

CORRELATION AMONG PARTIAL ORDERS*

P. M. WINKLER†

Abstract. If A is a poset and P is a (finite) poset whose underlying set contains the elements of A , then A is said to *occur* in a linear extension L of P if each relation in A is realized in L ; if L is chosen at random, A can be regarded as an *event* whose probability is the number of linear extensions in which it occurs divided by the total number of linear extensions of P .

We give a complete characterization of the pairs of partial orders which are never negatively correlated, i.e., the pairs A, B with the following property: for any poset P whose underlying set contains the elements of A and of B , $\Pr(A \text{ and } B) \geq \Pr(A) \Pr(B)$.

1. Introduction. In the study of algorithms for sorting, a poset P (henceforth always assumed finite) may be thought of as the set of known relations (at some point in an algorithm) among a set of elements which have an unknown linear order. Thus, if x and y are elements of P , it is natural to define the *probability* of the poset $\{x < y\}$ to be the number of linear extensions of P in which the relation $x < y$ occurs, divided by the total number of linear extensions of P . Thus, if A is a poset whose underlying set is contained in the underlying set of P , then $\Pr(A)$ in P is the probability that all of the relations in A will occur in a random permutation of the elements of P , subject to the constraints in P .

Ivan Rival and Bill Sands [2] conjectured that if x, y and z are three elements of a poset P , then the occurrence of $x < y$ could never diminish the probability that x is below z ; this means that the posets $\{x < y\}$ and $\{x < z\}$ are never negatively correlated, i.e.,

$$\Pr(x < y \text{ and } x < z) \geq \Pr(x < y) \Pr(x < z) \quad \text{in any poset } P$$

and has become known as the *xyz conjecture*. Since a very clever proof was found recently by Shepp [4], it will be referred to here as the *xyz inequality*.

Shepp, in [4], asked the following more general question: for which posets A, B is it always the case that

$$\Pr(A \text{ and } B) \geq \Pr(A) \Pr(B)?$$

Such a pair A, B will be said to be *universally correlated*, denoted by $A \uparrow B$, so that the *xyz inequality* becomes simply $\{x < y\} \uparrow \{x < z\}$. We give below a complete (and easily implemented) characterization of the universally correlated pairs of posets, which shows that all nontrivial cases are ultimately deducible from the *xyz inequality*.

It should be noted that certain other correlations, which are not universal but hold in some important special cases, have been proved by Graham, Yao and Yao [1] and Shepp [3]. Some consequences of the *xyz inequality* can be found in [5].

2. Terminology. We may assume that the posets A and B are defined on a common underlying set S , so that A and B are each subsets of S^2 . The intersection $A \cap B$ will then be a poset, and the transitive closure $(A \cup B)^*$ of the union $A \cup B$ will be a poset unless A and B are inconsistent. If P is a poset whose underlying set contains S , and L is a linear extension of P , then clearly $A \cup B$ occurs in L if and only if A and B both occur in L ; thus $\Pr(A \text{ and } B)$ could be written $\Pr(A \cup B)$. We will do this even though it conflicts with the notion of the union of *events*, for which we reserve the written disjunction "or".

* Received by the editors November 13, 1981, and in final form December 28, 1981.

† Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322.

If P is a poset let $\Delta(P)$ be the set of *covering pairs* in P , i.e., the set of pairs (x, y) of elements of P such that $x < y$ in P but there is no z with $x < z$ and $z < y$ both in P . Note that $(\Delta(P))^* = P$; and if A and B are posets then

$$\Delta((A \cup B)^*) = \Delta(A \cup B) = (\Delta(A) \dot{\cap} \Delta(B)) \cup (\Delta(A \cup B) - \Delta(B)) \cup (\Delta(A \cup B) - \Delta(A))$$

since covering pairs cannot arise from transitive closure.

3. The characterization.

THEOREM. *Let A and B be finite posets. Then A and B are universally correlated if and only if $A \cup B$ is consistent and for every pair $(x, y) \in (\Delta(A \cup B) - \Delta(B))$ and every pair $(u, v) \in (\Delta(A \cup B) - \Delta(A))$, either $x = u$ or $y = v$.*

Several lemmas concerning correlation of events will be useful; in each case the proof requires only elementary probability theory, and is trivial whenever any of the relevant conditional probabilities is undefined.

LEMMA 1. *If C, D and E are events with $\Pr(C \text{ and } D|E) \geq \Pr(C|E) \Pr(D|E)$ and $C \rightarrow E$ and $D \rightarrow E$, then $\Pr(C \text{ and } D) \geq \Pr(C) \Pr(D)$.*

Proof. $\Pr(C \text{ and } D) = \Pr(C \text{ and } D \text{ and } E) = \Pr(C \text{ and } D|E) \Pr(E) \geq \Pr(C \text{ and } D|E) \Pr(E) \Pr(E) \geq \Pr(C|E) \Pr(E) \Pr(D|E) \Pr(E) = \Pr(C \text{ and } E) \Pr(D \text{ and } E) = \Pr(C) \Pr(D)$.

LEMMA 2. *If C, D and E are events with $\Pr(C \text{ and } D) \geq \Pr(C) \Pr(D)$ and $(C \text{ and } D) \rightarrow E \rightarrow D$, then $\Pr(C \text{ and } E) \geq \Pr(C) \Pr(E)$.*

Proof. $\Pr(C \text{ and } E) = \Pr(C \text{ and } D) \geq \Pr(C) \Pr(D) \geq \Pr(C) \Pr(E)$.

LEMMA 3. *If C, D and E are events with $\Pr(C \text{ and } D) \geq \Pr(C) \Pr(D)$ and $\Pr(C \text{ and } E|D) \geq \Pr(C|D) \Pr(E|D)$ then $\Pr(C \text{ and } (D \text{ and } E)) \geq \Pr(C) \Pr(D \text{ and } E)$.*

Proof. $\Pr(C \text{ and } D \text{ and } E) = \Pr(C \text{ and } E|D) \Pr(D) \geq \Pr(C|D) \Pr(E|D) \Pr(D) = \Pr(C \text{ and } D) \Pr(D \text{ and } E) / \Pr(D) \geq \Pr(C) \Pr(D) \Pr(D \text{ and } E) / \Pr(D) = \Pr(C) \Pr(D \text{ and } E)$.

4. Proof of the theorem. We begin by assuming the condition of the theorem holds, with the object of showing that $A \uparrow_P B$. Let S be the common underlying set of A and B , and let P be an arbitrary finite poset whose underlying set contains S ; we wish to show that

$$\Pr(A \cup B) \geq \Pr(A) \Pr(B) \quad \text{in } P,$$

which we abbreviate by $A \uparrow_P B$.

Let $A' = \Delta(A \cup B) - \Delta(B)$ and $B' = \Delta(A \cup B) - \Delta(A)$. The condition of the theorem forces at least one of the following three cases to hold:

Case 1. A' or B' is empty.

Case 2. For some fixed $x \in S$, the pairs in $A' \cup B'$ are all of the form $x < y$ or all of the form $y < x$.

Case 3. One of the sets A', B' contains just one pair $x < v$, and all of the pairs in the other are of form $x < y$ or $u < v$.

Case 4. There exist u, v, x, y such that $A' = \{u < v, x < y\}$ and $B' = \{x < v, u < y\}$.

Cases 2, 3 and 4 are diagrammed in Figs. 1, 2, 3. The diagrams are unique up to duality and exchange of A and B .

Case 1 is easy. If, say, A' is empty, then $\Delta(A \cup B) = \Delta(B)$, thus $B = (A \cup B)^*$ so that $\Pr(A \cup B) = \Pr(B)$ and it follows that $A \uparrow_P B$.

For Cases 2, 3 and 4 let $Q = (P \cup (\Delta(A) \dot{\cap} \Delta(B)))^*$, and suppose we can show that $A' \uparrow_Q B'$. Since $(A'$ and B' and $(\Delta(A) \dot{\cap} \Delta(B)))$ is equivalent to $\Delta(A \cup B)$ and hence to $A \cup B$, we have that $(A' \text{ and } B') \rightarrow A \rightarrow A'$ and $(A' \text{ and } B') \rightarrow B \rightarrow B'$ in Q . Using these

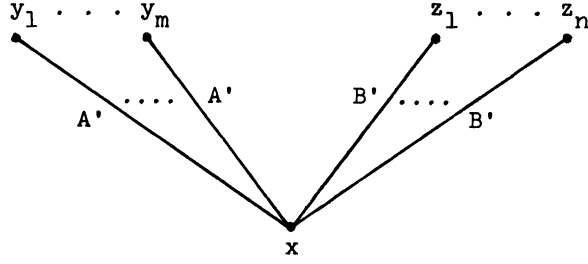


FIG. 1. Case 2.

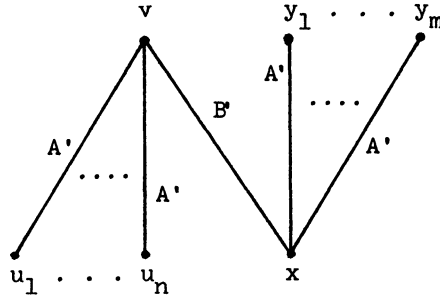


FIG. 2. Case 3.

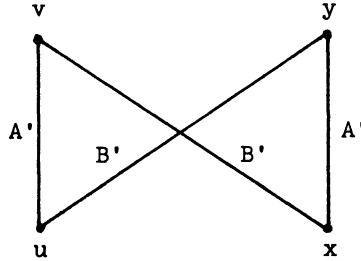


FIG. 3. Case 4.

implications two applications of Lemma 2 yields $A \uparrow_Q B$. Now we observe that in P , $A \rightarrow Q$ and $B \rightarrow Q$; thus $A \uparrow_P B$ as required.

To show that $A' \uparrow_Q B'$ we prove that in fact A' and B' are universally correlated. Note that if C, D and P are arbitrary posets then $C \uparrow_P D$ and $C \uparrow_{P \cup D} E$ together imply $C \uparrow_P (D \cup E)$ by Lemma 3, thus $(C \uparrow_D$ and $C \uparrow_E)$ implies $C \uparrow (D \cup E)$.

In Case 2, we have $\{x < y_i\} \uparrow \{x < z_j\}$ for each i and j , directly from the xyz inequality. One multiple application of the above argument now yields $\{x < y_i\} \uparrow B'$ for each i , and a second multiple application yields $A' \uparrow B'$.

In Case 3, we have $\{x < v\} \uparrow \{x < y_i\}$ and $\{x < v\} \uparrow \{u_j < v\}$ for each i and j via the xyz inequality and its dual, and thus $\{x < v\} \uparrow A'$ but $B' = \{x < v\}$.

