=0}^{\infty}(A^h\otimes A^h).$$

Thus, by Lemma 2.1 (v),

$$d\text{vec}S = 2\sum_h (A^h \otimes A^h) N(AS \otimes I) d\text{vec}A + (I \otimes I - A \otimes A)^{-1} d\text{vec}V$$

$$= 2N\sum_h (A^h \otimes A^h)(AS \otimes I) d\text{vec}A + (I \otimes I - A \otimes A)^{-1} d\text{vec}V$$

$$= 2N(I \otimes I - A \otimes A)^{-1}(AS \otimes I) d\text{vec}A + (I \otimes I - A \otimes A)^{-1} d\text{vec}V,$$

and $$dv(S) = L d\text{vec}S = 2LN(I \otimes I - A \otimes A)^{-1}(AS \otimes I) d\text{vec}A$$

$$+ L(I \otimes I - A \otimes A)^{-1} d\text{vec}V. \qquad \square$$

**Appendix: proofs of lemmas in §4.**

*Proof of Lemma* 4.1. Let $A$ and $B$ be $(n, n)$ matrices. We shall write $L(A \otimes B)L'$ in terms of unit vectors, using the explicit definition of $L$ and (2.5).

$$L(A \otimes B)L' = \sum_{i \geqq j} u_{ij}(\text{vec}E_{ij})'(A \otimes B) \sum_{s \geqq t} (\text{vec}E_{st}) u'_{st}$$

$$= \sum_{i \geqq j} \sum_{s \geqq t} (\text{vec}E_{ij})'(A \otimes B)(\text{vec}E_{st}) u_{ij} u'_{st}$$

$$= \sum_{i \geqq j} \sum_{s \geqq t} \text{tr}(E_{ji} B E_{st} A') u_{ij} u'_{st} = \sum_{i \geqq j} \sum_{s \geqq t} a_{jt} b_{is} u_{ij} u'_{st}.$$

If $A = \Lambda$, $B = M$, and $\delta_{ij}$ denotes the Kronecker delta symbol ($\delta_{ij} = 0$ if $i \neq j$, $\delta_{ii} = 1$), we find

$$L(\Lambda \otimes M)L' = \sum_{i \geqq j} \sum_{s \geqq t} \delta_{jt} \delta_{is} \lambda_j \mu_i u_{ij} u'_{st} = \sum_{i \geqq j} \lambda_j \mu_i u_{ij} u'_{ij},$$

which is a diagonal matrix (since $u_{ij} u'_{ij}$ is diagonal) with elements $\mu_i \lambda_j$, $i \geqq j$, and determinant $\prod_{i \geqq j} \mu_i \lambda_j = \prod_i \mu_i^i \lambda_i^{n-i+1}$.

If $A = P$ and $B = Q$, we have $L(P \otimes Q)L' = \sum_{i \geqq j} \sum_{s \geqq t} p_{jt} q_{is} u_{ij} u'_{st}$. Because $P$ and $Q$ are lower triangular, we may restrict the summation to $i \geqq j \geqq t$, $i \geqq s \geqq t$. This implies that the matrix $u_{ij} u'_{st}$ is lower triangular. Hence, $L(P \otimes Q)L'$ is lower triangular. By putting $s = i$ and $t = j$, we find that its diagonal elements are $q_{ii} p_{jj}$, $i \geqq j$, and its determinant is $\prod_{i \geqq j} q_{ii} p_{jj} = \prod_i q_{ii}^i p_{ii}^{n-i+1}$.

Similarly, we can express $L(A \otimes B)D$ in terms of unit vectors, using the explicit definitions of $L$ and $D$. One verifies that

$$L(A \otimes B)D = L(A \otimes B)L' + \Gamma(A, B),$$

where

$$\Gamma(A, B) = \sum_{i \geqq j} \sum_{s > t} a_{js} b_{it} u_{ij} u'_{st}.$$

Consider the matrix $\Gamma(A, B)$. It is easy to see that $\Gamma(\Lambda, M) = 0$. If $A = P$ and $B = Q$, we may restrict the summation to $i \geqq j \geqq s > t$, so that $\Gamma(P, Q)$ is *strictly* lower triangular. If $A = P'$ and $B = Q'$, we may restrict the summation to $s > t \geqq i \geqq j$, so that $\Gamma(P', Q')$ is *strictly* upper triangular. The properties of $L(\Lambda \otimes M)D$, $L(P \otimes Q)D$, and $L(P' \otimes Q')D$ then follow from the properties of $L(\Lambda \otimes M)L'$ and $L(P \otimes Q)L'$. $\square$

*Proof of Lemma* 4.2. Let $P$ and $Q$ be lower triangular and $v(X)$ arbitrary. Remembering that $L'v(X)=\text{vec}\,\overline{X}$ (Lemma 3.3 (i)), we have

$$L'L(P'\otimes Q)L'v(X)=L'L(P'\otimes Q)\text{vec}\,\overline{X}=L'L\text{vec}\,Q\overline{X}P$$
$$=\text{vec}\,Q\overline{X}P=(P'\otimes Q)\text{vec}\,\overline{X}=(P'\otimes Q)L'v(X).$$

Thus,

$$L'L(P'\otimes Q)L'=(P'\otimes Q)L'.$$

Property (ii) follows from repeated application of (i). Further, $D'(P'\otimes Q)L'=D'L'L(P'\otimes Q)L'=L(P'\otimes Q)L'$, since $LD=I$ (Lemma 3.5(i)). Let us now determine the eigenvalues of $L(P'\otimes Q)L'$. We will need the following result.

*Result* A.1. Let $P$ be a lower triangular matrix with *distinct* diagonal elements. Then there exists a lower triangular matrix $S$ with ones on the diagonal such that $S^{-1}PS=\text{dg}(P)$.

*Proof.* Consider the $\frac{1}{2}n(n-1)$ equations in $\frac{1}{2}n(n-1)$ unknowns $(s_{ij}, i>j)$ given by $PS=S\text{dg}(P)$. This gives $p_{ij}+\sum_{h=j+1}^{i}p_{ih}s_{hj}=s_{ij}p_{jj}$, $i>j$, from which we can sequentially solve for $s_{j+1,j}$ $(j=1\cdots n-1)$, $s_{j+2,j}$ $(j=1\cdots n-2),\cdots,s_{n1}$. □

Assume that both $P$ and $Q$ have distinct diagonal elements. Then, by Result A.1, there exist lower triangular matrices $S$ and $T$ with ones on the diagonal such that

$$S^{-1}PS=\text{dg}(P)\quad\text{and}\quad T^{-1}QT=\text{dg}(Q).$$

By repeated application of (i) we see that

$$L(S'\otimes T^{-1})L'L(P'\otimes Q)L'L(S'^{-1}\otimes T)L'$$
$$=L(S'P'S'^{-1}\otimes T^{-1}QT)L'=L(\text{dg}(P)\otimes\text{dg}(Q))L',$$

and

$$L(S'\otimes T^{-1})L'L(S'^{-1}\otimes T)L'=LL'=I.$$

From Lemma 4.1 we know that $L(\text{dg}(P)\otimes\text{dg}(Q))L'$ is a diagonal matrix with elements $q_{ii}p_{jj}$, $i\geqq j$. These, therefore, are the eigenvalues of $L(P'\otimes Q)L'$.

If not all diagonal elements of $P$ and $Q$ are distinct, we can obtain (iii) by way of a limiting relation, starting with $P+\Delta$ and $Q+\Delta$, where $\Delta$ is a diagonal matrix with $\delta^h$ as its $h$th diagonal element. If $\delta$ is sufficiently small, $P+\Delta$ and $Q+\Delta$ will each have distinct diagonal elements. Hence the eigenvalues of $L[(P+\Delta)'\otimes(Q+\Delta)]L'$ are $(q_{ii}+\delta^i)(p_{jj}+\delta^j)$, $i\geqq j$. Letting $\delta\to 0$, we find the desired result. □

*Proof of Lemma* 4.3. Using Lemma 4.2, we have

$$|L(PQ'\otimes R'S)L'|=|L(P\otimes R')(Q'\otimes S)L'|=|L(P\otimes R')L'L(Q'\otimes S)L'|$$
$$=|L(P'\otimes R)L'||L(Q'\otimes S)L'|=\Big(\prod_{i\geqq j}r_{ii}p_{jj}\Big)\Big(\prod_{i\geqq j}s_{ii}q_{jj}\Big)$$
$$=\prod_{i}(r_{ii}s_{ii})^{i}(p_{ii}q_{ii})^{n-i+1}.$$

To prove (ii) we first assume that $P$ and $Q$ are nonsingular. Then, by Lemma 4.2 and $LL'=I$,

$$L(P'\otimes Q+R'\otimes S)L'=L(I\otimes I+R'P'^{-1}\otimes SQ^{-1})(P'\otimes Q)L'$$
$$=L(I\otimes I+R'P'^{-1}\otimes SQ^{-1})L'L(P'\otimes Q)L'$$
$$=(I+L(R'P'^{-1}\otimes SQ^{-1})L')(L(P'\otimes Q)L').$$

Hence,

$$|L(P'\otimes Q+R'\otimes S)L'|=\prod_{i\geqq j}\left(1+\frac{s_{ii}}{q_{ii}}\frac{r_{jj}}{p_{jj}}\right)\prod_{i\geqq j}(q_{ii}p_{jj})=\prod_{i\geqq j}(q_{ii}p_{jj}+s_{ii}r_{jj}).$$

If $P$ or $Q$ is singular, we obtain (ii) starting with $P+\delta I$ or $Q+\delta I$, where $\delta$ is small and $P+\delta I$ or $Q+\delta I$ is nonsingular. To prove (iii) and (iv) we use Lemmas 4.1 (iii) and 4.2 (i) and (iii):

$$|L(P\otimes Q'R)D|=|L(P\otimes Q')(I\otimes R)D|=|L(P\otimes Q')L'L(I\otimes R)D|$$
$$=|L(P'\otimes Q)L'||L(I\otimes R)D|=\Pi_i q_{ii}^i p_{ii}^{n-i+1}\Pi_i r_{ii}^i=\Pi_i(q_{ii}r_{ii})^i p_{ii}^{n-i+1},$$
$$|L(PQ'\otimes R')D|=|L(P\otimes R')(Q'\otimes I)D|=|L(P\otimes R')L'L(Q'\otimes I)D|$$
$$=|L(P'\otimes R)L'||L(Q'\otimes I)D|=\Pi_i r_{ii}^i p_{ii}^{n-i+1}\Pi_i q_{ii}^{n-i+1}=\Pi_i r_{ii}^i(p_{ii}q_{ii})^{n-i+1}.$$

For nonsingular $P, Q, R, S$, we again use Lemma 4.2 (i) to prove (v) as follows.

$$L(PQ'\otimes R'S)L'L(Q'^{-1}\otimes S^{-1})L'L(P^{-1}\otimes R'^{-1})L'$$
$$=L(P\otimes R')L'L(P^{-1}\otimes R'^{-1})L'=LL'=I.$$

Let us now show (vi). Assume that $P$ has distinct diagonal elements. Then there exists a lower triangular matrix $S$ with ones on the diagonal such that $S^{-1}PS=\Lambda$, with $\Lambda=\mathrm{dg}(P)$ to simplify notation (see Result A.1). Thus,

$$L\sum_{h=1}^{H}\left[(P')^{H-h}\otimes P^{h-1}\right]L'=\sum_h L\left[S'^{-1}\Lambda^{H-h}S'\otimes S\Lambda^{h-1}S^{-1}\right]L'$$
$$=\sum_h L(S'^{-1}\otimes S)L'L(\Lambda^{H-h}\otimes\Lambda^{h-1})L'L(S'\otimes S^{-1})L'$$
$$=\left(L(S'^{-1}\otimes S)L'\right)\sum_h\left(L(\Lambda^{H-h}\otimes\Lambda^{h-1})L'\right)\left(L(S'\otimes S^{-1})L'\right).$$