In Case 4 again $\{u < v\} \uparrow \{u < y\}$ and $\{u < v\} \uparrow \{x < v\}$ by the xyz inequality, thus $\{u < v\} \uparrow B'$; similarly $\{x < y\} \uparrow B'$ and hence finally $A' \uparrow B'$.

We now assume that A and B are universally correlated, with the intent of proving that the condition of the theorem holds. First, note that $A \cup B$ must indeed be

consistent; otherwise let P be the totally unordered poset on S . Then A and B each have positive probability but $\Pr(A \cup B) = 0$, forcing a negative correlation.

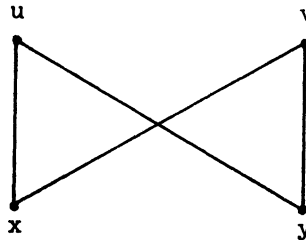
For the condition of the theorem to fail when $A \cup B$ is consistent, at least one of the following two cases would have to obtain:

Case 5. There are distinct elements $x, y, u, v \in S$ such that $x < y$ is in A' and $u < v$ is in B' .

Case 6. There are distinct elements $x, y, z \in S$ such that $x < y$ is in A' and $y < z$ is in B' (or symmetrically, $x < y$ is in B' and $y < z$ is in A').

The proof proceeds in each case by constructing a poset P which provides a counterexample to the presumed universal correlation of A and B . P will be chosen so as to satisfy every relation in $(A \cup B)^*$ except the two specified in each case; thus negative correlation of A and B will be reduced to negative correlation between the two relations.

In Case 5, let $C = (A \cup B)^* - \{x < y, u < v\}$. Since the two excluded relations are both covering relations in $(A \cup B)^*$, C is already closed under transitivity. By duality we may assume that the relation $u < x$ is not in C ; in that case also $v < x$ is not in C . Assume the poset D (see in Fig. 4) on $\{x, y, u, v\}$ is consistent with C .



D

FIG. 4

We partition $S - \{x, y, u, v\}$ as follows:

$$S_1 = \{w \in S : w < x \text{ or } w < y \text{ is in } C\},$$

$$S_2 = \{w \in S : u < w \text{ or } v < w \text{ is in } C\}$$

and

$$S_3 = S - (S_1 \cup S_2 \cup \{x, y, u, v\}).$$

Notice that S_1 and S_2 must be disjoint, otherwise a forbidden relation is implied in C . For $i = 1, 2, 3$ let L_i be a linear ordering of S_i which is consistent with C , and let $z \notin S$. We construct a poset P with underlying set $S \cup \{z\}$ according to the Hasse diagram in Fig. 5. It is easily checked that every relation in C already holds in P , thus in P the event A is equivalent to $\{x < y\}$ and the event B to $\{u < v\}$; therefore we need only show $\{x < y\}$ and $\{u < v\}$ are negatively correlated in P . Letting $k = |L_2|$ and counting the places into which z can fall, we have $\Pr(x < y) = \Pr(u < v) = (2k + 3)/(4k + 8)$ and $\Pr(x < y \text{ and } u < v) = (k + 1)/(4k + 8) = (4k^2 + 2k + 8)/(4k + 8)^2 = ((2k + 3)^2 - 1)/(4k + 8)^2$, thus $\Pr(x < y \text{ and } u < v) < \Pr(x < y) \Pr(u < v)$ as required.

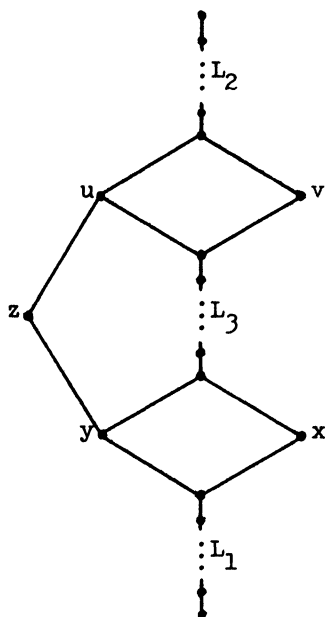


FIG. 5

If C is not consistent with D then C must contain the relation $u < y$. If $v < y$ is also in C then no further relation among x, y, u, v can appear in C , so the dual of the above argument suffices here and also when only $u < y$ appears.

The only other possibility is that the restriction of C to $\{x, y, u, v\}$ is exactly $\{u < y, x < v\}$. In that case we partition $S - \{x, y, u, v\}$ as follows:

$$\begin{aligned} S_1 &= \{w \in S: w < x \text{ or } w < u \text{ in } C\}, \\ S_2 &= \{w \in S: x < w \text{ and } w < v \text{ in } C\}, \\ S_3 &= \{w \in S: u < w \text{ and } w < y \text{ in } C\}, \\ S_4 &= S - (S_1 \cup S_2 \cup S_3 \cup \{x, y, u, v\}). \end{aligned}$$

No element of S is between x and y or between u and v in C , hence there is nothing in C to prevent all of S_4 from lying above $\{x, y, u, v\}$. As before let L_i be a linear ordering of S_i consistent with C , $1 \leq i \leq 4$, and this time P will have underlying set S and the Hasse diagram shown in Fig. 6. Again all relations of C are true in P so it suffices to show $\{x < y\}$ and $\{u < v\}$ are negatively correlated in P . But at least one of these events must occur in P ; hence $\Pr(x < y \text{ and } u < v) = 1 - \Pr(x > y \text{ or } u > v) = 1 - \Pr(x > y) - \Pr(u > v) = \Pr(x < y) + \Pr(u < v) - 1 = \Pr(x < y) \Pr(u < v) - (1 - \Pr(x < y))(1 - \Pr(u < v)) = \Pr(x < y) \Pr(u < v) - \Pr(x > y) \Pr(u > v) < \Pr(x < y) \Pr(u < v)$, and Case 5 is completed.

In Case 6, where $x < y$ is in A' and $y < z$ in B' , the relation $z < x$ cannot be in C nor is there any element of S between x and y or between y and z in C ; else $x < y$ and $y < z$ could not both be covering relations in $(A \cup B)^*$. Therefore in this case we partition $S - \{x, y, z\}$ into $S_1 = \{w | w < x \text{ or } w < y \text{ in } C\}$, $S_2 = \{w | w > z \text{ or } w > y \text{ in } C\}$, and $S_3 = S - (S_1 \cup S_2 \cup \{x, y, z\})$; let L_1, L_2 and L_3 be as before, and take P to have underlying set S and the Hasse diagram shown in Fig. 7.

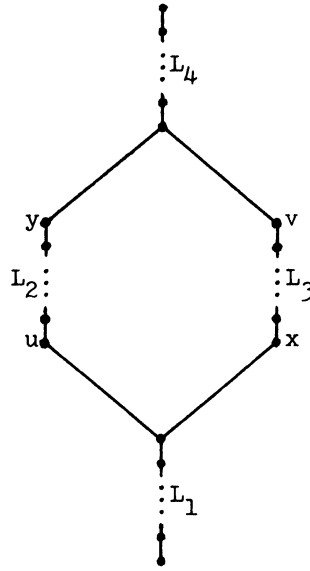


FIG. 6

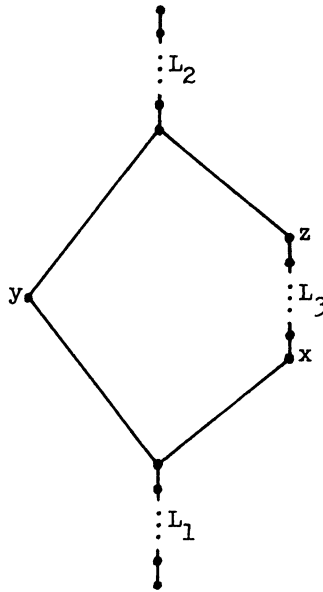


FIG. 7

Here there are only $k + 1$ linear extensions, where $k = |L_3|$; and $\Pr(x < y < z) = (k - 1)/(k + 1) = (k^2 - 1)/(k + 1)^2 < k^2/(k + 1)^2 = \Pr(x < y) \Pr(y < z)$, so the proof of the theorem is complete.

5. Some examples. It follows from the theorem that the following pairs of posets are universally correlated (Cases 1, 2, 3 and 4 respectively):

$$\begin{aligned} A &= \{x < y < z\}, & B &= \{x < z\}; \\ A &= \{x < y, x < z\}, & B &= \{x < u, x < v\}; \\ A &= \{u < v, x < y, x < z\}, & B &= \{x < v\}; \\ A &= \{u < v, x < y\}, & B &= \{x < v, u < y\}. \end{aligned}$$

On the other hand the following reasonable-looking pairs are *not* universally correlated:

$$\begin{aligned} A &= \{x < y\}, & B &= \{x < u < v\}; \\ A &= \{x < y < z, u < w\}, & B &= \{x < z, u < v < w\}; \\ A &= \{x < u, y < u\}, & B &= \{x < v, y < v\}. \end{aligned}$$

In each of the last three cases the proof of the theorem constructs a simple counterexample; in fact we have the following:

COROLLARY. *If A and B are posets on a common underlying set of n elements and A and B are not universally correlated, then there is a poset P having at most $n + 1$ elements on which A and B are negatively correlated.*

This is best possible, since for example the pair

$$A = \{x < u, x < v, y < u, y < v, x < y\}, \quad B = \{x < u, x < v, y < u, y < v, u < v\}$$

is not universally correlated but has no 4-element counterexample.

Acknowledgment. The author wishes to express his gratitude to Ivan Rival and Larry Shepp for bringing this problem to his attention, and to the referee for corrections.

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] *Proceedings of the Symposium on Ordered Sets*, Banff, 1981, D. Reidel, Boston, 1982.
- [3] L. A. SHEPP, *The FKG inequality and some monotonicity properties of partial orders*, this Journal, 1 (1980), pp. 205–299.
- [4] L. A. SHEPP, *The xyz conjecture and the FKG inequality*, Ann. of Probab., 10 (1982), to appear.
- [5] P. B. WINKLER, *Average height in a partially ordered set*, to appear.

ON LEVEL NUMBERS OF t -ARY TREES*

J. W. MOON†

Abstract. The level numbers of a t -ary tree are the distances from the root to the exterior nodes of the tree. Ruskey and Hu [SIAM J. Comput., 6 (1977), pp. 745-758] have considered certain problems involving these level numbers. Alternate derivations and generalizations of some of their results are given here.

1. Introduction. A t -ary tree is a rooted plane tree each node of which is incident with 0 or t edges that lead away from the root, where $t \geq 2$; the nodes of these two types are called *exterior* and *interior* nodes, respectively. If a t -ary tree T_n has n interior nodes then it has $n(t-1)+1$ exterior nodes for $n = 0, 1, \dots$. The j th level number of a t -ary tree T_n is the distance d_j from the root to the j th exterior node of T_n (counting from left to right).