Because

$$\left(L(S'^{-1}\otimes S)L'\right)^{-1}=L(S'\otimes S^{-1})L',$$

we have

$$\left|L\sum_{h=1}^{H}\left[(P')^{H-h}\otimes P^{h-1}\right]L'\right|=\left|\sum_h L(\Lambda^{H-h}\otimes\Lambda^{h-1})L'\right|.$$

Now, from Lemma 4.1 we know that $L(\Lambda^{H-h}\otimes\Lambda^{h-1})L'$ is a diagonal matrix with diagonal elements $\lambda_i^{h-1}\lambda_j^{H-h}$, $i\geqq j$. Hence $\Sigma_{h=1}^H L(\Lambda^{H-h}\otimes\Lambda^{h-1})L'$ is diagonal with elements $\Sigma_h\lambda_i^{h-1}\lambda_j^{H-h}$, $i\geqq j$, and determinant

$$\left|\sum_h L(\Lambda^{H-h}\otimes\Lambda^{h-1})L'\right|=\prod_{i\geqq j}\left(\sum_{h=1}^{H}\lambda_i^{h-1}\lambda_j^{H-h}\right)=\prod_i(H\lambda_i^{H-1})\prod_{i>j}\left(\sum_{h=1}^{H}\lambda_i^{h-1}\lambda_j^{H-h}\right)$$
$$=H^n\left(\prod_i\lambda_i\right)^{H-1}\prod_{i>j}\left(\sum_{h=1}^{H}\lambda_i^{h-1}\lambda_j^{H-h}\right)=H^n|P|^{H-1}\prod_{i>j}\mu_{ij},$$

where $\mu_{ij}=\Sigma_{h=1}^H p_{ii}^{h-1}p_{jj}^{H-h}$ (since $\lambda_i=p_{ii}$)$=(p_{ii}^H-p_{jj}^H)/(p_{ii}-p_{jj})$. The case where not all diagonal elements of $P$ are distinct, say $p_{ii}=p_{jj}$, can be considered to be a limiting case of the situation where $p_{jj}$ approaches $p_{ii}$. Taking the limit as $p_{jj}\to p_{ii}$, we find $\mu_{ij}=Hp_{ii}^{H-1}$. □

*Proof of Lemma* 4.4. Using the properties $DLN=N$, $D=ND$, and $(A\otimes A)N=N(A\otimes A)$—see Lemmas 3.5 (ii), 3.6 (ii), and 2.1 (v)—we have

$$DL(A\otimes A)D=DL(A\otimes A)ND=DLN(A\otimes A)D=N(A\otimes A)D$$
$$=(A\otimes A)ND=(A\otimes A)D.$$

This proves (i). By repeated application of (i) we find (ii). To prove (iii) we note that $L(A\otimes A)D$ and $DL(A\otimes A)$ have the same set of eigenvalues apart from $\frac{1}{2}n(n-1)$ zeroes which belong to the latter matrix. Let $A$ have eigenvalues $\lambda_i$ and eigenvectors $x_i$; then

$$\begin{aligned}DL(A\otimes A)(x_i\otimes x_j+x_j\otimes x_i)&=DL(Ax_i\otimes Ax_j+Ax_j\otimes Ax_i)\\&=\lambda_i\lambda_j DL(x_i\otimes x_j+x_j\otimes x_i)=\lambda_i\lambda_j DL\operatorname{vec}(x_jx_i'+x_ix_j')\\&=\lambda_i\lambda_j\operatorname{vec}(x_jx_i'+x_ix_j') \text{ (by the implicit definition of } D)\\&=\lambda_i\lambda_j(x_i\otimes x_j+x_j\otimes x_i).\end{aligned}$$

Hence, $DL(A\otimes A)$ has eigenvalues $\lambda_i\lambda_j$, $i\geqq j$, plus $\frac{1}{2}n(n-1)$ zeroes, and $L(A\otimes A)D$ has eigenvalues $\lambda_i\lambda_j$, $i\geqq j$. Its determinant is

$$|L(A\otimes A)D|=\prod_{i\geqq j}\lambda_i\lambda_j=\prod_i\lambda_i^{n+1}=|A|^{n+1}.$$

Let us now prove (v) and (vi). Since $D=NL'(LNL')^{-1}$ (Lemma 3.5 (iii)), and again using Lemmas 3.6 (ii) and 2.1 (v), we can write

$$D'(A\otimes A)D=(LNL')^{-1}LN(A\otimes A)D=(LNL')^{-1}L(A\otimes A)D.$$

The properties of $D'(A\otimes A)D$ thus follow from the properties of $LNL'$ (Lemma 3.4) and $L(A\otimes A)D$ (this lemma).

To prove (vii) we first assume that $A$ has distinct eigenvalues. In that case there exists a matrix $T$ such that $T^{-1}AT=\Lambda$, where $\Lambda$ is a diagonal matrix containing the eigenvalues of $A$. From $AB=BA$ we have $T\Lambda T^{-1}B=BT\Lambda T^{-1}$, or $\Lambda M=M\Lambda$, where $M=T^{-1}BT$. Since all $\lambda$'s are distinct by assumption, $M$ is diagonal. Hence it contains the eigenvalues of $B$. We may then write

$$D'(A\otimes B)D=D'(T\Lambda T^{-1}\otimes TMT^{-1})D=D'(T\otimes T)(\Lambda\otimes M)(T^{-1}\otimes T^{-1})D$$
$$=D'(T\otimes T)L'D'(\Lambda\otimes M)DL(T^{-1}\otimes T^{-1})D,$$

by (i). Hence, using (iv), the explicit definition of $N$, and Lemmas 3.5 (iii), 2.1 (ii), 3.4 (i), and 3.6 (ii),

$$\begin{aligned}|D'(A\otimes B)D|&=|D'(\Lambda\otimes M)D|=|(LNL')^{-1}LN(\Lambda\otimes M)D|\\&=|LNL'|^{-1}|\frac{1}{2}L(I+K)(\Lambda\otimes M)D|\\&=|LNL'|^{-1}2^{-(1/2)n(n+1)}|L(\Lambda\otimes M)D+L(M\otimes\Lambda)KD|\\&=2^{-n}|L(\Lambda\otimes M+M\otimes\Lambda)D|.\end{aligned}$$

From Lemma 4.1 we know that $L(\Lambda\otimes M)D$ and $L(M\otimes\Lambda)D$ are diagonal matrices with elements $\mu_i\lambda_j$ and $\lambda_i\mu_j$, $i\geqq j$. The determinant of their sum is $\prod_{i\geqq j}(\mu_i\lambda_j+\lambda_i\mu_j)$,

and thus

$$|D'(A\otimes B)D| = 2^{-n}\prod_{i\geqq j}(\mu_i\lambda_j+\lambda_i\mu_j) = 2^{-n}\prod_i(2\lambda_i\mu_i)\prod_{i>j}(\mu_i\lambda_j+\lambda_i\mu_j)$$
$$= |A||B|\prod_{i>j}(\mu_i\lambda_j+\lambda_i\mu_j).$$

If $A$ has multiple eigenvalues, say $\lambda_i=\lambda_j$, we consider this as a limiting case of the situation where $\lambda_j$ approaches $\lambda_i$. Taking the limit as $\lambda_j\to\lambda_i$ the result follows. □

*Proof of Lemma* 4.5. From Lemma 2.1 (vi) and the properties $ND=D$ and $DLN=N$, follows (i). The proof of (ii) is similar to that of Lemma 4.4 (iii). Property (iii) follows from (ii). Property (iv) follows from (i) since $LD=I$. We find (v) by repeated application of $DL(I\otimes A+A\otimes I)D=(I\otimes A+A\otimes I)D$, and Lemma 4.4 (i). Let us prove (vi). We proceed as in the proof of Lemma 4.3 (vi). If $A$ has distinct eigenvalues (or if $A=A'$), there exists a nonsingular matrix $S$ such that $S^{-1}AS=\Lambda$, where $\Lambda$ is a diagonal matrix containing the eigenvalues of $A$. Thus,

$$L\sum_{h=1}^{H}(A^{H-h}\otimes A^{h-1})D = L\sum_h(S\Lambda^{H-h}S^{-1}\otimes S\Lambda^{h-1}S^{-1})D$$
$$= L(S\otimes S)\sum_h(\Lambda^{H-h}\otimes\Lambda^{h-1})(S^{-1}\otimes S^{-1})D$$
$$= L(S\otimes S)DL\sum_h(\Lambda^{H-h}\otimes\Lambda^{h-1})DL(S^{-1}\otimes S^{-1})D,$$

by Lemmas 4.4 (i) and 4.5 (v). Since $L(S^{-1}\otimes S^{-1})D=(L(S\otimes S)D)^{-1}$, and using Lemma 4.4 (ii), we have

$$\left|L\sum_{h=1}^{H}(A^{H-h}\otimes A^{h-1})D\right| = \left|L\sum_h(\Lambda^{H-h}\otimes\Lambda^{h-1})D\right|.$$

Lemma 4.1 tells us that $L(\Lambda^{H-h}\otimes\Lambda^{h-1})D$ is a diagonal matrix with elements $\lambda_i^{h-1}\lambda_j^{H-h}$, $i\geqq j$. Hence,

$$\left|L\sum_{h=1}^{H}(A^{H-h}\otimes A^{h-1})D\right| = \prod_{i\geqq j}\left(\sum_{h=1}^{H}\lambda_i^{h-1}\lambda_j^{H-h}\right) = H^n|A|^{H-1}\prod_{i>j}\mu_{ij},$$

with

$$\mu_{ij} = \sum_{h=1}^{H}\lambda_i^{h-1}\lambda_j^{H-h} = \frac{(\lambda_i^H-\lambda_j^H)}{(\lambda_i-\lambda_j)}.$$

If $A$ has multiple eigenvalues, $\lambda_i=\lambda_j$ say, we again consider this as a limiting case of the situation where $\lambda_j$ approaches $\lambda_i$. Taking the limit as $\lambda_j\to\lambda_i$ we find $\mu_{ij}=H\lambda_i^{H-1}$.

□

*Proof of Lemma* 4.6. We shall only consider the determinant of the sum of $L(A\otimes A)D$ and $L(B\otimes B)D$. The determinant of their difference is proved in the same way. By Lemma 4.4 (i),

$$L(A\otimes A+B\otimes B)D = \big(I+L(BA^{-1}\otimes BA^{-1})D\big)L(A\otimes A)D.$$

If $BA^{-1}$ has eigenvalues $\lambda_i$, $i=1\cdots n$, $L(BA^{-1}\otimes BA^{-1})D$ has eigenvalues $\lambda_i\lambda_j$, $i\geqq j$,

by Lemma 4.4 (iii), so that, using Lemma 4.4 (iv),

$$|L(A\otimes A+B\otimes B)D|=\prod_{i\geqslant j}(1+\lambda_i\lambda_j)|A|^{n+1}.$$

To prove (ii) we first assume that $A$ is nonsingular. Then,

$$|L(A\otimes A+B\otimes B)D|=|A|^{n+1}\prod_{i\geqslant j}\left(1+\frac{b_{ii}b_{jj}}{a_{ii}a_{jj}}\right)=\prod_{i\geqslant j}(a_{ii}a_{jj}+b_{ii}b_{jj}),$$

since $A$ and $B$ are now lower triangular. If $A$ is singular, we obtain (ii) starting with $A+\delta I$, where $\delta$ is small and $A+\delta I$ is nonsingular.

Consider now case (iii) where $AB=BA$. This result can be proved applying the same method as in the proof of Lemma 4.4 (vii). □

*Proof of Lemma* 4.7. We shall only show (iii) and (iv), as (i) and (ii) can be proved similarly. Since $A$ is symmetric and nonsingular by assumption, we have from the implicit definition of $D$ and Lemma 4.4, $DL\text{vec}A=\text{vec}A$, $DL(A\otimes A)D=(A\otimes A)D$, $|L(A\otimes A)D|=|A|^{n+1}$, and $(L(A\otimes A)D)^{-1}=L(A^{-1}\otimes A^{-1})D$. Thus,

$$\begin{aligned}L(A\otimes A+\alpha\text{vec}A(\text{vec}A)')D&=\left[I+\alpha L\text{vec}A(\text{vec}A)'DL(A^{-1}\otimes A^{-1})D\right]L(A\otimes A)D\\&=\left[I+\alpha L\text{vec}A(\text{vec}A)'(A^{-1}\otimes A^{-1})D\right]L(A\otimes A)D\\&=\left[I+\alpha(L\text{vec}A)(D'\text{vec}A^{-1})'\right]L(A\otimes A)D.\end{aligned}$$

Since for any two vectors $x$ and $y$ of the same order,

$$|I+xy'|=1+y'x\quad\text{and}\quad(I+xy')^{-1}=I-\frac{xy'}{1+y'x},$$

we find

$$\begin{aligned}|I+\alpha(L\text{vec}A)(D'\text{vec}A^{-1})'|&=1+\alpha(\text{vec}A^{-1})'DL\text{vec}A\\&=1+\alpha(\text{vec}A^{-1})'\text{vec}A=1+\alpha\text{tr}A^{-1}A=1+\alpha n,\end{aligned}$$

and

$$\left[I+\alpha(L\text{vec}A)(D'\text{vec}A^{-1})'\right]^{-1}=I-\frac{\alpha}{1+\alpha n}L\text{vec}A(D'\text{vec}A^{-1})'.$$

Hence,

$$|L(A\otimes A+\alpha\text{vec}A(\text{vec}A)')D|=(1+\alpha n)|L(A\otimes A)D|=(1+\alpha n)|A|^{n+1},$$

and

$$\begin{aligned}&\left[L(A\otimes A+\alpha\text{vec}A(\text{vec}A)')D\right]^{-1}\\&\quad=L(A^{-1}\otimes A^{-1})D\left[I-\frac{\alpha}{1+\alpha n}L\text{vec}A(\text{vec}A^{-1})'D\right]\\&\quad=L\left[A^{-1}\otimes A^{-1}-\frac{\alpha}{1+\alpha n}(A^{-1}\otimes A^{-1})DL\text{vec}A(\text{vec}A^{-1})'\right]D\\&\quad=L\left[A^{-1}\otimes A^{-1}-\frac{\alpha}{1+\alpha n}(\text{vec}A^{-1})(\text{vec}A^{-1})'\right]D.\quad\square\end{aligned}$$

## REFERENCES


[1] T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.

[2] P. BALESTRA, *La dérivation matricielle*, Collection de l'IME, no. 12, Sirey, Paris, 1976.

[3] M. BROWNE, *Generalized least squares estimation in the analysis of covariance structures*, South African Statist. J., 8 (1974), pp. 1–24.

[4] J. CONLISK, *The equilibrium covariance matrix of dynamic econometric models*, J. Amer. Statist. Assoc., 64 (1969), pp. 277–279.

[5] W. L. DEEMER AND I. OLKIN, *The Jacobians of certain matrix transformations useful in multivariate analysis*, Biometrika, 38 (1951), pp. 345–367.

[6] H. V. HENDERSON AND S. R. SEARLE, *Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics*, Canad. J. Statist., 7 (1979) pp. 65–81.

[7] H. JACK, *Jacobians of transformations involving orthogonal matrices*, Proc. Roy. Soc. Edinburgh Sect. A, 67 (1966), pp. 81–103.

[8] J. R. MAGNUS AND H. NEUDECKER, *The commutation matrix: some properties and applications*, Ann. Statist., 7 (1979), pp. 381–394.

[9] D. G. NEL, *On the symmetric multivariate normal distribution and the asymptotic expansion of a Wishart matrix*, South African Statist. J., 12 (1978), pp. 145–159.

[10] H. NEUDECKER, *Some theorems on matrix differentiation with special reference to Kronecker matrix products*, J. Amer. Statist. Assoc., 64 (1969), pp. 953–963.

[11] I. OLKIN, *Note on 'The Jacobians of certain matrix transformations useful in multivariate analysis'*, Biometrika, 40 (1953), pp. 43–46.

[12] I. OLKIN AND A. R. SAMPSON, *Jacobians of matrix transformations and induced functional equations*, Linear Algebra and Appl., 5 (1972), pp. 257–276.

[13] R. PENROSE, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

[14] C. R. RAO, *Linear Statistical Inference and its Applications*, 2nd ed., John Wiley, New York, 1973.

[15] J. F. RICHARD, *A note on the information matrix of the multivariate normal distribution*, J. Econometrics, 3 (1975), pp. 57–60.

[16] D. S. TRACY AND P. S. DWYER, *Multivariate maxima and minima with matrix derivatives*, J. Amer. Statist. Assoc., 64 (1969), pp. 1576–1594.

[17] D. S. TRACY AND R. P. SINGH, *Some modifications of matrix differentiation for evaluating Jacobians of symmetric matrix transformations*, Symmetric Functions in Statistics, D. S. Tracy, ed., University of Windsor, Windsor, Ontario, 1972.

[18] W. J. VETTER, *Vector structures and solutions of linear matrix equations*, Linear Algebra and Appl., 10 (1975), pp. 181–188.

[19] A. ZELLNER, *An Introduction to Bayesian Inference in Econometrics*, John Wiley, New York, 1971.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980


0196-5212/80/0104-0009 $1.00/0

# DECOMPOSING A PERMUTATION INTO TWO LARGE CYCLES: AN ENUMERATION*

EDWARD A. BERTRAM† AND VICTOR K. WEI‡

**Abstract.** Let $c_{l,m,\tau}^{(n)}$ denote the number of ways a permutation $\tau$ can be expressed as the product of an $l$-cycle and an $m$-cycle, all in the symmetric group on $n$ symbols. In 1972, the first author gave a necessary and sufficient condition on $l$ such that $c_{l,l,\tau}^{(n)}>0$ for every even permutation $\tau$. In 1978, G. Boccara gave a necessary and sufficient condition on $l$, $m$, and $\tau$ such that $c_{l,m,\tau}^{(n)}>0$. More recently, D. W. Walkup developed a recursion for $c_{n,n,\tau}^{(n)}$. In this paper, we show how to recursively calculate the values of $c_{n,n-i,\tau}^{(n)}$. Theorem 1 states that $c_{n,n-1,\tau}^{(n)}=2\cdot(n-2)!$ for every odd $\tau$. Theorem 2 exhibits $c_{n+1,n-i,\sigma}^{(n+1)}$ as a linear combination (with easily obtained integral coefficients) of a specified set of $c_{n,n-i,\tau}^{(n)}$. Applications include a method to evaluate, by inverting an integral triangular matrix, all values in $\{c_{n,n,\tau}^{(n)}: \tau$ has exactly $k$ disjoint cycles$\}$, for arbitrary $k\leqq n$.

**1. Introduction.** Let $C_\rho$ denote the conjugacy class containing the permutation $\rho$ in the symmetric group $S_n$ on the symbols $\{1,2,\ldots,n\}$. If $C_\rho$ and $C_\sigma$ are two classes in $S_n$, $C_\rho C_\sigma$ is the collection of all permutations (counting repetitions) which are the product (composition) $\alpha\beta$ of permutations $\alpha\in C_\rho$ and $\beta\in C_\sigma$. In the group algebra of $S_n$ over the complex numbers, $K_\rho$ will denote the sum of permutations in $C_\rho$. Multiplying in the center of the group algebra, we know that $K_\rho K_\sigma=\sum_{K_\tau} c_{\rho\sigma\tau}^{(n)} K_\tau$, where each $c_{\rho\sigma\tau}^{(n)}$ is a nonnegative integer, the number of decompositions of *each* permutation in $C_\tau$ into a product $\alpha\beta$ of permutations $\alpha\in C_\rho$ and $\beta\in C_\sigma$.

For integers $l, m$: $1\leqq l, m\leqq n$, let $c_{l,m,\tau}^{(n)}$ denote the number of ways of expressing $\tau\in S_n$ as a product $\rho\sigma$ of an $l$-cycle $\rho$ and an $m$-cycle $\sigma$, where $\rho,\sigma\in S_n$. In [1] the first author proved that $c_{l,l,\eta}^{(n)}>0$, for *each* even permutation $\eta\in A_n$, if and only if $\lfloor \frac{3}{4}n \rfloor \leqq l\leqq n$ (here $\lfloor x \rfloor$ denotes the greatest integer $\leqq x$). Similar necessary and sufficient conditions were also given on $l$, in order that $c_{l,l-1,\omega}^{(n)}>0$ for each odd permutation $\omega\in S_n$. However, very recently a recursion for $c_{n,n,\eta}$ was developed by D. W. Walkup [10].

Extending the above results, G. Boccara [2] has recently given necessary and sufficient conditions on $s, d, \tau$ in order that $\tau$ admit a decomposition into a product of two (possibly nondisjoint) cycles, the sum and difference of whose lengths are $s$ and $d$, respectively. When $\rho$ and $\sigma$ are both reflections in $S_n$ (i.e., involutions: $\rho^2=\sigma^2=$ identity), G. Moran [7] first gave necessary and sufficient conditions on $\tau$ in order that $c_{\rho\sigma\tau}^{(n)}>0$, and in [8] obtained generating functions for such $c_{\rho\sigma\tau}^{(n)}$.

In this paper we show how to recursively determine the value of $c_{n+1,n-i,\sigma}^{(n+1)}$ for each $i$: $0\leqq i\leqq n-1$ and each $\sigma\in S_{n+1}$, from certain earlier values $c_{n,n-i,\tau}^{(n)}$. Using direct combinatorial reasoning in §2 (in contrast to the use of character theory in §3), we prove in Theorem 1 that $c_{n,n-1,\tau}^{(n)}=2[(n-2)!]$, independent of the odd permutation $\tau\in S_n$. Our main result is Theorem 2, where we show that for each $\sigma\in S_{n+1}, 0\leqq i\leqq n-1, (i+1)c_{n+1,n-i,\sigma}^{(n+1)}$ is a linear combination of certain previously obtained $c_{n,n-i,\tau}^{(n)}$, where the permutations $\tau$ belong to specified conjugacy classes in the group $S_n$ (or any subgroup of permutations in $S_{n+1}$ which fix a symbol $k$) and where the coefficients are easily obtained integers.

---

*Received by the editors November 30, 1979 and in final form February 28, 1980.

†Department of Mathematics, University of Hawaii at Manoa, 2565 The Mall, Honolulu, Hawaii 96822.

‡Bell Laboratories, Murray Hill, New Jersey 07974. The research of this author was supported by the National Science Foundation under Grant ENG78-05151.

As a first application of Theorem 2, when $i=0$, we find (in Corollaries 2.1 and 2.2) the number $c^{(n)}_{n,n,\tau}$ of decompositions of $\tau$ (in $A_n$, the subgroup of even permutations) into the product of two $n$-cycles, for a variety of classes $C_\tau$. For example, $c^{(n)}_{n,n,\tau}=2[(n-1)!]/(n-l+1)$ whenever $\tau$ belongs to a class in $S_n$ represented by any partition of the form $[k^m, k-1, 1^l]$ with $k$ and $m$ even. We also show, by examples, how one may proceed to recursively develop expressions for $c^{(n)}_{n,n,\tau}$ when $\tau$ belongs to any class represented by a partition of the form $[k, 2^m, 1^l]$.