Ruskey and Hu [7] and Ruskey [8], [9] have considered some enumeration problems involving level numbers of binary and t -ary trees that arose in analyzing algorithms for generating and ranking these trees. They derived certain formulae by developing recurrence relations and then appealing to identities for binomial coefficients. Our object here is to give alternate derivations and generalizations of some of their results by means of generating functions. Our arguments lean heavily on the well-known facts that if

$$y = y(x) = \sum_0^{\infty} y_n x^n,$$

where y_n denotes the number of t -ary trees T_n with n interior nodes, then

$$(1) \quad y = 1 + xy^t$$

and

$$(2) \quad y^k = \sum_0^{\infty} \frac{k}{tn+k} \binom{tn+k}{n} x^n$$

for $k = 1, 2, \dots$. Relation (1) follows immediately from the definition of t -ary trees (see [4] or [5, p. 584]), and (2) follows from (1) by induction [2, p. 30] or by applying the Bürmann-Lagrange formula [6, p. 348].

Our main result in § 2 is a formula for the expected position of the first level number that concludes a run of h equal consecutive level numbers, where $1 \leq h \leq t$. In § 3 we determine the asymptotic behaviour of the expected value of the j th level number for fixed j as $n \rightarrow \infty$.

2. Equal consecutive level numbers. If $n \geq 1$ and $1 \leq h \leq t$, let $\gamma_h = \gamma_h(T_n)$ denote the least integer j such that $d_j = d_{j-1} = \dots = d_{j-h+1}$; that γ_h exists follows from the fact that any interior node at maximum distance from the root of T_n is joined to t exterior nodes which perforce determine t equal consecutive level numbers. Let $\mu(n, h)$ denote the expected value of $\gamma_h(T_n)$ over all the y_n t -ary trees T_n . Ruskey [8, p. 439] derived a formula for $\mu(n, t) - t + 1$; this quantity provided an estimate for the average running time of an algorithm he was considering. (Ruskey and Hu [7, p. 758] dealt

* Received by the editors October 23, 1981. The preparation of this paper was assisted by a grant from the Natural Sciences and Engineering Research Council of Canada.

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

with the case $t = 2$ earlier.) We shall derive a formula for $\mu(n, h)$ from a relation for the generating function

$$F_{hk}(x, z) = \sum_{n, j \geq 1} f(n, j; h, k) z^j x^n,$$

where $f(n, j; h, k)$ denotes the number of t -ary trees T_n such that $\gamma_h(T_n) = j$ and $d_j = k$, for positive integers j and k ; let $F_h(x, z)$ and $f(n, j; h)$ denote the corresponding quantities when the value of d_j is not specified.

THEOREM 1.

$$(3) \quad F_{hk}(x, z) = x^k z^h y^{t-h} \{y^{t-1} + zy^{t-2} + \dots + z^{h-1} y^{t-h}\}^{k-1},$$

and

$$(4) \quad F_h(x, z) = \frac{(y-z)(y-1)z^h}{(y-1)z^h - (z-1)y^h}.$$

Proof. Consider the path P from the root to the j th exterior node in any tree T_n such that $\gamma_h(T_n) = j$ and $d_j = k$. There are k interior nodes in P ; these contribute the factor x^k to the right-hand side of (3). Each of these k nodes is joined to $t - 1$ other nodes not in P , and these other nodes are either exterior nodes or the roots of nontrivial subtrees. There can be no nontrivial subtrees nor more than $h - 1$ exterior nodes joined to any of the k interior nodes of P from the left side of P ; for, if there were, it would follow that $\gamma_h(T_n) < j$, contrary to our hypothesis. Hence, each of the first $k - 1$ interior nodes of P is joined to some number i of exterior nodes on the left, where $0 \leq i \leq h - 1$, and to the roots of $t - 1 - i$ (trivial or nontrivial) subtrees on the right. This accounts for the factor $\{y^{t-1} + zy^{t-2} + \dots + z^{h-1} y^{t-h}\}^{k-1}$ in (3). Finally, since $\gamma_h(T_n) = j$, the k th interior node of P must be joined to $h - 1$ exterior nodes on the left of P and to the roots of $t - h$ additional subtrees on the right of P ; this accounts for the factor $z^h y^{t-h}$ in (3). It is not difficult to see that when the factors in the right-hand side of (3) are multiplied out, each tree of the required type contributes one to the coefficient of $z^j x^n$.

Equation (3) may be rewritten as

$$F_{hk}(x, z) = z^h (xy^{t-h})^k \left\{ \frac{y^h - z^h}{y - z} \right\}^{k-1}.$$

When we sum this over $k = 1, 2, \dots$ and appeal to (1) we obtain (4). As a partial check notice that $F_h(x, 1) = y - 1$, the trivial tree T_0 being the only tree not counted.

Before deriving a formula for $\mu(n, h)$ we give some other results that can be deduced as corollaries of Theorem 1; they can also be derived readily from first principles, but separate arguments using generating functions would be somewhat repetitious.

Let $T(n, k, h)$ denote the number of t -ary trees T_n such that $d_1 = \dots = d_h = k$; we assume, as usual, that $n, k \geq 1$ and $1 \leq h \leq t$. The following result was given by Ruskey [8, p. 431].

COROLLARY 1.

$$T(n, k, h) = \frac{tk - k - h + 1}{tn - k - h + 1} \binom{tn - k - h + 1}{n - k}.$$

Proof. Since $T(n, k, h) = f(n, h; h, k)$, it follows that $T(n, k, h)$ is the coefficient of $z^h x^n$ in (3), or the coefficient of x^n in $x^k y^{k(t-1)-h+1}$. This implies the required result upon appealing to (2).

Notice, in particular, that

$$(5) \quad T(n, k, 1) = \frac{tk - k}{tn - k} \binom{tn - k}{n - k}.$$

The case $t = 2$ of this formula was given earlier by Ruskey and Hu [7, p. 752].

Let $R(n, j)$ denote the number of t -ary trees T_n for which j is the least integer such that $d_j = d_{j-1}$.

COROLLARY 2.

$$R(n, j) = \frac{(t-1)(j-1)}{tn-j+1} \binom{tn-j+1}{n-j+1}.$$

Proof. If we appeal to (1) we find that, when $h = 2$, (4) may be rewritten as

$$F_2(x, z) = \frac{(y-1)z^2}{y-z(y-1)} = \frac{z^2xy^{t-1}}{1-zxy^{t-1}} = \sum_{j \geq 2} z^j (xy^{t-1})^{j-1}.$$

Since $R(n, j) = f(n, j; 2)$, it follows that $R(n, j)$ is the coefficient of x^n in $(xy^{t-1})^{j-1}$. This implies the required result upon appealing to (2).

If we compare Corollary 2 with (5), we find that

$$(6) \quad R(n, j) = T(n, j-1, 1).$$

The case $t = 2$ of this relation was given in [7, p. 757].

Let $G(n, j)$ denote the number of t -ary trees T_n such that $d_1 < \dots < d_j$.

COROLLARY 3.

$$G(n, j) = \frac{tj-j+1}{tn-j+1} \binom{tn-j+1}{n-j}.$$

Proof. Since $G(n, j) = R(n, j+1) + R(n, j+2) + \dots$, it follows from the proof of Corollary 2 that $G(n, j)$ is the coefficient of x^n in

$$\sum_{i \geq j} (xy^{t-1})^i = (xy^{t-1})^j (1 - xy^{t-1})^{-1} = y(xy^{t-1})^j,$$

where we have used (1) again. This implies the required result upon appealing to (2).

If we compare Corollaries 1 and 3 we find that

$$(7) \quad G(n, j) = T(n+1, j+1, t-1).$$

The case $t = 2$ of this relation was given in [7, p. 756].

We now derive a formula for $\mu(n, h)$. We adopt the convention that $(x)_0 = 1$ and that $(x)_j = x(x-1) \cdots (x-j+1)$ for $j = 1, 2, \dots$.

THEOREM 2.

$$\mu(n, h) = h \frac{(tn+h-1)_{h-1}}{(tn-n+h)_{h-1}}.$$

Proof. It follows readily from the definitions of $\mu(n, h)$ and $F_h(x, z)$ and Theorem 1 that

$$\sum_1^\infty \mu(n, h) y_n x^n = \left(\frac{\partial}{\partial z} F_h(x, z) \right)_{z=1} = y^h - 1.$$

This implies the required result upon appealing to (2).

Notice, in particular, that

$$\mu(n, t) = \frac{y_{n+1}}{y_n} = t \frac{(tn + t - 1)_{t-1}}{(tn - n + t)_{t-1}} \quad \text{and} \quad \mu(n, 2) = \frac{2tn + 2}{tn - n + 2}.$$

It follows from Theorem 2 that

$$\lim_{n \rightarrow \infty} \mu(n, h) = h\tau^{h-1}$$

for fixed values of h and t , where $\tau = t(t-1)^{-1}$. Furthermore, if $\sigma^2(n, h)$ denotes the variance of $\gamma_n(T_n)$, then it can be shown that

$$\sigma^2(n, t) = \frac{2y_{n+2}}{y_n} - \frac{(2t-1)y_{n+1}}{y_n} - \left(\frac{y_{n+1}}{y_n}\right)^2$$

and that

$$\lim_{n \rightarrow \infty} \sigma^2(n, h) = [h^2 - 2(h-t)^2]\tau^{2h-2} - 2t(h-t)\tau^{h-2} - h(2h-1)\tau^{h-1}.$$

3. The expected value of d_j . Let $e(n, j)$ denote the expected value of the j th level number d_j over all the y_n t -ary trees T_n . It follows from (6) and Theorem 2 that

$$e(n, 1) = \mu(n, 2) - 1 = \frac{nt + n}{nt - n + 2},$$

as was shown by Ruskey and Hu [7, p. 758] when $t = 2$. Ruskey [9] derived a formula for

$$e(j) = \lim_{n \rightarrow \infty} e(n, j)$$

for fixed positive integers j when $t = 2$. Let

$$P_k(x, z) = \sum_{n,j} p_{kjn} y_n z^j x^n,$$

where p_{kjn} denotes the probability that $d_j(T_n) = k$ over all the y_n t -ary trees T_n ; we shall assume until further notice that $t = 2$ so that $y = 1 + xy^2$, whence

$$(8) \quad y = \frac{\{1 - (1 - 4x)^{1/2}\}}{2x}.$$

LEMMA 1.

$$P_k(x, z) = zx^k \{zy(zx) + y(x)\}^k.$$

Proof. We may assume that $k \geq 1$ since the result certainly holds when $k = 0$. Consider the path P from the root to the j th exterior node u in any binary tree T_n such that $d_j(T_n) = k$, for any fixed positive integer j ; there are k interior nodes in this path and these contribute the factor x^k to $P_k(x, z)$. Furthermore, each of these nodes is joined to the root of a subtree lying either to the left or to the right of P . We must take into account the number of exterior nodes in the subtrees that lie to the left of P . Since a binary tree with m interior nodes has $m + 1$ exterior nodes, it follows that the contribution of these k subtrees to $P_k(x, z)$ is $\{zy(zx) + y(x)\}^k$. Finally, the exterior node u itself contributes the factor z to $P_k(x, z)$. It is not difficult to see that when these factors are multiplied out, each binary tree T_n such that $d_j(T_n) = k$ contributes one to the coefficient of $z^j x^n$.