Next, let $p(k,n)$ denote the number of (unordered) partitions of $n$ into exactly $k$ parts. Then there are $p(k,n)$ classes $C_\tau$ (in $S_n$) of permutations $\tau$ with exactly $k$ disjoint cycles (including, possibly, 1-cycles). Note that when $k\not\equiv n \pmod 2$, then $\tau$ is an odd permutation and $c^{(n)}_{n,n,\tau}=0$. When $k\equiv n \pmod 2$ we exhibit $p(k,n)$ linearly independent equations with integer coefficients, in the $p(k,n)$ variables $c^{(n)}_{n,n,\tau}$, for the $C_\tau$ just described. The computation of these $c^{(n)}_{n,n,\tau}$ is shown to require only the inverting of an integral triangular matrix.

As further implications of Theorem 2 we derive, in Corollary 2.3, explicit formulas for $c^{(n)}_{n,n-2,\sigma}$, $c^{(n)}_{n,n-3,\sigma}$, and $c^{(n)}_{n,n-4,\sigma}$, and show that $c^{(n)}_{n,n-i,\sigma}$, $i\geqq 2$, depends only on ($n$ and) the number of $j$-cycles, for $1\leqq j\leqq i-1$, in the disjoint cycle decomposition of $\sigma$ in $S_n$.

On the other hand, the integers $c^{(n)}_{\rho\sigma\tau}$ can in principle be computed from the character tables of $S_n$, and in §3 we illustrate how this method yields special cases of some of our earlier results. We show how classical recurrence relations give expressions for the necessary character values, and how we are led to a variety of sums involving binomial coefficients. Our point here is to contrast the lack of unity in these computational methods with the scope and simplicity of the methods and results of §2.

**2. Combinatorial derivation of $c^{(n)}_{n,n-i,\tau}$.** For $n\geqq 3$, there are clearly $(n-1)!\cdot n\cdot(n-2)!$ (odd) permutations in $S_n$, counting repetitions, which are a product $\alpha\beta$ of an $n$-cycle $\alpha$, and $(n-1)$-cycle $\beta, \alpha, \beta\in S_n$. If *every* odd permutation in $S_n$ has at least $2[(n-2)!]$ decompositions into such a product then, since $(n!/2)\cdot 2[(n-2)!]=(n-1)!n(n-2)!$, every odd permutation has *exactly* $2[(n-2)!]$ decompositions.

THEOREM 1. *For $n\geqq 3$, every odd permutation in $S_n$ has at least, and therefore exactly, $2(n-2)!$ decompositions into a product $\alpha\beta$ of an $n$-cycle $\alpha$ and an $(n-1)$-cycle $\beta, \alpha, \beta\in S_n$.*

*Proof.* We need only prove the result for a fixed representative of each class of odd permutations in $S_n$. For $n=3$ the transposition $(12)(3)=(132)(13)=(123)(23)$ (composing will always be done from right to left) has $2(n-2)!=2$ such decompositions. Now suppose $n\geqq 4$, and let $\sigma$ be an *odd* permutation in $S_n$; we proceed by induction. Since $\sigma$ is not the identity permutation we may assume that $\sigma(1)=2$. Now put $\sigma'=(t\ 2\ 1)\sigma$, where $t\in\{3,4,\ldots,n\}$; then $\sigma'(1)=1$ and $\sigma'$ is an odd permutation on the set $\{2,3,\ldots,n\}$. By our induction hypothesis $\sigma'$ has $2(n-3)!$ decompositions of the form $\sigma'=\gamma\delta, \gamma(1)=\delta(1)=1$, $\gamma$ is an $(n-1)$-cycle and $\delta$ is an $(n-2)$-cycle. Suppose $\delta(t)\neq t$. Then $\sigma=(12t)\gamma\delta=(2t)[(1t)\gamma(1t)](1t)\delta$. Also, $(1t)\delta$ is an $(n-1)$-cycle and $(1t)\gamma(1t)$ is an $(n-1)$-cycle which fixes (only) $t$; hence $(2t)[(1t)\gamma(1t)]$ is an $n$-cycle. On the other hand, suppose $\delta(t)=t$. Then $\delta(2)\neq 2$, and $\sigma=(12t)\gamma\delta=(12)[(2t)\gamma(2t)](2t)\delta$. Here $(2t)\delta$ is an $(n-1)$-cycle, and $(2t)\gamma(2t)$ is an $(n-1)$-cycle which fixes (only) 1; hence $(12)[(2t)\gamma(2t)]$ is an $n$-cycle. Thus, for *each* $t\in\{3,4,\ldots,n\}$, we have found $2(n-3)!$ decompositions of $\sigma$ into the product of an $n$-cycle and an $(n-1)$-cycle. We will have found $(n-2)\cdot 2\cdot(n-3)!$ decompositions of $\sigma$, once it is shown that they are all different. Suppose, for example, that $\sigma=(1\ 2\ s)\gamma_1\delta_1=\varepsilon_1(1\ s)\delta_1$, and $\sigma=(1\ 2$

$t)\gamma_2\delta_2 = \varepsilon_2(1\ t)\delta_2$, where the $\varepsilon_i$ are $n$-cycles, the $\delta_i$ are $(n-2)$-cycles which fix the symbols 1 and 2, and $s \neq t$, $s, t \in \{3,4,\ldots,n\}$. Then the $(n-1)$-cycles $(1\ s)\delta_1$ and $(1\ t)\delta_2$ are *different*, since they each map 1 to a different symbol. On the other hand, suppose $\sigma = (1\ 2\ s)\gamma_1\delta_1 = \varepsilon_1(1\ s)\delta_1$, and $\sigma = (1\ 2\ t)\gamma_2\delta_2 = \varepsilon_2(2\ t)\delta_2$, where the $\varepsilon_i$ are $n$-cycles, $\delta_1$ is an $(n-2)$-cycle which fixes the symbols 1 and 2, $\delta_2$ is an $(n-2)$-cycle which fixes the symbols 1 and $t$, and $s \neq t$ are in $\{3,4,\ldots,n\}$. Then $(1\ s)\delta_1$ and $(2\ t)\delta_2$ are *different* $(n-1)$-cycles since the latter fixes 1, while the former maps 1 to $s$. The cases where $s = t$ are handled in a similar manner, and there are $(n-2)\cdot 2\cdot(n-3)! = 2(n-2)!$ decompositions of $\sigma$. □

Let $S_{n+1}(k)$ denote the subgroup of all those permutations in $S_{n+1}$ which fix the symbol $k, 1 \leqq k \leqq n+1$. Then $S_{n+1}(k)$ is isomorphic to $S_n$ for each $k$, a simple but important fact. Thus the number of representations of $\tau \in S_{n+1}(k)$ as a product $\rho\sigma$ of an $l$-cycle $\rho$ and $m$-cycle $\sigma$, $\rho$ and $\sigma$ in $S_{n+1}(k)$, is equal to $c^{(n)}_{l,m,\tau'}$, where $\tau'$ is any permutation in $S_n$ with the same disjoint cycle structure as $\tau$.

Now suppose that $D$ is a class in $S_{n+1}(k)$, $k$ arbitrary: $1 \leqq k \leqq n+1$. We call the class $C \subseteq S_{n+1}$ an *ascendant class* of $D$ (and $D$ a *descendant* of $C$) if $\sigma \in C$ and $\tau \in D$ have the same number of cycles of each length, except that exactly one cycle of $\sigma$ has length one more than a corresponding cycle of $\tau$. Thus, whenever $n = \lambda_1 + \lambda_2 + \cdots + \lambda_r$ is a partition of $n$ into $r$ parts, $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_r \geqq 1$, and the class $D$ is represented by $[\lambda_1, \lambda_2, \ldots, \lambda_r]$, then the ascendant classes of $D$ are the classes in $S_{n+1}$ given by: $[\lambda_1 + 1, \lambda_2, \ldots, \lambda_r]$, $[\lambda_1, \lambda_2 + 1, \lambda_3, \ldots, \lambda_r]$, $[\lambda_1, \lambda_2, \lambda_3 + 1, \lambda_4, \ldots, \lambda_r]$, etc. Of course, if there are duplications among some of the $\lambda_i$ for $D$, then there will be fewer than $r$ ascendant classes of $D$. For example, if $D$ (in any $S_{14}(k)$) is represented by the partition [5, 3, 2, 2, 1], then $D$ has only 4 ascendant classes: [6, 3, 2, 2, 1], [5, 4, 2, 2, 1], [5, 3, 3, 2, 1], [5, 3, 2, 2, 2] in $S_{14}$, and $D$ is a descendant class of each of these. Note also that [6, 3, 2, 2] (in any $S_{14}(k)$) is *not* a descendant class of [6, 3, 2, 2, 1].

Alternatively, we may represent $C \subseteq S_{n+1}$ by the partition $[1^{n_1}2^{n_2}\ldots]$ and $D \subseteq S_{n+1}(k)$ by the partition $[1^{n'_1}\ 2^{n'_2}\ldots]$. Then $C$ is an ascendant class of $D$, and $D$ a descendant class of $C$, if and only if, for exactly one value of $r$, $n'_r = n_r - 1$ and $n'_{r-1} = n_{r-1} + 1$.

We define the *ascendancy factor*: $a(D, C)$ from $D \subseteq S_{n+1}(k)$ to $C$ in $S_{n+1}$ by:

$$a(D,C) = \begin{cases} 0, & \text{if } C \text{ is not an ascendant class of } D, \\ s \cdot n_s, & \text{if } C = \left[\ldots, s^{n_s - 1}, (s+1)^{n_{s+1}+1}, \ldots\right] \text{ is an ascendant} \\ & \text{class of } D = \left[\ldots, s^{n_s}, (s+1)^{n_{s+1}}, \ldots\right]. \end{cases}$$

Correspondingly, we define the *descendancy factor*: $d(C, D)$ from $C$ to $D$ by:

$$d(C,D) = \begin{cases} 0, & \text{if } D \text{ is not a descendant class of } C, \\ r \cdot n_r, & \text{if } D = \left[\ldots, (r-1)^{n_{r-1}+1}, r^{n_r - 1}, \ldots\right] \text{ is a descendant} \\ & \text{class of } C = \left[\ldots, (r-1)^{n_{r-1}}, r^{n_r}, \ldots\right]. \end{cases}$$

*Example.* If $D \subseteq S_{14}(k)$ is represented by [5, 3, 2, 2, 1] and $C \subseteq S_{14}$ is represented by [5, 3, 3, 2, 1], then $a(D, C) = 2 \cdot 2 = 4$ and $d(C, D) = 3 \cdot 2 = 6$.

LEMMA 1. *For each pair of classes, $C \subseteq S_{n+1}$ and $D \subseteq S_{n+1}(k)$, $|C|d(C, D) = (n+1)|D|a(D, C)$, where $|C|$ denotes the cardinality of $C$.*

*Proof.* Let $C = [1^{n_1}\ 2^{n_2}\ldots i^{n_i}\ldots]$ and $D = [1^{n'_1}\ 2^{n'_2}\ldots]$. If $D$ is not a descendant class of $C$, then both sides $= 0$. If $D$ is a descendant class of $C$, then $n'_i = n_i$ except for exactly one value of $r$, where $n'_r = n_r - 1$ and $n'_{r-1} = n_{r-1} + 1$. But $|C| =$

$(n+1)!/(\Pi i^{n_i}\Pi n_i!)$ and $|D|=n!/(\Pi i^{n'_i}\Pi n'_i!)$ (see, e.g., [6, page 40]). Hence $|C|/|D|=[(n+1)(r-1)(n_{r-1}+1)]/(r\cdot n_r)$. But $a(D,C)=(r-1)(n_{r-1}+1)$ and $d(C,D)=r\cdot n_r$, from the definitions of $a(D,C)$ and $d(C,D)$, so the proof is complete. □

In the following preparation for the proof of Theorem 2, fix $n\geqq 2$, $i$: $0\leqq i\leqq n-1$, and $k$: $1\leqq k\leqq n+1$.

DEFINITIONS. If the permutation $\nu=(x_1x_2\dots x_{n+1})(y_1y_2\dots y_{n-i})$, with both cycles in $S_{n+1}$, then the ordered pair $((x_1x_2\dots x_{n+1}),(y_1y_2\dots y_{n-i}))$ is called an $S_{n+1}$-*decomposition* of $\nu$. If $\mu=(a_1a_2\dots a_n)(b_1b_2\dots b_{n-i})$, with both cycles in $S_{n+1}(k)$, then ($\mu\in S_{n+1}(k)$ and) the ordered pair $((a_1a_2\dots a_n),(b_1b_2\dots b_{n-i}))$ is called a $(k)$-*decomposition* of $\mu$. Such $S_{n+1}$-decompositions of $\nu$ and $(k)$-decompositions of $\mu$ are said to be *coupled* if (as permutations):

(1) $(a_1a_2\dots a_n)=(x_1\dots x_{j-1}x_{j+1}\dots x_{n+1})$, for some $j$, and

(2) $(b_1b_2\dots b_{n-i})=(y_1y_2\dots y_{n-i})$.

Thus, if these decompositions are coupled, $(x_jx_{j+1})\nu=\mu$ and the class $C_\mu$ is a descendant class of $C_\nu$.

Suppose $\sigma$ is any permutation in $S_{n+1}$, and $C$ the conjugacy class of $\sigma$. We let $\mathfrak{Q}(\sigma)$ denote the set of all $S_{n+1}$-decompositions of $\sigma$. Clearly, $|\mathfrak{Q}(\sigma)|=|\mathfrak{Q}(\sigma')|=c^{(n)}_{n+1,n-i,\sigma}$, and $\mathfrak{Q}(\sigma)\cap\mathfrak{Q}(\sigma')=\varnothing$, if $\sigma\neq\sigma'$ are both in $C$. Let $\mathfrak{Q}(C):=\cup_{\sigma'\in C}\mathfrak{Q}(\sigma')$; we thus have $|\mathfrak{Q}(C)|=|C|c^{(n+1)}_{n+1,n-i,\sigma}$ (as defined in the introduction, $c^{(n+1)}_{n+1,n-i,\sigma}$ is the number of $S_{n+1}$-decompositions of $\sigma$). Fix $k$: $1\leqq k\leqq n+1$, and let $\mathfrak{D}_k(C)$ denote the set of all descendant classes $D$ of $C$, $D\subseteq S_{n+1}(k)$. For $\tau\in D$, let $\mathfrak{P}_k(\tau)$ denote the set of all $(k)$-decompositions of $\tau$. Set $\mathfrak{P}_k(D):=\cup_{\tau\in D}\mathfrak{P}_k(\tau)$ and $\mathfrak{P}_k:=\cup\mathfrak{P}_k(D)$, the latter union over all descendant classes $D\in\mathfrak{D}_k(C)$. If $\tau\neq\tau'$ are any two permutations in $D$, $\mathfrak{P}_k(\tau)\cap\mathfrak{P}_k(\tau')=\varnothing$, and $|\mathfrak{P}_k(\tau)|=|\mathfrak{P}_k(\tau')|=c^{(n)}_{n,n-i,\tau}$. Thus $|\mathfrak{P}_k(D)|=|D|c^{(n)}_{n,n-i,\tau}$. If $D\neq D'$ are any two descendant classes in $\mathfrak{D}_k(C)$, then $\mathfrak{P}_k(D)\cap\mathfrak{P}_k(D')=\varnothing$, so that $|\mathfrak{P}_k|=\Sigma_D|\mathfrak{P}_k(D)|=\Sigma_{D\in\mathfrak{D}_k(C)}|D|c^{(n)}_{n,n-i,\tau}$, where $\tau$ is any permutation in $D$. Finally, let $\mathfrak{P}(C):=\cup_{k=1}^{n+1}\mathfrak{P}_k$. Since $\mathfrak{P}_j\cap\mathfrak{P}_k=\varnothing$, $j\neq k$, and $|\mathfrak{P}_j|=|\mathfrak{P}_k|$, we have $|\mathfrak{P}(C)|=(n+1)|\mathfrak{P}_k|$, for *each* $k$, so $|\mathfrak{P}(C)|=(n+1)\Sigma_{D\in\mathfrak{D}_1}|D|c^{(n)}_{n,n-i,\tau}$, $\tau$ any permutation in $D$.

THEOREM 2. *For each* $i$: $0\leqq i\leqq n-1$, *and* $\sigma\in S_{n+1}$, *we have*

$$\sum_{D\subseteq S_n} d(C,D)c^{(n)}_{n,n-i,\tau}=(i+1)c^{(n+1)}_{n+1,n-i,\sigma},$$

*where* $C$ *is the class* ($\subseteq S_{n+1}$) *of* $\sigma$ *and* $\tau\in D$ *is arbitrary.*

*Remark* 1. Since $d(C,D)=0$ when $D$ is not a descendant class of $C$, the left-hand side is a summation over descendant classes of $C$. Also, since $S_{n+1}(k)\cong S_n$ for every $k$: $1\leqq k\leqq n+1$, we may replace $\Sigma_{D\subseteq S_n}$ by $\Sigma_{D\in\mathfrak{D}_1(C)}$, or $\Sigma_{D\in\mathfrak{D}_2(C)}$, or $\dots\Sigma_{D\in\mathfrak{D}_{n+1}(C)}$.

*Remark* 2. Note that $c^{(n+1)}_{n+1,n-i,\sigma}>0$ if and only if $\operatorname{sgn}(\sigma)=(-1)^{i+1}$. Otherwise, both sides are zero and equality holds.

*Proof of the theorem.* Recall that $\mathfrak{Q}(C)$ is the set of all $S_{n+1}$-decompositions (into an $(n+1)$-cycle and an $(n-i)$-cycle) of all permutations in the class $C$ of $\sigma$; thus $|\mathfrak{Q}(C)|=|C|c^{(n+1)}_{n+1,n-i,\sigma}$. $\mathfrak{P}(C)$ is the set union, over all $k$: $1\leqq k\leqq n+1$, of all $(k)$-decompositions of all permutations in all descendant classes of $C$ which belong to $S_{n+1}(k)$; and $|\mathfrak{P}(C)|=(n+1)\Sigma_{D\in\mathfrak{D}_k(C)}|D|c^{(n)}_{n,n-i,\tau}$. We will use our coupling relation between $\mathfrak{Q}(C)$ and $\mathfrak{P}(C)$ to prove the theorem.

First, we claim that each decomposition in $\mathfrak{Q}(C)$ is coupled with exactly $i+1$ decompositions in $\mathfrak{P}(C)$. For $((x_1\dots x_{n+1}),(y_1\dots y_{n-i}))\in\mathfrak{Q}(C)$ is coupled to $((x_1\dots x_{j-1}x_{j+1}\dots x_{n+1}),(y_1\dots y_{n-i}))\in\mathfrak{P}_{x_j}(C)\subseteq\mathfrak{P}(C)$ if and only if $x_j\notin\{y_1,\dots,y_{n-i}\}$, i.e., if and only if $x_j$ is one of the $i+1$ members of $\{1,2,\dots,n+1\}-\{y_1,y_2,\dots,y_{n-i}\}$. Secondly, if $D$ is a descendant class of $C$, $D\subseteq S_{n+1}(k)$, then each $(k)$-decomposition

of each permutation $\mu \in D$ is coupled with exactly $a(D, C)$-decompositions in $\mathfrak{Q}(C)$: for, suppose $((a_1 \dots a_n),(b_1 \dots b_{n-i}))$ is a $(k)$-decomposition of $\mu$. Then $k \notin \{a_1, \dots, a_n\}$, $(k\ a_j)\mu = (a_1 \dots a_{j-1}\ k\ a_j \dots a_n)(b_1 \dots b_{n-i})$, and the latter pair are the *only* type of $S_{n+1}$-decompositions in $\mathfrak{Q}(C)$ which may be coupled to the given $(k)$-decomposition. Furthermore, the disjoint cycle structure of $(k\ a_j)\mu$ differs from that of $\mu$ precisely by the addition of the symbol $k$ to that cycle of $\mu$ which contains $a_j$. Since $C$ is an ascendant class of $D$, there are exactly $a(D, C)$ choices of $a_j \in \{1, 2, \dots, k-1, k+1, \dots, n+1\}$ such that $(k\ a_j)\mu \in C$. Hence there are exactly $a(D, C)$ decompositions in $\mathfrak{Q}(C)$ coupled with each $(k)$-decomposition of each $\mu \in D$. We now form a bipartite graph, joining by an edge a decomposition $q \in \mathfrak{Q}(C)$ with a decomposition $p \in \mathfrak{P}(C)$ if and only if these decompositions are coupled, and then count the total number of edges. This is: $\Sigma_{p \in \mathfrak{P}(C)} |\{\text{edges incident to } p\}| = \Sigma_{q \in \mathfrak{Q}(C)} |\{\text{edges incident to } q\}|$. But we've seen above that the right hand sum $= (i+1)|\mathfrak{Q}(C)| = (i+1) \cdot |C| \cdot c^{(n+1)}_{n+1, n-i, \sigma}$. Also, the left-hand sum $= (n+1)\ \Sigma_{D \in \mathfrak{D}_k(C)} |D| \cdot c^{(n)}_{n, n-i, \tau} a(D, C)$. From Lemma 1, the latter sum $= \Sigma_{D \in \mathfrak{D}_k(C)} |C| \cdot d(C, D) \cdot c^{(n)}_{n, n-i, \tau}$. The proof of the theorem is complete.

An immediate corollary of Theorems 1 and 2 is

COROLLARY 2.1. *Let $C_\sigma$ be any class of odd permutations in $S_{n+1}$. Then $\Sigma_{D_\tau} d(C_\sigma, D_\tau) c^{(n)}_{n, n, \tau} = 2[(n-1)!]$ where the summation is over all classes $D_\tau$ in $S_n$.*

*Example.* Let $C_\sigma$ be the class in $S_{14}$ represented by the partition [5, 3, 3, 2, 1] of $n+1 = 14$. Then $C_\sigma$ has three descendant classes, in $S_{13}$, given by $C_{\tau_1} = [4, 3, 3, 2, 1]$, $C_{\tau_2} = [5, 3, 2, 2, 1]$ and $C_{\tau_3} = [5, 3, 3, 1, 1]$. Since $d(\sigma, \tau_1) = 5 \cdot 1$, $d(\sigma, \tau_2) = 3 \cdot 2$, and $d(\sigma, \tau_3) = 2 \cdot 1$, we obtain the equation $5c^{(13)}_{13, 13, \tau_1} + 6c^{(13)}_{13, 13, \tau_2} + 2c^{(13)}_{13, 13, \tau_3} = 2 \cdot (12!)$.

In the following, $(x)_j$ denotes $x(x-1)\dots(x-j+1)$, for $j \geqq 1$.

LEMMA 2. *Let $C_\tau \subseteq S_n$ be represented by the partition $[\dots 3^{n_3}\ 2^{n_2}\ 1^{n_1}]$, and let $C_{\tau'} \subseteq S_{n+j}$ be represented by $[\dots 3^{n_3}\ 2^{n_2}\ 1^{n_1+j}]$. Then $c^{(n+j)}_{n+j, n+j-i, \tau'} = (n+j-i-1)_j c^{(n)}_{n, n-i, \tau}$, $0 \leqq i \leqq n-1$.*

*Proof.* It suffices to prove the lemma for $j = 1$; the other cases follow easily by repeated applications of the case $j = 1$.