We now give an alternate derivation of the formula for $e(j)$ given in [9]. We let $\mathcal{C}_m\{f\}$ denote the coefficient of x^m in the power series $f(x)$.

THEOREM 3.

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

Proof. When we expand the expression for $P_k(x, z)$ in Lemma 1 and pick off the coefficient of $z^j x^n$, we find that

$$(9) \quad p_{kjn} y_n = \sum_{l=0}^k \binom{k}{l} \mathcal{C}_{j-1}\{(xy)^l\} \cdot \mathcal{C}_{n-j+1}\{(xy)^{k-l}\}.$$

(We remark that this relation is equivalent to [9, Lemma 1].) It is not difficult to see, upon appealing to (2) with $t = 2$, that

$$(10) \quad \lim_{n \rightarrow \infty} \mathcal{C}_{n-j+1} \frac{\{(xy)^{k-l}\}}{y_n} = (k-l) \left(\frac{1}{2}\right)^{k-l+2j-1}$$

for fixed l, k and j , and that

$$(11) \quad \mathcal{C}_{n-j+1} \frac{\{(xy)^{k-l}\}}{y_n} \leq (k-l) \left(\frac{1}{2}\right)^{k-l+2j-4},$$

when $2j \leq n + 1$.

If p_k denotes the limit of p_{kjn} as $n \rightarrow \infty$ for fixed values of k and j , then (9) and (10) imply that

$$(12) \quad \begin{aligned} p_k &= \lim_{n \rightarrow \infty} \sum_{l=0}^k \binom{k}{l} \mathcal{C}_{j-1}\{(xy)^l\} \cdot \frac{\mathcal{C}_{n-j+1}\{(xy)^{k-l}\}}{y_n} \\ &= \sum_{l=0}^k \binom{k}{l} \mathcal{C}_{j-1}\{(xy)^l\} \cdot (k-l) \left(\frac{1}{2}\right)^{k-l+2j-1} \\ &= 4^{-j} \mathcal{C}_{j-1}\left\{k \left(\frac{1}{2} + xy\right)^{k-1}\right\}. \end{aligned}$$

It now follows from Tannery's theorem [1, p. 136] and (10)–(12) and (8) that

$$(13) \quad \begin{aligned} e(j) + 1 &= \sum_{k=0}^{\infty} (k+1) p_k = 4^{-j} \mathcal{C}_{j-1} \left\{ \sum_{k=0}^{\infty} (k+1) k \left(\frac{1}{2} + xy\right)^{k-1} \right\} \\ &= 4^{2-j} \mathcal{C}_{j-1}\{(1-2xy)^{-3}\} = 4^{2-j} \mathcal{C}_{j-1}\{(1-4x)^{-3/2}\} \\ &= 2j \binom{2j}{j} / 4^{j-1}, \end{aligned}$$

as required.

It can be shown that the limiting value $v(j)$ of the variance of d_j , for fixed values of j as $n \rightarrow \infty$, is given by the formula

$$v(j) = 24j - (e(j) + 1)(e(j) + 2)$$

but we shall not pursue this further here.

We also point out that the foregoing argument can be extended to t -ary trees for arbitrary fixed values of t . The details are rather more complicated, but in the general case the factors $zy(zx) + y(x)$ in the formula for $P_k(x, z)$ in Lemma 1 are replaced by

factors

$$(zy(xz^{t-1}))^{t-1} + y(x)(zy(xz^{t-1}))^{t-2} + \dots + y^{t-1}(x)$$

and (13) becomes

$$(14) \quad e(j) + 1 = 2r^j \mathcal{C}_{j-1} \left\{ \frac{r\tau - xy(x^{t-1})}{(r-x)^2} \right\},$$

where $\tau = t(t-1)^{-1}$ and $r = (t-1)t^{-t/(t-1)}$. It follows from (14) that

$$(15) \quad e(j) + 1 = 2\{\tau j - \sum'((j-1) - m(t-1))y_m \rho^m\},$$

where $\rho = r^{t-1}$ and the sum is over all nonnegative integers m such that $m(t-1) < j-1$. (We remark that ρ is the radius of convergence of $y(x)$ and $\tau = y(\rho)$.) It can be shown that

$$e(j) + 1 \sim \frac{(2t)^{1/2}}{t-1} \cdot \frac{j}{4^{j-1}} \binom{2j}{j} \sim \frac{4}{t-1} \left(\frac{2tj}{\pi}\right)^{1/2}$$

for large j and arbitrary fixed values of t .

In conclusion we mention, for purposes of comparison, three related results that hold when $t = 2$. Let δ_n denote the average value of all the level numbers of a binary tree T_n , Δ_n denote the $[\frac{1}{2}(n+2)]$ nd level number of a binary tree T_n , and D_n denote the maximum level number of a binary tree T_n . Finally, let $E(\delta_n)$, $E(\Delta_n)$ and $E(D_n)$ denote the expected values of these parameters over all the y_n binary trees T_n . It follows from a result of Knuth [5, p. 590] that

$$E(\delta_n) \sim (\pi n)^{1/2} = (1.77 \dots)n^{1/2}.$$

It follows from [9, Thm. 1], upon passing to the limit and approximating the resulting sum by an integral, that

$$E(\Delta_n) \sim 4 \left(\frac{n}{\pi}\right)^{1/2} = (2.25 \dots)n^{1/2}.$$

Finally, Flajolet and Odlyzko [3] have shown that

$$E(D_n) \sim 2(\pi n)^{1/2} = (3.54 \dots)n^{1/2}.$$

REFERENCES

[1] T. BROMWICH, *An Introduction to the Theory of Infinite Series*, Macmillan, London, 1931.
 [2] N. R. C. DOCKERAY, *An extension of van der Mond's theorem and some applications*, Math. Gazette, 17 (1933), pp. 26-35.
 [3] P. FLAJOLET AND A. ODLYZKO, *Exploring binary trees and other simple trees*, 21st Annual Symposium on the Foundations of Computer Science, IEEE, Piscataway, NJ, 1980, pp. 207-216.
 [4] D. A. KLARNER, *Correspondences between plane trees and binary sequences*, J. Combin. Theory, 9 (1970), pp. 401-411.
 [5] D. E. KNUTH, *The Art of Computer Programming. Vol. 1: Fundamental Algorithms*, 2nd ed., Addison-Wesley, Reading, MA, 1973.
 [6] G. PÓLYA AND G. SZEGÖ, *Problems and Theorems in Analysis, Vol. I*, Springer, Berlin, 1972.
 [7] F. RUSKEY AND T. C. HU, *Generating binary trees lexicographically*, SIAM J. Comput., 6 (1977), pp. 745-758.
 [8] F. RUSKEY, *Generating t -ary trees lexicographically*, SIAM J. Comput., 7 (1978), pp. 424-439.
 [9] ———, *On the average shape of binary trees*, this Journal, 1 (1980), pp. 43-59.

ON SYMMETRIC REPRESENTATIONS OF FINITE FIELDS*

G. SEROUSSI† AND A. LEMPEL†

Abstract. This paper presents a complete characterization of symmetric representations of finite fields, and representations that are closed under transposition. It is also shown that every finite field has a symmetric representation, and conditions are given under which closure under transposition is equivalent to symmetry.

Key words. representations of finite fields, symmetric representations, trace-dual bases

1. Introduction. Let $F = GF(q)$ be a finite field of q elements, and let $\Phi = GF(q^n)$ be an extension of degree n of F . A set R of square matrices of order n over F is called a *representation* of Φ if, under the operations of matrix addition and matrix multiplication, R forms a field isomorphic to Φ . Characterizations of fields of matrices over finite fields can be found in [1], [2], and [3], where the order of the matrices is not restricted to the degree n of the extension. However, the results of [3] imply that the case where the matrix order is other than n amounts, essentially, to direct-sum constructions of matrices of order n . In this paper, the term "representation" is used only for a field of matrices of order n .

R is called a *symmetric representation* if every element of R is a symmetric matrix; R is *closed under transposition* if for every element B of R , B' , the transpose of B , is also in R . Clearly, every symmetric representation is closed under transposition, but the converse is not always true.

Representations of Φ are closely related to bases of Φ when this field is viewed as a vector space over F . In § 2 this relation is shown to be a many-to-one correspondence under which every basis Ω gives rise to a representation $R(\Omega)$ and every representation corresponds to a subset of bases.

The main result of this paper is a complete characterization of the representations of Φ which are closed under transposition. The characterization utilizes the concept of trace-dual bases. A basis $\Lambda = (\lambda_1 \lambda_2 \cdots \lambda_n)$ is called a *trace-dual* (or a *complementary basis*) of $\Omega = (\omega_1 \omega_2 \cdots \omega_n)$, denoted by $\Lambda = \Omega^*$, if for $1 \leq i, j \leq n$,

$$T(\omega_i \lambda_j) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

where $T: \Phi \rightarrow F$ is the *trace operator* defined by $T(\alpha) = \sum_{i=0}^{n-1} \alpha^{q^i}$ for every $\alpha \in \Phi$. Ω is called a *trace-orthonormal basis* if $\Omega = \Omega^*$. The main properties of trace-dual bases are presented in § 3.

In § 4 we show that a representation $R(\Omega)$ is symmetric if and only if $\Omega^* = \alpha \Omega = (\alpha \omega_1 \alpha \omega_2 \cdots \alpha \omega_n)$ for some $\alpha \in \Phi$. If Φ has a trace-orthonormal basis over F , α must be a quadratic residue, and $\sqrt{\alpha} \Omega$ is such a basis. This result is used to show that every finite extension Φ of a finite field F has a symmetric representation. Note that this is not necessarily true when F is infinite; the field of complex numbers does not have a symmetric representation over the reals.

In § 5 we deal with representations that are closed under transposition. We show that if q is even or both q and n are odd, then a representation of Φ is closed under transposition if and only if it is symmetric; if q is odd and n is even, then $R(\Omega)$ is

* Received by the editors May 6, 1981, and in revised form January 12, 1982.