Without loss of generality, assume $\tau' \in S_{n+1}$ (on the symbols $\{1, 2, \dots, n, n+1\}$) and $\tau \in S_n$ (on the symbols $\{1, 2, \dots, n\}$) satisfy $\tau'(k) = \tau(k)$ for each $k$: $1 \leqq k \leqq n$, and $\tau'(n+1) = n+1$. There are $c^{(n)}_{n, n-i, \tau}$ ways to decompose $\tau$ as a product of an $n$-cycle and an $(n-i)$-cycle in $S_n$. For each of these decompositions we will generate $n-i$ different decompositions of $\tau'$ as a product of an $(n+1)$-cycle and an $(n+1-i)$-cycle in $S_{n+1}$. Consider an arbitrary decomposition $\tau = (x_1 \dots x_n)(y_1 \dots y_{n-i})$ in $S_n$. For each $y_k$, $1 \leqq k \leqq n-i$, we obtain

$$\begin{aligned} \tau' &= (x_1 \dots x_n)(n+1\ y_k)(n+1\ y_k)(y_1 \dots y_{n-i}) \\ &= [(x_1 \dots x_n)(n+1\ y_k)](y_1 \dots y_{k-1}\ n+1\ y_k \dots y_{n-i}), \end{aligned}$$

a decomposition of $\tau'$ in $S_{n+1}$, where the bracket contains an $(n+1)$-cycle, since $y_k \in \{x_1, \dots, x_n\} = \{1, 2, \dots, n\}$.

It is easily verified that distinct decompositions of $\tau \in S_n$ generate, in this way, distinct decompositions of $\tau' \in S_{n+1}$, and we have obtained $(n-i)c^{(n)}_{n, n-i, \tau}$ decompositions of $\tau'$. On the other hand, every decomposition of $\tau'$ has been obtained. For, if $\tau' = (x_1 \dots x_{n+1})(y_1 \dots y_{n-i+1})$ where, say, $x_{n+1} = n+1$, then $\tau = \tau' = (x_1 \dots x_n)[(x_n\ x_{n+1})(y_1 \dots y_{n-i+1})]$. Since $\tau'$ fixes $n+1$, the bracket must contain an $(n-i)$-cycle which fixes $x_{n+1}$.

We are now in a position to obtain expressions for $c^{(n)}_{n, n, \tau}$, for a variety of $\tau \in A_n$. We continue to write "$C_\tau$ = [a partition of $n$]", in place of "$C_\tau$ is represented by...".

COROLLARY 2.2.

(a) *If* $C_\tau = [k^m, k-1, 1^l] \subseteq A_n$; $k \geqq 2, l \geqq 0$ *and* $m \geqq 0$, *then* $c^{(n)}_{n,n,\tau} = 2[(n-1)!]/(n-l+1)$.

(b) *If* $C_\tau = [k, 2, 1^l] \subseteq A_n$ (*so* $k$ *is even*), *then* $c^{(n)}_{n,n,\tau} = 2k[(n-1)!]/[(k+1)(k+2)]$.

(c) *If* $C_\tau = [3, 2^m, 1^l] \subseteq A_n$, $l \geqq 0$, *then* $c^{(n)}_{n,n,\tau} = (n-1)!(2m+3)/[2(m+1)(m+3)]$.

(d) *If* $C_\tau = [2^m] \subseteq A_n$ (*so* $n \equiv 0 \mod 4$) *then* $c^{(n)}_{n,n,\tau} = (n-1)!/(m+1) = (n-1)!/(n/2+1)$.

*Proofs.*

(a) Let $C_\sigma = [k^{m+1}, 1^l]$. Then $C_\sigma \subseteq S_{n+1}$ and $C_\sigma$ has only $C_\tau$ as a descendant class. Since $d(C_\sigma, C_\tau) = k(m+1) = n-l+1$, Corollary 2.1 yields $(n-l+1)c^{(n)}_{n,n,\tau} = 2[(n-1)!]$

(b) Let $C_\sigma = [k+1, 2, 1^l] \subseteq S_{n+1}$. Then $C_\sigma$ has only the *two* descendant classes: $C_\tau$ and $C_\rho = [k+1, 1^{l+1}]$. Also, $d(C_\sigma, C_\tau) = k+1$ and $d(C_\sigma, C_\rho) = 2$. From (a), with $m=0$, and $l$ replaced by $l+1$, we have $c^{(n)}_{n,n,\rho} = 2(n-1)!/(n-l)$. Applying Corollary 2.1, we have $d(C_\sigma, C_\tau) \cdot c^{(n)}_{n,n,\tau} + d(C_\sigma, C_\rho) \cdot c^{(n)}_{n,n,\rho} = 2[(n-1)!]$. Substituting, and using $n-l = k+2$, we find the stated expression for $c^{(n)}_{n,n,\tau}$.

(c) If $l \geqq 2$, let $C_\sigma = [3, 2^{m+1}, 1^{l-1}]$. Then $C_\sigma \subseteq S_{n+1}$ and $C_\sigma$ has only the *two* descendant classes: $C_\tau$ and $C_\rho = [2^{m+2}, 1^{l-1}]$. Also $d(C_\sigma, C_\tau) = 2(m+1)$, and $d(C_\sigma, C_\rho) = 3$. From (a) with $k=2$, and $[k^m, k-1, 1^l] = [2^m, 1^{l+1}]$ replaced by $C_\rho = [2^{m+2}, 1^{l-1}] = [2^{m+2}, 1, 1^{l-2}]$, we have $c^{(n)}_{n,n,\rho} = 2[(n-1)!]/[n-(l-2)+1]$. By Corollary 2.1, $d(C_\sigma, C_\tau) c^{(n)}_{n,n,\tau} + d(C_\sigma, C_\rho) c^{(n)}_{n,n,\rho} = 2[(n-1)!]$. Solving for $c^{(n)}_{n,n,\tau}$, we find the desired expression, when we use $n-l = 2m+3$. Thus (c) is true for each $l \geqq 2$. As for $l=1$, let $C_\tau = [3, 2^m, 1] \subseteq A_n$ and $C_{\tau'} = [3, 2^m, 1^2] \subseteq A_{n+1}$. From Lemma 2, with $i=0$ and $j=1$, we have $c^{(n+1)}_{n+1,n+1,\tau'} = n c^{(n)}_{n,n,\tau}$. Also, $c^{(n+1)}_{n+1,n+1,\tau'} = n!(2m+3)/[2(m+1)(m+3)]$, so $c^{(n)}_{n,n,\tau} = (n-1)!(2m+3)/[2(m+1)(m+3)]$. Thus (c) is true for $l=1$; similarly (c) is true for $l=0$ also.

(d) The proof of (d) follows directly from Lemma 2 (with $j=1$ and $i=0$), and (a) above (with $k=2$ and $l=0$). Alternatively, one can use $C_\sigma = [3, 2^{m-1}] \subseteq S_{n+1}$ and its descendant classes, and Corollary 2.1 and (c) above.

Note in particular that if $C_\tau \subseteq A_n$ ($n$ odd) is represented by $[k^m, k-1]$, then

$$c^{(n)}_{n,n,\tau} = \frac{2[(n-1)!]}{(n+1)} = \frac{\prod_{j=1}^{n} j}{\sum_{j=1}^{n} j}.$$

We have thus found a combinatorial interpretation of the latter ratio, for odd $n$. Also, our examples show how we can proceed recursively, beginning with either (b) or (c) of the corollary, to develop expressions for $c^{(n)}_{n,n,\tau}$ when $C_\tau = [k, 2^m, 1^l]$, $k \geqq 3$. We do not pursue this.

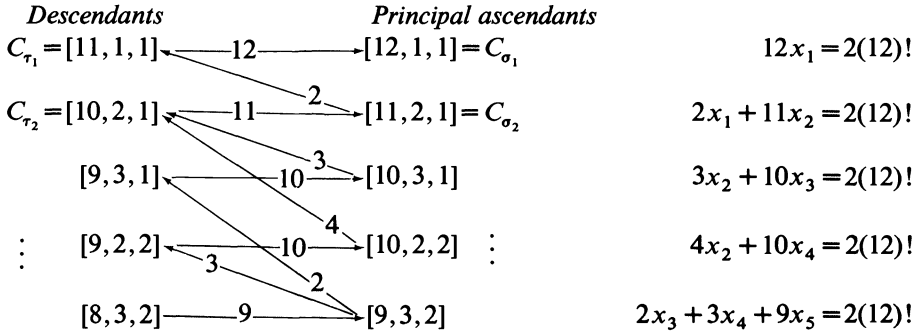
As before, $p(r, n)$ denotes the number of partitions of $n$ into exactly $r$ parts, and there are $p(r, n)$ classes of permutations in $S_n$ with exactly $r$ disjoint cycles. These are even permutations if and only if $r \equiv n \pmod 2$. As $C_\sigma \subseteq S_{n+1}$ runs over all the classes of odd permutations with exactly $r$ disjoint cycles, Corollary 2.1 yields $p(r, n+1)$ linear equations in the $p(r, n)$ unknowns $c^{(n)}_{n,n,\tau}$, where $D_\tau \subseteq S_n$ is represented by a partition of $n$ into exactly $r$ parts. For example, with $n=5$ and $r=3$: Let $C_{\sigma_1} = [4, 1, 1]$, $C_{\sigma_2} = [3, 2, 1]$ and $C_{\sigma_3} = [2, 2, 2]$. $C_{\sigma_1}$ has descendant class $[3, 1, 1] = D_{\tau_1}$; $C_{\sigma_2}$ has descendant classes $[2, 2, 1] = D_{\tau_2}$, and $D_{\tau_1}$; $C_{\sigma_3}$ has descendant class $D_{\tau_2}$. From Corollary 2.1 we have the system: $4 \cdot c^{(5)}_{5,5,\tau_1} = 2 \cdot 4!$, $3 \cdot c^{(5)}_{5,5,\tau_1} + 2 \cdot c^{(5)}_{5,5,\tau_2} = 2 \cdot 4!$, and $6 \cdot c^{(5)}_{5,5,\tau_3} = 2 \cdot 4!$.

We now show that it is possible in general to pick out $p(r,n)$ linearly independent equations in the $p(r,n)$ unknowns, and solve for all $c^{(n)}_{n,n,\tau}$. We say that $C_\sigma(\subseteq S_{n+1})$ is the principal ascendant class of $C_\tau=[\lambda_1,\lambda_2,\dots,\lambda_r]\subseteq S_n$, if $C_\sigma=[\lambda_1+1,\lambda_2,\dots,\lambda_r]$, that is, the length of the *longest* cycle in $\sigma$ exceeds the length of the longest cycle in $\tau$ by exactly one. Furthermore, we may order this collection of $p(r,n)$ partitions of $n$ by putting $[\lambda_1,\lambda_2,\lambda_3,\dots,\lambda_r]>[\lambda'_1,\lambda'_2,\lambda'_3,\dots,\lambda'_r]$ if and only if $\lambda_1=\lambda'_1,\lambda_2=\lambda'_2,\dots,\lambda_{i-1}=\lambda'_{i-1}$, and $\lambda_i>\lambda'_i$. If $C_\tau$ is represented by $[\lambda_1,\lambda_2,\dots,\lambda_r]$, etc., we have clearly induced an ordering on the classes: $C_{\tau_1}>C_{\tau_2}\dots C_{\tau_p}(p=p(r,n))$. If $C_{\sigma_i}$ is the principal ascendant class of $C_{\tau_i}$, $1\leqq i\leqq p$, then we also have $C_{\sigma_1}>C_{\sigma_2}>C_{\sigma_3}>\cdots>C_{\sigma_p}$. Let the column vector $x=(x_1,x_2,\dots,x_p)^t$, $x_i=c^{(n)}_{n,n,\tau}$, $1\leqq i\leqq p$; also let $D=(d_{ij})$ be the $p\times p$ matrix with entries $d_{ij}=d(\sigma_i,\tau_j)$. Then, if $c$ is the column vector of length $p$ with each entry the integer $2[(n-1)!]$, $Dx=c$ follows immediately from Corollary 2.1. Clearly each $d_{ii}>0$. Furthermore $d_{ij}>0, i\neq j$, if and only if $C_{\tau_j}$ is a descendant class of $C_{\sigma_i}$, which implies that $C_{\tau_j}>C_{\tau_i}$ and $j<i$. Thus $D$ is a lower triangular matrix with positive diagonal entries, $D$ is nonsingular, and $Dx=c$ can be solved uniquely for each $c^{(n)}_{n,n,\tau}$.

On the other hand, if a *particular* $c^{(n)}_{n,n,\tau_m}$ is desired, where $C_{\tau_1}>C_{\tau_2}>\cdots>C_{\tau_m}$, it is only necessary to solve a system of *no more than m* equations, obtaining many (but not necessarily all) of the $c^{(n)}_{n,n,\tau_i}$, $1\leqq i\leqq m-1$, along with $c^{(n)}_{n,n,\tau_m}$. The following steps outline the procedure, putting $C_\tau=C_{\tau_m}$:

1) Find the principal ascendant class $C_\sigma(\subseteq S_{n+1})$ of $C_\tau$.
2) List all the descendant classes $(\subseteq S_n)$ of $C_\sigma$.
3) List all the principal ascendant classes of the classes in Step 2.
4) List all the descendant classes of the classes found in Step 3.
5) Continue in this way until reaching $C_{\sigma_1}$, the principal ascendant class of $C_{\tau_1}$.

*Example.* $C_\tau=[8,3,2]\subseteq S_{13}$ and $C_\sigma=[9,3,2]$.




The arrows in the diagram point to the relevant descendant and principal ascendant classes found in the steps above.

Here $d(C_{\sigma_1},C_{\tau_1})=12,\dots,d(C_\sigma,C_\tau)=9$, and the $x_i=c^{(13)}_{13,13,\tau_i}$ are easily found.

COROLLARY 2.3. *Let $C_\sigma\subseteq S_n$ be represented by the partition $[\dots 3^{n_3}\,2^{n_2}\,1^{n_1}]$, and let $|M(\sigma)|=n-n_1$.*

(a) *If $\sigma\in A_n$, $n\geqq 3$, then $c^{(n)}_{n,n-2,\sigma}=|M(\sigma)|(n-3)!$.*

(b) *If $\sigma\in S_n-A_n$, $n\geqq 4$, then $c^{(n)}_{n,n-3,\sigma}=[(n-4)!/3][|M(\sigma)|(|M(\sigma)|-1)-2n_2]$;*

(c) *If $\sigma\in A_n$, $n\geqq 5$, then*

$$c^{(n)}_{n,n-4,\sigma}=\frac{(n-5)!}{4\cdot 3}\left[|M(\sigma)|(|M(\sigma)|-1)(|M(\sigma)|-2)-6n_2(|M(\sigma)|-2)-6n_3\right].$$

*Proofs.*

(a) From Theorem 2 we have, with $i=1$:

$$\sum_{D\subseteq S_{n-1}} d(C_\sigma, D_\tau)c^{(n-1)}_{n-1,n-2,\tau} = 2c^{(n)}_{n,n-2,\sigma}.$$

From Theorem 1 we have $c^{(n-1)}_{n-1,n-2,\tau} = 2[(n-3)!]$ whenever $d(C_\sigma, D_\tau)\neq 0$ (for then $\tau$ is an odd permutation). But

$$\sum_{D_\tau \subseteq S_{n-1}} d(C_\sigma, D_\tau) = \sum_{k\geqq 2} k\cdot n_k = n - n_1 = |M(\sigma)|,$$

and the stated formula follows immediately.

(b) When $i=2$, Theorem 2 yields

$$\sum_{D_\tau \subseteq S_{n-1}} d(C_\sigma, D_\tau)c^{(n-1)}_{n-1,n-3,\tau} = 3c^{(n)}_{n,n-3,\sigma}.$$

From (a), $c^{(n-1)}_{n-1,n-3,\tau} = (n-4)!\,|M(\tau)|$. If $\sigma$ has one more 2-cycle than $\tau$, and $d(C_\sigma, D_\tau) = 2\cdot n_2$, then $|M(\tau)| = |M(\sigma)|-2$; if one more $k$-cycle, and $d(C_\sigma, D_\tau) = k\cdot n_k$, $k\geqq 3$, then $|M(\tau)| = |M(\sigma)|-1$. Thus

$$\begin{aligned} 3c^{(n)}_{n,n-3,\sigma} &= \sum_{k\geqq 3} k\cdot n_k\cdot(n-4)!(|M(\sigma)|-1) + 2\cdot n_2\cdot(n-4)!(|M(\sigma)|-2) \\ &= (n-4)!\left[(|M(\sigma)|-1)\sum_{k\geqq 2} k\cdot n_k - 2n_2\right]. \end{aligned}$$

The expression for $c^{(n)}_{n,n-3,\sigma}$ follows immediately.

(c) When $i=3$, Theorem 2 yields

$$\sum_{D_\tau \subseteq S_{n-1}} d(C_\sigma, D_\tau)c^{(n-1)}_{n-1,n-4,\tau} = 4c^{(n)}_{n,n-4,\sigma}.$$

From (b), $c^{(n-1)}_{n-1,n-4,\tau} = [(n-5)!/3][|M(\tau)|(|M(\tau)|-1) - 2(\#\text{ of 2-cycles of }\tau)]$. When $d(C_\sigma, D_\tau) = 2n_2, |M(\tau)| = |M(\sigma)|-2$ and $\tau$ has $n_2-1$ 2-cycles. When $d(C_\sigma, D_\tau) = 3n_3, |M(\tau)| = |M(\sigma)|-1$ and $\tau$ has $n_2+1$ 2-cycles. When $d(C_\sigma, D_\tau) = k\cdot n_k$, $k\geqq 4, |M(\tau)| = |M(\sigma)|-1$ and $\tau$ has $n_2$ 2-cycles. Thus

$$\begin{aligned} 4c^{(n)}_{n,n-4,\sigma} &= \frac{(n-5)!}{3}\Big\{\big[(|M(\sigma)|-2)(|M(\sigma)|-3) - 2(n_2-1)\big]2n_2 \\ &\qquad + \big[(|M(\sigma)|-1)(|M(\sigma)|-2) - 2(n_2+1)\big]3n_3 \\ &\qquad + \sum_{k\geqq 4}\big[(|M(\sigma)|-1)(|M(\sigma)|-2) - 2n_2\big]kn_k\Big\} \\ &= \frac{(n-5)!}{3}\Big\{\sum_{k\geqq 2}\big[(|M(\sigma)|-1)(|M(\sigma)|-2) - 2n_2\big]kn_k \\ &\qquad - 2n_2(2|M(\sigma)|-6) - 6n_3\Big\} \\ &= \frac{(n-5)!}{3}\{|M(\sigma)|(|M(\sigma)|-1)(|M(\sigma)|-2) - 6n_2(|M(\sigma)|-2) - 6n_3\}, \end{aligned}$$

since $\sum_{k\geqq 2} kn_k = |M(\sigma)|$.

We see in (a), (b), (c) above that, for $i=2, 3, 4$ and $n \geqq i+1$, $c^{(n)}_{n,n-i,\sigma}$ depends only on $n, i$, the parity of $\sigma$, $|M(\sigma)|$, and on the number $n_j$ of $j$-cycles in the disjoint cycle decomposition of $\sigma$ for $j \leqq i-1$ $(i-1=n-(n-i)-1)$. We can now show that this dependence of $c^{(n)}_{n,n-i,\sigma}$ extends to all $i \geqq 5$ (and $n \geqq i+1$). From Theorem 2,

$$\sum_{D_\tau \subseteq S_{n-1}} d(C_\sigma, D_\tau) c^{(n-1)}_{n-1,n-i,\tau} = i c^{(n)}_{n,n-i,\sigma}, \quad \text{for } 1 \leqq i \leqq n-1.$$

Let $i \geqq 5$, $n \geqq i+1$, and suppose $c^{(n-1)}_{n-1,n-i,\tau} (= c^{(n-1)}_{n-1,(n-1)-(i-1),\tau})$ depends only on $n-1$, $i-1$, the parity of $\tau$, $|M(\tau)|$, and the number of $j$-cycles in the disjoint cycle decomposition of $\tau$, for $j \leqq i-2$ $(=(n-1)-(n-i)-1)$. But $d(C_\sigma, D_\tau) \neq 0$ if and only if $D_\tau$ is a descendant class of $C_\sigma$. If $D_\tau$ is a descendant class of $C_\sigma$, then $|M(\tau)|$ equals either $|M(\sigma)|-1$ or $|M(\sigma)|-2$. Furthermore, for each descendant class $D_\tau$ of $C_\sigma = [\ldots j^{n_j} \ldots]$, the number of $j$-cycles in the disjoint cycle decomposition of $\tau$ is one of $\{n_j+1, n_j-1, n_j\}$, depending on whether the cycle length where $\sigma$ and $\tau$ differ is, respectively, a $(j+1)$-cycle, $j$-cycle, or other $k$-cycle (of $\sigma$). Thus $c^{(n)}_{n,n-i,\sigma}$ also depends only on $n, i$, the parity of $\sigma$, $|M(\sigma)|$, and the number of $(j+1)$-cycles of $\sigma$, $j+1 \leqq (i-2)+1$.

**3. Use of the characters of $S_n$.** As mentioned in the introduction, in principle it is possible to compute the integers $c^{(n)}_{\rho\sigma\tau}$, using the character table of $S_n$. For background, including methods for computing the irreducible characters of the symmetric groups, see, e.g., [5] and [6].

Let $[\lambda]=[\lambda_1, \lambda_2, \ldots, \lambda_r]$, $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_r \geqq 1$, $\lambda_1+\lambda_2+\cdots+\lambda_r=n$, denote a partition of the integer $n$. $\chi^\lambda_\rho$ denotes the *character* of the conjugacy class $C_\rho \subseteq S_n$, in the irreducible representation of $S_n$ corresponding to $[\lambda]$. $\chi^\lambda_{1^n}$ denotes the *degree* of the irreducible character $\chi^\lambda$.

Our purpose in this section is to illustrate, using some of the classical results on the ordinary irreducible characters of $S_n$, how one may compute certain of the $c^{(n)}_{\rho\sigma\tau}$ obtained earlier. We wish to contrast the diverse summation problems which arise here with the unity and scope of the results of §2. Our starting points are Theorems A, B, and C below. We then exhibit the variety of sums involving binomial coefficients necessary to evaluate: (i) $c^{(n)}_{n,n-1,\omega}$, where $C_\omega$ is any class of odd permutations in $S_n$ (compare Theorem 1 of §2); (ii) $c^{(n)}_{n,n,\tau}$, where $C_\tau$ is the class of all $n$-cycles in $S_n$, and $n$ is odd (compare Corollary 2.2a of §2 with $l=m=0$ and $k=n+1$); (iii) $c^{(n)}_{n,n,\tau}$, when $n=2^m$, $m$ even, and $C_\tau$ is the class of involutions $[2^m]$ (compare Corollary 2.2d).

THEOREM A (see, e.g., [4, p. 128]).

$$c^{(n)}_{\rho\sigma\tau} = \frac{|C_\rho||C_\sigma|}{n!} \sum_{[\lambda]} \frac{\chi^\lambda_\rho \chi^\lambda_\sigma \chi^\lambda_\tau}{\chi^\lambda_{1^n}},$$

*where the sum is over all partitions* $[\lambda]$ *of* $n$.

THEOREM B (see, e.g., [6, pp. 140–141]).