† Department of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel.

closed under transposition and asymmetric if and only if for some $\alpha \in \Phi$ $\Omega^* = ((\alpha\omega_1)^{q^{n/2}}(\alpha\omega_2)^{q^{n/2}} \cdots (\alpha\omega_n)^{q^{n/2}})$, where $\Omega = (\omega_1\omega_2 \cdots \omega_n)$. We end the section with an example of such a representation of $\Phi = GF(3^2)$ over $F = GF(3)$.

In the sequel we shall make free use of basic results in algebra such as can be found in any standard textbook, e.g., [4].

2. Bases and representations. For every element $\beta \in \Phi$ the mapping $m_\beta: \Phi \rightarrow \Phi$ defined by $m_\beta\alpha = \beta \cdot \alpha$, $\alpha \in \Phi$, is a linear transformation on Φ (viewed as a vector space over F). Let Ω be a basis of Φ over F , and let $M(\beta, \Omega)$ be the matrix of the linear transformation m_β with respect to the basis Ω . It can be readily verified that the set $R(\Omega) = \{M(\beta, \Omega) | \beta \in \Phi\}$ is a representation of Φ . ($R(\Omega)$ is called a *regular representation*.) The following theorem and corollary are direct consequences of the results of [3].

THEOREM 1. *R is a representation of Φ if and only if there exists a basis Ω of Φ over F such that $R = R(\Omega)$.*

COROLLARY 1. *If R is a representation of Φ then for every $a \in F$, the image of a in R is aI , where I is the identity matrix.*

The correspondence between bases and representations of Φ is many-to-one. The following theorem establishes the conditions under which two bases give rise to the same representation of Φ . Given a vector $X = (x_1x_2 \cdots x_n)$, we denote by X^k the vector resulting from raising each component of X to the k th power, namely, $X^k = (x_1^kx_2^k \cdots x_n^k)$.

THEOREM 2. *Let Ω and Λ be bases of Φ over F . Then $R(\Omega) = R(\Lambda)$ if and only if there exist $\alpha \in \Phi - \{0\}$ and an integer $0 \leq k \leq n - 1$ such that $\Lambda = (\alpha\Omega)^{q^k}$.*

The proof of this theorem is relegated to the Appendix so that we can proceed without diversion to our main topic of symmetric representations.

3. Trace-dual bases.

THEOREM 3. *Every basis Ω of Φ over F has a unique trace-dual. Moreover, if W is the matrix over Φ defined by*

$$W = \begin{bmatrix} \Omega \\ \Omega^q \\ \vdots \\ \Omega^{q^{n-1}} \end{bmatrix},$$

then W is nonsingular and its inverse is of the form $[\Lambda'(\Lambda^q)' \cdots (\Lambda^{q^{n-1}})']$ where $\Lambda = \Omega^$.*

A proof of the above theorem can be found in [5, pp. 13–15]. Another proof, valid for fields of characteristic 2 only, can be found in [6, pp. 117–118]. The following lemma summarizes the main properties of trace-duality.

LEMMA 1. *Let $\Omega = (\omega_1\omega_2 \cdots \omega_n)$ be a basis of Φ over F and let $\Omega^* = (\omega_1^*\omega_2^* \cdots \omega_n^*)$.¹ Then*

- (d1) $(\Omega^*)^* = \Omega$.
- (d2) Let $\alpha \in \Phi - \{0\}$. Then $(\alpha\Omega)^* = \alpha^{-1}\Omega^*$.
- (d3) Let k be an integer, $0 \leq k \leq n - 1$. Then $(\Omega^{q^k})^* = (\Omega^*)^{q^k}$.
- (d4) Let $\alpha = \sum_{i=1}^n a_i\omega_i$ and $\beta = \sum_{i=1}^n b_i\omega_i^*$ be arbitrary elements of Φ . Then $T(\alpha\beta) = \sum_{i=1}^n a_ib_i = a'_\Omega\beta\Omega^*$, where x_Λ denotes the column-vector representation of $x \in \Phi$ with respect to a basis Λ .
- (d5) For every $\beta \in \Phi$, $\beta = \sum_{i=1}^n \omega_i^*T(\omega_i\beta)$.
- (d6) Let L be a nonsingular matrix of order n over F . Then $(\Omega L)^* = \Omega^*(L^{-1})'$.

¹ This notation does not imply that ω_i^* depends solely on ω_i .

Proof. (d1) to (d5) follow directly from the definition of trace-duality, and from the basic properties of the trace operator [5, p. 13], [6, p. 116]. To prove (d6), let

$$U = \begin{bmatrix} \Omega L \\ (\Omega L)^q \\ \vdots \\ (\Omega L)^{q^{n-1}} \end{bmatrix}.$$

By Theorem 3, $(\Omega L)^*$ is the first row of $(U^{-1})'$. Since $(\Omega L)^q = \Omega^q L$, it follows that $U = WL$, where W is the same as in Theorem 3. Since Ω^* is the first row of $(W^{-1})'$, and $(U^{-1})' = (L^{-1}W^{-1})' = (W^{-1})'(L^{-1})'$, we have $(\Omega L)^* = \Omega^*(L^{-1})'$. Q.E.D.

In the sequel we shall often refer to a trace-orthonormal basis. The conditions under which such bases exist were derived in [7] and are stated here in the form of the following theorem.

THEOREM 4. $\Phi = GF(q^n)$ has a trace-orthonormal basis over $F = GF(q)$ if and only if either q is even or both q and n are odd.

4. Symmetric representations.

THEOREM 5. (i) Every finite extension Φ of a finite field F has a symmetric representation.

(ii) $R(\Omega)$ is a symmetric representation of Φ over F if and only if there exists an element $\alpha \in \Phi$ such that $\Omega^* = \alpha\Omega$.

(iii) If Φ has a trace-orthonormal basis over F then α and Ω satisfy $\Omega^* = \alpha\Omega$ if and only if $\alpha = \beta^2$ for some $\beta \in \Phi$, and $\beta\Omega$ is trace-orthonormal.

The proof of this theorem is presented in a different order. First, we prove (ii), then (iii) and then (i). We begin with the following lemma.

LEMMA 2. Let $\Omega = (\omega_1\omega_2 \cdots \omega_n)$, $\Omega^* = (\omega_1^*\omega_2^* \cdots \omega_n^*)$, and let B be the image of $\beta \in \Phi$ in $R(\Omega)$. Then

$$B_{ij} = T(\omega_i^*\beta\omega_j), \quad 1 \leq i, j \leq n.$$

Proof. Since $B = M(\beta, \Omega)$, we have

$$(\beta\omega_j)_\Omega = M(\beta, \Omega)(\omega_j)_\Omega = B(\omega_j)_\Omega = BI_j,$$

where I_j , $1 \leq j \leq n$, is the j th column of the identity matrix. By (d4) of Lemma 1, we obtain

$$T(\omega_i^*\beta\omega_j) = T(\beta\omega_j\omega_i^*) = (\beta\omega_j)_\Omega(\omega_i^*)_{\Omega^*} = I_j B' I_i = B_{ij}, \quad 1 \leq i, j \leq n. \quad \text{Q.E.D.}$$

Proof of (ii). Assume $\Omega^* = \alpha\Omega$ for some $\alpha \in \Phi$, and let B be an arbitrary element of $R(\Omega)$. B is the image of some $\beta \in \Phi$ and, by Lemma 2,

$$B_{ij} = T(\omega_i^*\beta\omega_j) = T(\alpha\omega_j\beta\omega_i) = T(\alpha\omega_j\beta\omega_i) = T(\omega_j^*\beta\omega_i) = B_{ji}, \quad 1 \leq i, j \leq n.$$

This implies that $R(\Omega)$ is symmetric.

For the "only if" part, assume that $R(\Omega)$ is symmetric. Then $B_{ij} = B_{ji}$ for every $1 \leq i, j \leq n$ and every $B \in R(\Omega)$. Thus, by Lemma 2,

$$T(\omega_i^*\beta\omega_j) = T(\omega_j^*\beta\omega_i) \quad \text{or} \quad T(\beta(\omega_i^*\omega_j - \omega_j^*\omega_i)) = 0, \quad 1 \leq i, j \leq n, \quad \beta \in \Phi.$$

Since this equality holds for every $\beta \in \Phi$, and since Φ contains elements with a nonzero trace [6, p. 116], we have

$$\omega_i^*\omega_j - \omega_j^*\omega_i = 0 \quad \text{or} \quad \frac{\omega_i^*}{\omega_i} = \frac{\omega_j^*}{\omega_j}, \quad 1 \leq i, j \leq n.$$

Hence, for $\alpha = \omega_1^*/\omega_1$, we have $\Omega^* = \alpha\Omega$. Q.E.D.

Proof of (iii). By Theorem 4 we are dealing with the case where q is even or both q and n are odd. Assume that $\alpha = \beta^2$ and $\beta\Omega$ is trace-orthonormal. Then, $(\beta\Omega)^* = \beta\Omega$, and by (d2) of Lemma 1,

$$\Omega^* = (\beta^{-1}\beta\Omega)^* = \beta(\beta\Omega)^* = \beta \cdot \beta\Omega = \alpha\Omega.$$

Assume now that Ω is a basis of Φ over F such that $\Omega^* = \alpha\Omega$ for some $\alpha \in \Phi$. First, we shall show that α is a quadratic residue. If q is even then for $\beta = \alpha^{q^{n/2}}$ we have $\beta^2 = \alpha^{q^n} = \alpha$. Suppose that both q and n are odd, and consider the matrices

$$W = \begin{bmatrix} \omega_1 & \cdots & \omega_n \\ \omega_1^q & \cdots & \omega_n^q \\ \vdots & & \vdots \\ \omega_1^{q^{n-1}} & \cdots & \omega_n^{q^{n-1}} \end{bmatrix}, \quad X = \begin{bmatrix} \alpha\omega_1 & (\alpha\omega_1)^q & \cdots & (\alpha\omega_1)^{q^{n-1}} \\ \vdots & \vdots & & \vdots \\ \alpha\omega_n & (\alpha\omega_n)^q & \cdots & (\alpha\omega_n)^{q^{n-1}} \end{bmatrix}.$$

Since $\Omega^* = \alpha\Omega$, we have $XW = I$ and therefore $|X||W| = 1$, where $|M|$ denotes the determinant of a matrix M . Observing that $|X| = \alpha^{1+q+\cdots+q^{n-1}}|W|$, we obtain $\alpha^{1+q+\cdots+q^{n-1}} = |X|^2$. Since both q and n are odd, $1+q+\cdots+q^{n-1}$ is odd, and therefore α must be a quadratic residue of Φ . Let $\beta \in \Phi$ be such that $\alpha = \beta^2$. Then, by (d2) of Lemma 1, we obtain

$$(\beta\Omega)^* = \beta^{-1}\Omega^* = \beta^{-1}\alpha\Omega = \beta\Omega.$$

Hence, $\beta\Omega$ is trace-orthonormal. Q.E.D.

We need the following two lemmas for the proof of (i).

LEMMA 3. *Let γ be a primitive element of Φ , let Γ denote the basis $(1 \ \gamma \ \gamma^2 \ \cdots \ \gamma^{n-1})$ of Φ over F , and let S be the symmetric matrix of order n over F defined by $S_{ij} = T(\gamma^{i+j-1})$, $1 \leq i, j \leq n$. Then $\Gamma = \gamma^{-1}\Gamma^*S$.*

Proof. We have for all $1 \leq j \leq n$,

$$(\Gamma^*S)_j = \Gamma^*S I_j = \sum_{i=1}^n \gamma_i^* S_{ij} = \sum_{i=1}^n \gamma_i^* T(\gamma^{i+j-1}) = \sum_{i=1}^n \gamma_i^* T(\gamma^{i-1}\gamma^j) = \gamma^j,$$

where the last equality follows from (d5) of Lemma 1. Hence, $\gamma^{-1}\Gamma^*S = \Gamma$. Q.E.D.

LEMMA 4. *If q is odd and n is even then the matrix S of Lemma 3 can be factored into $S = LL'$, where L is a nonsingular matrix with entries from F .*

Proof. By Lemma 3, S transforms the basis $\gamma^{-1}\Gamma^*$ into the basis Γ , and, therefore, S is nonsingular. By Theorem 2 of [7], S can be factored, over F , into $S = LL'$ if and only if $|S|$ is a quadratic residue of F . Clearly, since S is nonsingular, L must be nonsingular when S is so factorable. Let V be the square matrix of order n with $V_{ij} = \gamma^{(i-1)q^{j-1}}$, $1 \leq i, j \leq n$, and let D be the diagonal matrix with $D_{ii} = \gamma^{q^{i-1}}$, $1 \leq i \leq n$. It can be readily verified that $S = VDV'$ and therefore $|S| = |VV'||D|$. It was proved in [7] that VV' is a matrix over F , and that when q is odd and n is even, its determinant $|VV'|$ is a nonresidue of F . The determinant of the diagonal matrix D is given by $|D| = \prod_{i=1}^n D_{ii} = \gamma^{1+q+\cdots+q^{n-1}} = \gamma^{(q^n-1)/(q-1)}$. Thus, $|D|^{q-1} = 1$ and $|D| \in F$. Since γ is primitive in Φ , $|D|$ is primitive in F and, therefore, it is a nonresidue of F . Finally, since the product of two nonresidues of F is a quadratic residue of F , it follows that $|S| = |VV'||D|$ is a quadratic residue of F . Q.E.D.

Proof of (i). By (ii), it suffices to prove the existence of a basis Ω of Φ over F such that $\Omega^* = \alpha\Omega$ for some $\alpha \in \Phi$.

If q is even or both q and n are odd, Φ has a trace-orthonormal basis Ω which satisfies the required equality with $\alpha = 1$.

For the case where q is odd and n is even, let S, L , and Γ be as defined in Lemmas 3 and 4, and let $\Omega = \Gamma^*L$. Since L is nonsingular Ω is a basis of Φ over F . Also, by (d1) and (d6) of Lemma 1, we have

$$\Omega^* = (\Gamma^*L)^* = \Gamma(L^{-1})' = \Gamma(L')^{-1}.$$

By Lemma 3, we obtain

$$\Omega^* = \Gamma(L')^{-1} = \gamma^{-1}\Gamma^*S(L')^{-1} = \gamma^{-1}\Gamma^*LL'(L')^{-1} = \gamma^{-1}\Gamma^*L = \gamma^{-1}\Omega.$$

Thus, the required equality is satisfied with $\alpha = \gamma^{-1}$. Q.E.D.

5. Representations closed under transposition.

THEOREM 6. *If q is even or both q and n are odd then a representation of Φ is closed under transposition if and only if it is symmetric; if q is odd and n is even then a representation $R(\Omega)$ is closed under transposition but is not symmetric if and only if $\Omega^* = (\alpha\Omega)^{q^{n/2}}$ for some $\alpha \in \Phi$.*

Again, we prove a few lemmas first.

LEMMA 5. *Let Ω be a basis of Φ over F , and let $R'(\Omega) = \{B' \mid B \in R(\Omega)\}$. Then $R'(\Omega) = R(\Omega^*)$.*

Proof. Let B and \bar{B} be the images of $\beta \in \Phi$ in $R(\Omega)$ and $R(\Omega^*)$, respectively. Then, by Lemma 2, and by (d1) of Lemma 1, we have

$$B_{ji} = T(\omega_j^* \beta \omega_i) = T(\omega_i \beta \omega_j^*) = \bar{B}_{ij}, \quad 1 \leq i, j \leq n.$$

Thus $B' = \bar{B}$, and since this holds for every $B \in R(\Omega)$, we have $R'(\Omega) = R(\Omega^*)$. Q.E.D.

LEMMA 6. *$R(\Omega)$ is closed under transposition if and only if $\Omega^* = (\alpha\Omega)^{q^k}$ for some $\alpha \in \Phi$ and some integer $0 \leq k \leq n-1$.*

Proof. This lemma follows directly from Lemma 5 and Theorem 2.

LEMMA 7. *If $\Omega^* = (\alpha\Omega)^{q^k}$ then either $k = 0$, or $k = n/2$ and $\alpha \in GF(q^{n/2})$.*

Proof. Assume $\Omega^* = (\alpha\Omega)^{q^k}$. By (d3) of Lemma 1, we have

$$(\Omega^{q^{n-k}})^* = (\Omega^*)^{q^{n-k}} = (\alpha\Omega)^{q^k \cdot q^{n-k}} = (\alpha\Omega)^{q^n} = \alpha\Omega,$$

or

$$(\alpha\Omega)^* = \Omega^{q^{n-k}}.$$

By (d2) of Lemma 1, $\alpha^{-1}\Omega^* = \Omega^{q^{n-k}}$, or $\Omega^* = \alpha\Omega^{q^{n-k}} = (\alpha^{q^k}\Omega)^{q^{n-k}}$. Thus, we can assume, without loss of generality, that $k \leq n/2$ and, since the trace-dual of Ω is unique (Theorem 3), we must have $(\alpha\Omega)^{q^k} = \alpha\Omega^{q^{n-k}}$. Raising both sides of this equality to the q^k th power we obtain

$$\varphi\Omega^{q^{2k}} = \Omega,$$

where $\varphi = \alpha^{(q^k-1)q^k}$. Let $\beta = \sum_{i=1}^n b_i\omega_i$ be an arbitrary element of Φ . Then

$$\beta = \sum_{i=1}^n b_i\omega_i = \sum_{i=1}^n b_i(\varphi\omega_i^{q^{2k}}) = \varphi \left(\sum_{i=1}^n b_i\omega_i \right)^{q^{2k}} = \varphi\beta^{q^{2k}}.$$

Hence, every element of Φ is a root of the polynomial $\varphi x^{q^{2k}} - x$, which implies that $x^{q^n} - x$ divides $\varphi x^{q^{2k}} - x$. Recalling that $0 \leq k \leq n/2$, this is possible only if $\varphi = 1$ and either $k = 0$ or $k = n/2$. If $k = n/2$ then the equality $\varphi = 1$ implies that $\alpha^{q^{n/2-1}} = 1$ and therefore $\alpha \in GF(q^{n/2})$. Hence, either $k = 0$ or $k = n/2$ and $\alpha \in GF(q^{n/2})$. Q.E.D.

LEMMA 8. *If q is even or both q and n are odd, then $k = 0$.*

Proof. When n is odd, Lemma 7 implies $k = 0$. Therefore, it remains to consider only the case where both q and n are even. Assume, contrary to our claim, that $k \neq 0$. Then, by Lemma 7, we must have $k = n/2$ and $\alpha \in GF(q^{n/2})$. Hence $\Omega^* = (\alpha\Omega)^{q^{n/2}}$, and we have

$$1 = T(\omega_i\omega_i^*) = T(\alpha^{q^{n/2}}\omega_i^{q^{n/2+1}}).$$

Observing that $(\omega_i^{q^{n/2+1}})^{(q^{n/2-1})} = \omega_i^{q^{n-1}} = 1$, we can see that $\omega_i^{q^{n/2+1}} \in GF(q^{n/2})$ which, together with $\alpha \in GF(q^{n/2})$, implies that $\sigma_i = \alpha^{q^{n/2}}\omega_i^{q^{n/2+1}}$ belongs to $GF(q^{n/2})$. Therefore, $\sigma_i^{q^{n/2}} = \sigma_i$, and recalling that we are dealing with a field of characteristic 2, we have

$$T(\sigma_i) = \sum_{j=0}^{n-1} \sigma_i^{q^j} = \sum_{j=0}^{n/2-1} \sigma_i^{q^j} + \sum_{j=0}^{n/2-1} \sigma_i^{q^j} = 0.$$

This contradicts the previous result of $T(\sigma_i) = T(\omega_i\omega_i^*) = 1$, and invalidates the assumption that $k \neq 0$. Q.E.D.

Proof of Theorem 6. By Lemmas 6 and 7, $R(\Omega)$ is closed under transposition if and only if $\Omega^* = (\alpha\Omega)^{q^k}$ for some $\alpha \in \Phi$ and $k = 0$ or $k = n/2$. If q is even, or both q and n are odd then, by Lemma 8, $k = 0$ and, by Theorem 5(ii), $R(\Omega)$ is symmetric. On the other hand, it is obvious that if $R(\Omega)$ is symmetric, $R(\Omega)$ is closed under transposition. This proves the first part of Theorem 6.

If q is odd and n is even, then both values of k are possible ($k = 0$ and $k = n/2$). If $k = 0$, $R(\Omega)$ is symmetric by Theorem 5(ii). If $k = n/2$, $R(\Omega)$ is closed under transposition, and we claim that $R(\Omega)$ cannot be symmetric in this case. If it were symmetric, then, by Theorem 5(ii), there would exist an element $\beta \in \Phi$ such that $\Omega^* = \beta\Omega$ and, hence, $\beta\Omega = (\alpha\Omega)^{q^{n/2}} = \alpha\Omega^{q^{n/2}}$. As in the proof of Lemma 6, this would imply that every element of $\Phi = GF(q^n)$ is a root of $\alpha\beta^{-1}x^{q^{n/2}} - x$, contradicting the fact that this polynomial cannot have more than $q^{n/2}$ roots in Φ . This completes the proof of Theorem 6. Q.E.D.

We conclude this section with an example of a representation of $\Phi = GF(3^2)$ over $F = GF(3)$ which is closed under transposition but is not symmetric.