(1) $\chi^\mu_{\rho'} = \Sigma \chi^\lambda_\rho$, *where* $C_{\rho'} \subseteq S_n$, *and* $\rho'$ *has the same disjoint cycle structure as* $\rho \in S_{n-1}$, *plus one extra 1-cycle. If* $[\mu]=[\mu_1, \mu_2, \ldots, \mu_k]$, *the summation is over all partitions of* $n-1$: $[\lambda]=[\mu_1-1, \mu_2, \ldots, \mu_k]$, $[\mu_1, \mu_2-1, \ldots, \mu_k], \ldots, [\mu_1, \mu_2, \ldots, \mu_k-1]$, *omitting those for which the descending order of the parts is destroyed.*

(2) $\chi^\mu_{\rho'} = \Sigma \pm \chi^\lambda_\rho$, *where* $C_{\rho'} \subseteq S_{n+r}$, *and* $\rho'$ *has the same cycle structure as* $\rho$, *plus one extra r-cycle. If* $[\mu]=[\mu_1, \mu_2, \ldots, \mu_k]$, *the summation is over all partitions* $[\lambda]$ *of* $n$ *such that the sequence* $\lambda_1+k-1, \lambda_2+k-2, \ldots, \lambda_k$ *is obtained from the sequence* $\mu_1+k-$

$1, \mu_2+k-2, \ldots, \mu_k$ *by decreasing one term by r and rearranging in descending order. The minus sign is used if and only if the rearrangement corresponds to a negative permutation.*

Theorems A and B together with the results in [6, Chap. 5] and [5, Chap. 7] using lattice permutations, the Frobenius character formula, and the graph of a partition, make it possible to compute the necessary characters, which we do in Theorem C. In particular, let $[\lambda_i]$ be the partition $[n-i, 1^i]$, $0 \leqq i \leqq n-1$; $[2^m]$ is the *class* of involutions in $S_n$, $n=2^m \geqq 4$. $\chi_j^\lambda$ denotes the character of the class of $j$-cycles ($\subseteq S_n$) in the irreducible representation corresponding to the partition $[\lambda]$ of $n$.

THEOREM C.

(1) $\chi_n^\lambda = \begin{cases} (-1)^i, & \text{if } [\lambda]=[\lambda_i] \\ 0, & \text{otherwise,} \end{cases}$

(2) $\chi_j^{\lambda_i} = \left\{ \binom{n-j-1}{i} + (-1)^{j-1}\binom{n-j-1}{n-i-1}, \right. \qquad 0 \leqq i \leqq n-1, \quad 1 \leqq j \leqq n-1,$

(3) $\chi_{[2^m]}^{\lambda_i} = (-1)^{\lfloor (i+1)/2 \rfloor}\binom{m-1}{\lfloor \frac{i}{2} \rfloor},$

*where* $\lfloor x \rfloor$ *denotes the integer part of* $x$.

*Proof.* The formula in (1) is a direct application of the graphical methods for determining characters, and is referred to in [5, p. 205]. As for (2), for each $n$ the boundary cases $i=0, 1 \leqq j \leqq n-1$, and $j=1, 0 \leqq i \leqq n-1$, follow immediately from the graphical method. For, when $i=0$ we are asserting that $\chi_j^{\lambda_0}=\chi_j^n=\binom{n-j-1}{0}+(-1)^{j-1}\binom{n-j-1}{n-1}=1$, for $1 \leqq j \leqq n-1$. But $\chi_j^n=1$ is a special case of [6, Thm. III, p. 70]. When $j=1$ we are asserting for $1 \leqq i \leqq n-1$, that $\chi_{1^i}^{\lambda_i}=\binom{n-2}{i}+\binom{n-2}{n-i-1}=\binom{n-2}{i}+\binom{n-2}{i-1}=\binom{n-1}{i}$. But this follows immediately from [6, Thm. I, p. 68].

We may now prove formula (2) for each $n$, and $1 \leqq i \leqq n-1, 2 \leqq j \leqq n-1$, by induction. Assume that (2) holds with $n$ replaced by $n-1$, $i$ replaced by $i-1$, and $j$. By Theorem B, part (1), $\chi_j^{\lambda_i}=\chi_\rho^{n-1-i,1^i}+\chi_\rho^{n-i,1^{i-1}}$ where $C_\rho$ is the class ($\subseteq S_{n-1}$) of $j$-cycles. By induction, $\chi_\rho^{n-1-i,1^i}=\binom{n-j-2}{i}+(-1)^{j-1}\binom{n-j-2}{n-i-2}$ and $\chi_\rho^{n-i,1^{i-1}}=\binom{n-j-2}{i-1}+(-1)^{j-1}\binom{n-j-2}{n-i-1}$. Thus $\chi_j^{\lambda_i}=\binom{n-j-1}{i}+(-1)^{j-1}\binom{n-j-1}{n-i-1}$, for each $n$ and $0 \leqq i \leqq n-1, 1 \leqq j \leqq n-1$.

For a proof of (3), we use Theorem B, part (2), with $r=2$. A check of the character tables for $S_4$, and $S_6$ shows (3) to be correct when $m=2$ and $m=3$, and we proceed by induction, moving from $n-2$ to $n$. Let $C_\rho$ be the class of involutions ($\subseteq S_{n-2}$) with cycle structure $[2^{m-1}]$, $n=2m$. Theorem B part 2 yields $\chi_{[2^m]}^\lambda=\chi_\rho^{n-i-2,1^i}-\chi_\rho^{n-i,1^{i-2}}$. But the latter, by induction, gives

$$(-1)^{\lfloor (i+1)/2 \rfloor}\binom{m-2}{\lfloor \frac{i}{2} \rfloor}-(-1)^{\lfloor (i-1)/2 \rfloor}\binom{m-2}{\lfloor \frac{i-2}{2} \rfloor}=(-1)^{\lfloor (i+1)/2 \rfloor}\binom{m-1}{\lfloor \frac{i}{2} \rfloor}$$

We are now ready to evaluate: (i) $c_{n,n-1,\omega}^{(n)}, \omega \notin A_n$; (ii) $c_{n,n,\tau}^{(n)}, C_\tau$ the class of $n$-cycles, $n$ odd; and (iii) $c_{n,n,\tau}^{(n)}, C_\tau$ the class of involutions represented by $[2^m]$, $n=2m$

$\equiv 0 \pmod 4$. We begin by applying Theorem A in each case:

(i)
$$c^{(n)}_{n,n-1,\omega} = \frac{(n-1)!n(n-2)!}{n!}\sum_{[\lambda]}\frac{\chi^\lambda_n\chi^\lambda_{n-1}\chi^\lambda_\omega}{\chi^\lambda_{1^n}}$$
$$= (n-2)!\left(\frac{1\cdot 1\cdot 1}{1} + \frac{(-1)^{n-1}(-1)^{n-2}(-1)}{1}\right)$$
$$= 2[(n-2)!],$$

since, by Theorem C, part (2), $\chi^\lambda_n = 0$, if $[\lambda] \neq [\lambda_i]$; $\chi^{[\lambda_i]}_n = (-1)^i$, $\chi^{[\lambda_i]}_{n-1} = 0$, if $i \neq 0, n-1$;

$$\chi^{\lambda_i}_{n-1} = \begin{cases} (-1)^{n-2}, & i = n-1 \\ 1, & i = 0 \end{cases}.$$

We have also used the fact that $\chi^n_\omega = 1$ for every class $C_\omega$ ([6, Thm. III, p. 70]), and $\chi^{1^n}_\omega = -1$ since $C_\omega$ is a class of odd permutations in $S_n$.

(ii) $c^{(n)}_{n,n,\tau}$, $C_\tau$ the class of $n$-cycles in $S_n$, $n$ odd, is given by

$$c^{(n)}_{n,n,\tau} = \frac{[(n-1)!]^2}{n!}\sum_{[\lambda]}\frac{(\chi^\lambda_n)^3}{\chi^\lambda_{1^n}} = \frac{(n-1)!}{n}\sum_{i=0}^{n-1}\frac{(-1)^i}{\binom{n-1}{i}},$$

using Theorem C, parts 1 and 2. There are several ways to evaluate the latter sum. For example, it appears with hints for a nice suggested proof, as a special case in [9, p. 28, Exercise 3]. The result is, as in Corollary 2.2(a), that $c^{(n)}_{n,n,\tau} = 2[(n-1)!]/(n+1)$.

(iii) $c^{(n)}_{n,n,\tau}$, $C_\tau$ the class $[2^m]$ of involutions in $A_n$ with $m$ 2-cycles, $n = 2m$, $m$ even, is given by

$$c^{(n)}_{n,n,\tau} = \frac{[(n-1)!]^2}{n!}\sum_{[\lambda]}\frac{(\chi^\lambda_n)^2\chi^\lambda_\tau}{\chi^\lambda_{1^n}}$$
$$= \frac{(n-1)!}{n}\sum_{i=0}^{n-1}\frac{\chi^{\lambda_i}_{[2^m]}}{\chi^{\lambda_i}_{1^n}} = \frac{(n-1)!}{n}\sum_{i=0}^{n-1}(-1)^{\lfloor (i+1)/2\rfloor}\binom{m-1}{\lfloor \frac{i}{2}\rfloor}\binom{2m-1}{i}^{-1},$$

again by Theorem C, parts (1) and (3). The latter reduces to $\{2[(n-1)!]/n\}\sum_{i=0}^{m-1}(-1)^i\binom{m-1}{i}\binom{2m-1}{2i}^{-1}$. This sum appears in [4, p. 24, entry (4.25)]. Alternatively, as pointed out by Larry Wallen, one can use the relation between $\binom{2m-1}{2i}^{-1}$ and the beta function, and integrate by parts. In any case we obtain the same result as in Corollary 2.2(d) (along with the value 1 for the sum when $m$ is odd).

## REFERENCES


[1] E. A. BERTRAM, *Even permutations as a product of two conjugate cycles*, J. Combinatorial Theory, 12(1972), pp. 368–380.

[2] G. BOCCARA, *Décompositions d'une permutation d'un ensemble fini en produit de deux cycles*, Discrete Mathematics, 23(1978), pp. 189–205.

[3] D. GORENSTEIN, *Finite Groups*, Harper and Row, New York, 1968.

[4] H. W. GOULD, *Combinatorial Identities*, revised edition, Morgantown, WV, 1972.

[5] M. HAMERMESH, *Group Theory and its Application to Physical Problems*, Addison-Wesley, Reading MA, 1962.

[6] D. E. LITTLEWOOD, *The Theory of Group Characters*, Oxford University Press, Oxford, 1940.
[7] G. MORAN, *The product of two reflection classes of the symmetric group*, Discrete Mathematics, 15(1976), pp. 63–79.
[8] _____ *Some coefficients in the center of the group algebra of the symmetric group*, Discrete Mathematics, 21(1978), pp. 75–81.
[9] J. RIORDAN, *Combinatorial Identities*, John Wiley, New York, 1968.
[10] D. W. WALKUP, *How many ways can a permutation be factored into two n-cycles?*, Discrete Mathematics, 28(1979), pp. 315–319.

SIAM J. ALG. DISC. METH.
Vol. 1, No. 4, December 1980

# ACKNOWLEDGMENT OF PRIORITY: ON THE ORDER OF RANDOM CHANNEL NETWORKS*

A. MEIR,† J. W. MOON† AND J. R. POUNDER†

We have recently learned that some of the results in our paper were anticipated in papers [1] and [2].

## REFERENCES

[1] P. FLAJOLET, J. C. RAOULT AND J. VUILLEMIN, *The number of registers required for evaluating arithmetic expressions,* Theoret. Comput. Sci. 9(1979), pp. 99-125.

[2] R. KEMP, *The average number of registers needed to evaluate a binary tree optimally,* Acta Informatica, 11(1979), pp. 363-372.

* This Journal, 1(1980), pp. 25-33.

† Mathematics Department, University of Alberta, Edmonton, Canada T6G 2G1