Let α be a root of the polynomial $x^2 + 1 \in F[x]$. Then $\Phi = F(\alpha)$, and $\Omega = (1 \ \alpha)$ is a basis of Φ over F . The representation $R(\Omega)$ of Φ takes the form

$$\begin{aligned} 0 &\leftrightarrow \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & 1 &\leftrightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & 2 &\leftrightarrow \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, & \alpha &\leftrightarrow \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}, & 2\alpha &\leftrightarrow \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, \\ (1+\alpha) &\leftrightarrow \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}, & (1+2\alpha) &\leftrightarrow \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}, & (2+2\alpha) &\leftrightarrow \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix}, & (2+\alpha) &\leftrightarrow \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix}. \end{aligned}$$

This representation is closed under transposition and, as expected from Theorem 6, $\Omega^* = 2\Omega^3$.

Appendix.

Proof of Theorem 2. Assume first that $\Lambda = (\alpha\Omega)^{q^k}$ for some $\alpha \in \Phi - \{0\}$ and $0 \leq k \leq n-1$. Since $\alpha\Omega$ is also a basis of Φ over F , there exists a nonsingular matrix $V = [V_1 V_2 \cdots V_n]$ with entries from F , such that $\alpha\Omega = \Omega V$, or $(m_\alpha\omega_i)_\Omega = (\alpha\omega_i)_\Omega = V_i$, $1 \leq i \leq n$, where, as before, x_Ω denotes the column-vector representation of $x \in \Phi$ with respect to Ω .

Let β be an arbitrary element of Φ , and let $(b_1 b_2 \cdots b_n)' = \beta_\Omega$. Then, $(m_\alpha\beta)_\Omega = \sum_{i=1}^n b_i (m_\alpha\omega_i)_\Omega = \sum_{i=1}^n b_i V_i = V\beta_\Omega$. It follows that $V = M(\alpha, \Omega)$. Now, for every $\beta \in \Phi$

we have

$$M(\beta, \alpha\Omega) = V^{-1}M(\beta, \Omega)V = M(\beta, \Omega),$$

where the rightmost equality follows from the fact that both V and $M(\beta, \Omega)$ belong to $R(\Omega)$ and the commutativity of $R(\Omega)$. Since this holds for every $\beta \in \Phi$, we have $R(\Omega) = R(\alpha\Omega)$.

To complete the proof of the "if" part, it remains to show that $R(\Lambda) = R(\alpha\Omega)$. Let $\gamma \in \Phi$ and let $\gamma_{\alpha\Omega} = (c_1c_2 \cdots c_n)'$. Then

$$\gamma^{q^k} = \left(\sum_{i=1}^n c_i \alpha \omega_i \right)^{q^k} = \sum_{i=1}^n c_i (\alpha \omega_i)^{q^k} = \sum_{i=1}^n c_i \lambda_i,$$

which implies $(\gamma^{q^k})_{\Lambda} = \gamma_{\alpha\Omega}$ for all $\gamma \in \Phi$. Therefore, for every $\gamma, \beta \in \Phi$, we have

$$M(\beta, \alpha\Omega)\gamma_{\alpha\Omega} = (\beta\gamma)_{\alpha\Omega} = (\beta^{q^k}\gamma^{q^k})_{\Lambda} = M(\beta^{q^k}, \Lambda)(\gamma^{q^k})_{\Lambda} = M(\beta^{q^k}, \Lambda)\gamma_{\alpha\Omega},$$

which implies $M(\beta, \alpha\Omega) = M(\beta^{q^k}, \Lambda)$ which, in turn, implies $M(\beta, \alpha\Omega) \in R(\Lambda)$. Since $R(\alpha\Omega)$ and $R(\Lambda)$ have the same number of elements, we obtain $R(\Lambda) = R(\alpha\Omega) = R(\Omega)$.

To prove the "only if" part, assume $R(\Lambda) = R(\Omega)$, and denote both by R . Let I_l denote the l th column of the unit matrix. For any $l, 1 \leq l \leq n$, we have

$$M(\omega_l, \Omega)I_l = M(\omega_l, \Omega)(\omega_l)_{\Omega} = (\omega_l\omega_l)_{\Omega} = (\omega_l\omega_l)_{\Omega} = M(\omega_l, \Omega)(\omega_l)_{\Omega},$$

and therefore

$$M^{-1}(\omega_l, \Omega)M(\omega_l, \Omega)I_l = M^{-1}(\omega_l, \Omega)M(\omega_l, \Omega)(\omega_l)_{\Omega} = (\omega_l)_{\Omega} = I_l.$$

Similarly, one can show that

$$M^{-1}(\lambda_l, \Lambda)M(\lambda_l, \Lambda)I_l = I_l, \quad 1 \leq l \leq n.$$

Hence, $M^{-1}(\omega_l, \Omega)M(\omega_l, \Omega)$ and $M^{-1}(\lambda_l, \Lambda)M(\lambda_l, \Lambda)$ have identical first columns, and being elements of the same representation R of Φ , they must be equal (or else their difference would be a nonzero matrix in the representation without an inverse). Thus,

$$M^{-1}(\omega_l, \Omega)M(\omega_l, \Omega) = M^{-1}(\lambda_l, \Lambda)M(\lambda_l, \Lambda)$$

or

$$M(\lambda_l, \Lambda) = AM(\omega_l, \Omega), \quad 1 \leq l \leq n,$$

where $A = M(\lambda_1, \Lambda)M^{-1}(\omega_1, \Omega)$. Since $R(\Lambda) = R(\Omega) = R$, A is an element of $R(\Omega)$ representing some $\alpha \in \Phi$, i.e., $A = M(\alpha, \Omega)$. Let ψ_{Λ} and ψ_{Ω} be the isomorphisms from Φ into R such that $\psi_{\Lambda}(\gamma) = M(\gamma, \Lambda)$ and $\psi_{\Omega}(\gamma) = M(\gamma, \Omega)$ for every $\gamma \in \Phi$, and let $\psi = \psi_{\Lambda}^{-1}\psi_{\Omega}$. ψ is an automorphism of Φ that leaves F fixed, since for every $a \in F$ we have $\psi_{\Omega}(a) = \psi_{\Lambda}(a) = aI$, where I is the identity matrix. Hence, ψ is of the form $\psi(x) = x^{q^k}$ for some integer $0 \leq k \leq n-1$. Recalling that $M(\lambda_l, \Lambda) = AM(\omega_l, \Omega) = M(\alpha\omega_l, \Omega)$, we have $\lambda_l = \psi(\alpha\omega_l) = (\alpha\omega_l)^{q^k}$ for $1 \leq l \leq n$ and thus, $\Lambda = (\alpha\Omega)^{q^k}$ for some $\alpha \in \Phi$ and $k, 0 \leq k \leq n-1$. Q.E.D.

Acknowledgment. The authors are grateful to Dr. S. Winograd for many helpful discussions, and to the anonymous referees for their useful suggestions.

REFERENCES

- [1] J. T. B. BEARD, *Matrix fields over prime fields*, Duke Math. J., 39 (1972), pp. 313-321.
- [2] ———, *Matrix fields over finite extensions of prime fields*, Duke Math. J., 39 (1972), pp. 475-484.

- [3] M. WILLETT, *Matrix fields over $GF(q)$* , *Duke Math. J.*, 40 (1973), pp. 701–704.
- [4] I. N. HERSTEIN, *Topics in Algebra*, Blaisdell, New York, 1964.
- [5] I. F. BLAKE AND C. MULLIN, *The Mathematical Theory of Coding*, Academic Press, New York, 1975.
- [6] F. J. MCWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.
- [7] G. SEROUSSI AND A. LEMPEL, *Factorization of symmetric matrices and trace-orthogonal bases in finite fields*, *SIAM J. Comput.*, 9 (1980), pp. 758–767.

PARTITIONS OF Z_2^n *

PETER TANNENBAUM†

Abstract. Let $G = Z_2^n$ ($n > 1$) denote the additive group of the Galois field $GF(2^n)$ and G^* denote the nonzero elements in G . We consider partitions of G^* into disjoint subsets S_1, S_2, \dots, S_i of cardinalities k_1, k_2, \dots, k_i respectively such that the sum of the elements in each of the sets S_i is 0. We prove by constructive methods that such partitions exist if and only if: (1) $k + k_2 + \dots + k_i = 2^n - 1$ and (2) $k_i \geq 3$ for all i .

Applications of these constructions to the construction of single error correcting perfect mixed codes are discussed.

1. Introduction. Let G be a finite abelian group of order m (written additively), with the property that the sum of all of the elements in G is 0. It is easy to verify that either G is of odd order or it has more than one element of order two. Let G^* denote the nonzero elements of G and let S be a subset of G^* such that the sum of the elements of S is 0. Clearly, $|S| \geq 2$ unless S contains an element of order two in G , in which case $|S| \geq 3$. If $\pi: S_1, S_2, \dots, S_i$ is a partition of G^* such that each set S_i has zero sum then we have the following two necessary conditions on $|S_i| = k_i$:

- (1) $k_1 + k_2 + \dots + k_i = m - 1$,
- (2) $k_i \geq 2$ unless S_i contains an element of order two in G , in which case $k_i \geq 3$.

We now make some observations concerning the sufficiency of conditions (1) and (2).

(a) If G is a group of odd order then G has no elements of order two and condition (2) becomes: $k_i \geq 2$ for all i . In this case conditions (1) and (2) are known to be sufficient. This result was proved in [2] for cyclic groups and in [7] for (noncyclic) abelian groups.

(b) If G is a group of even order m with s elements of order two ($s \geq 2$), condition (2) has the following numerical interpretation: $k_i \geq 2$ and if $k_1 = \dots = k_t = 2$ then $t \leq (m - 1 - s)/2$. The following example illustrates the fact that in this case conditions (1) and (2) are not sufficient: Let $G = Z_4 \times Z_2 \times Z_2$, $k_1 = k_2 = k_3 = 2$, $k_4 = k_5 = k_6 = 3$. G^* consists of seven elements of order two in G (say h_1, \dots, h_7) and eight elements not of order two (say $g_1, -g_1, g_2, -g_2, g_3, -g_3, g_4, -g_4$). The desired partition of G^* has to be of the form: $S_1 = \{g_1, -g_1\}$, $S_2 = \{g_2, -g_2\}$, $S_3 = \{g_3, -g_3\}$, $S_4 = \{g_4, h_1, h_2\}$, $S_5 = \{-g_4, h_3, h_4\}$ and $S_6 = \{h_5, h_6, h_7\}$, where $h_1 + h_2 = -g_4$; $h_3 + h_4 = g_4$ and $h_5 + h_6 + h_7 = 0$. Clearly this is impossible.

When G is an abelian group of even order, the problem of finding a necessary and sufficient set of conditions for the existence of a partition of G^* into parts with zero sums is still open.

(c) Let G be the elementary abelian 2-group $G = Z_2 \times \dots \times Z_2$ (n times, $n > 1$). We will write $G = Z_2^n$ and identify the elements of G with binary vectors (words) of length n . Since every element of G^* is of order two, condition (2) becomes: $k_i \geq 3$ for all i . We will prove in § 2 of this paper that in this case conditions (1) and (2) are sufficient by actually constructing the desired partitions.

For a finite abelian group G , the following kind of "partition" of G is of particular interest: $G = G_1 \cup G_2 \cup \dots \cup G_i$, G_i is a subgroup of G , $G_i \cap G_j = \{0\}$. We call such a "partition" a *group partition* of G . Herzog and Schonheim [3], Lindstrom [5] have proved that the existence of a group partition of G is equivalent to the existence of a single error correcting perfect code in the group $G_1 \times \dots \times G_i$. Moreover it is known

* Received by the editors December 18, 1981, and in revised form February 23, 1982.

† Department of Mathematics, University of Arizona, Tucson, Arizona 85721.

[6], [3] that if G has a group partition then G must be an elementary abelian p -group. In § 3 of this paper we construct group partitions of $G = Z_2^n$ using some of the methods developed in § 2. The problem of determining necessary and sufficient conditions for the existence of a group partition in an abelian group G is still open.

2. Partitions of $(Z_2^n)^*$ into parts with zero sums. The main result of this section is:

THEOREM 2.1. *Let $G = Z_2^n$. Then G^* can be partitioned into disjoint subsets S_1, S_2, \dots, S_l of cardinalities k_1, k_2, \dots, k_l respectively and such that the sum of the elements in each S_i is 0 if and only if the following two conditions hold:*

- (1) $k_1 + k_2 + \dots + k_l = 2^n - 1$,
- (2) $k_i \geq 3$ for all i .

Since the necessity of these conditions has been established in § 1, we only need to prove their sufficiency. To do so, we will restrict our attention to the case in which the sets S_i are "small", i.e., $k_i = 3, 4$ or 5 . Once the existence of such zero-sum partitions is established, it is easy to see that any other zero-sum partition can be obtained by judiciously grouping together "small" zero-sum sets as needed to obtain larger ones.

Suppose now that $\pi: S_1, \dots, S_l$ is a partition of $(Z_2^n)^*$ such that $|S_i| = k_i = 3, 4$ or 5 . To each such π we will associate an ordered triple (p, q, t) , where p, q, t are the number of S_i 's of cardinality 5, 4 and 3 respectively, and we will say that π is a partition of type (p, q, t) .

It is clear from the above considerations that Theorem 2.1 is a direct corollary of the following theorem:

THEOREM 2.2. *Let (p, q, t) be any nonnegative integral solution of the equation $5p + 4q + 3t = 2^n - 1$. Then there exists a zero-sum partition of $(Z_2^n)^*$ of type (p, q, t) .*

Before giving the proof of Theorem 2.2, we will need a few preliminary lemmas and remarks.

LEMMA 2.3. *If n is even, then $2^n - 1 \equiv 0 \pmod{3}$ and there exists a zero-sum partition of $(Z_2^n)^*$ of type $(0, 0, (2^n - 1)/3)$.*

Proof. Lemma 2.3 is a special case of [4, Lemma 2]. The following proof is consistent with the other constructions developed in this section. We use induction. The lemma is clearly true for $n = 2$. We assume the lemma is true for $n - 2$. For each zero-sum triple $\bar{T} = \{\bar{a}, \bar{b}, \bar{c}\}$ ($\bar{a} + \bar{b} + \bar{c} = 0$) in Z_2^{n-2} form the following four triples in $Z_2^n = Z_2^2 \times Z_2^{n-2}$:

$$T_1 = \{00\bar{a}, 00\bar{b}, 00\bar{c}\}, \quad T_2 = \{01\bar{a}, 10\bar{b}, 11\bar{c}\},$$

$$T_3 = \{01\bar{b}, 10\bar{c}, 11\bar{a}\}, \quad T_4 = \{01\bar{c}, 10\bar{a}, 11\bar{b}\}.$$

In addition, we form the triple $T_0 = \{01\bar{0}, 10\bar{0}, 11\bar{0}\}$.

Clearly, the above triples are disjoint, have zero sum and partition $(Z_2^n)^*$, as can be seen from Fig. 1.

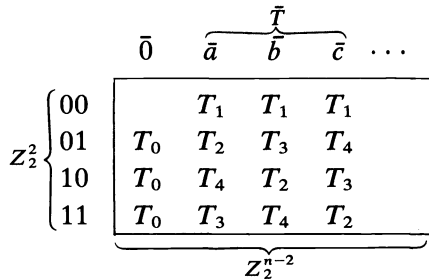


FIG. 1

LEMMA 2.4. *If n is odd ($n \geq 3$) then $2^n - 1 \equiv 1 \pmod{3}$ and there exists a zero-sum partition of $(Z_2^n)^*$ of type $(0, 1, (2^n - 5)/3)$. Moreover, one of the zero-sum triples (say T_0) and the zero-sum quadruple (Q_0) can be chosen so that $T_0 \cup Q_0 \cong (Z_2^2)^*$ (i.e., $T_0 = \{a_0, b_0, c_0 = a_0 + b_0\}$, $Q_0 = \{d_0, a_0 + d_0, b_0 + d_0, c_0 + d_0\}$).*

Proof. For $n = 3$ set $T_0 = \{001, 010, 011\}$, $Q_0 = \{100, 101, 110, 111\}$. Assume the lemma is true for $n - 2$, with the partition of $(Z_2^{n-2})^*$ consisting of $\bar{T}_0 = \{\bar{a}_0, \bar{b}_0, \bar{c}_0\}$, $\bar{Q}_0 = \{\bar{d}_0, \bar{a}_0 + \bar{d}_0, \bar{b}_0 + \bar{d}_0, \bar{c}_0 + \bar{d}_0\}$ and $\bar{T}_i = \{\bar{a}_i, \bar{b}_i, \bar{c}_i\}$ ($i = 1, \dots, (2^{n-2} - 8)/3$). Let $\bar{e}_0 = \bar{a}_0 + \bar{d}_0, \bar{f}_0 = \bar{b}_0 + \bar{d}_0, \bar{g}_0 = \bar{c}_0 + \bar{d}_0$.

We form the following sets in Z_2^n (see Fig. 2):

$$\begin{aligned} T_0 &= \{01\bar{a}_0, 10\bar{b}_0, 11\bar{c}_0\}, & Q_0 &= \{10\bar{0}, 11\bar{a}_0, 00\bar{b}_0, 01\bar{c}_0\}, \\ T_1 &= \{00\bar{c}_0, 00\bar{d}_0, 00\bar{g}_0\}, & T_2 &= \{01\bar{b}_0, 10\bar{g}_0, 11\bar{e}_0\}, \\ T_3 &= \{01\bar{g}_0, 10\bar{a}_0, 11\bar{f}_0\}, & T_4 &= \{00\bar{f}_0, 10\bar{c}_0, 10\bar{e}_0\}, \\ T_5 &= \{00\bar{e}_0, 11\bar{b}_0, 11\bar{g}_0\}, & T_6 &= \{00\bar{a}_0, 01\bar{d}_0, 01\bar{e}_0\}, \\ T_7 &= \{11\bar{0}, 01\bar{f}_0, 10\bar{f}_0\}, & T_8 &= \{01\bar{0}, 10\bar{d}_0, 11\bar{d}_0\}. \end{aligned}$$

		\bar{T}_0			\bar{Q}_0				\bar{T}_i			
		\bar{a}_0	\bar{b}_0	\bar{c}_0	\bar{d}_0	\bar{e}_0	\bar{f}_0	\bar{g}_0	\dots	\bar{a}_i	\bar{b}_i	\bar{c}_i
Z_2^2	00		T_6	Q_0	T_1	T_1	T_5	T_4	T_1	as in Lemma 2.3		
	01	T_8	T_0	T_2	Q_0	T_6	T_6	T_7	T_3			
	10	Q_0	T_3	T_0	T_4	T_8	T_4	T_7	T_2			
	11	T_7	Q_0	T_5	T_0	T_8	T_2	T_3	T_5			

Z_2^{n-2}

FIG. 2

A straightforward check shows that these sets have zero sums and partition the nonzero elements of $Z_2^2 \times (\bar{T}_0 \cup \bar{Q}_0 \cup \{0\})$. The remaining elements of $(Z_2^n)^*$ can be partitioned into zero-sum triples using the construction of Lemma 2.3.

In addition to Lemmas 2.3, 2.4, the following two remarks are essential.

Remark 1. A key construction we will exploit in the proof of Theorem 2.2 is based on the fact that under certain “favorable” circumstances, it is very easy to change a zero-sum partition π of $(Z_2^n)^*$ of type (p, q, t) into a zero-sum partition $\hat{\pi}$ of $(Z_2^n)^*$ of type $(p + 2, q - 1, t - 2)$. What are these “favorable” circumstances? Suppose that π contains a pair of triples $T_1 = \{x_1, x_2, x_3\}$, $T_2 = \{y_1, y_2, y_3\}$ and a quadruple $Q = \{z_1, z_2, z_3, z_4\}$ which satisfy the relation:

$$(*) \quad x_1 + y_1 = z_1 + z_2.$$

In this case we can rearrange the elements in $T_1 \cup T_2 \cup Q$ to form the sets $P_1 = \{x_1, y_2, y_3, z_1, z_2\}$ and $P_2 = \{x_2, x_3, y_1, z_3, z_4\}$. Since $y_2 + y_3 = y_1$ and $x_2 + x_3 = x_1$, it follows from (*) that P_1 and P_2 have zero sum.

When the sets T_1, T_2 and Q satisfy the relation (*) we will say that they form a *pivotal configuration* in π and will call the operation $\pi \rightarrow \hat{\pi}$ ($\{T_1, T_2, Q\} \rightarrow \{P_1, P_2\}$) a *pivot*.

Essential to the proof of Theorem 2.2 will be the fact that we will be able to construct zero-sum partitions having a large number of disjoint pivotal configurations.

Remark 2. Let σ_n be the set of all nonnegative integral solutions (p, q, t) of the equation $5p + 4q + 3t = 2^n - 1$. Define the following equivalence relation on σ_n : $(p_1, q_1, t_1) \equiv (p_2, q_2, t_2)$ if and only if $p_2 - p_1 = 2(q_1 - q_2) = t_1 - t_2$. This equivalence relation induces a partition of σ_n into equivalence classes. A typical equivalence class E can be listed in increasing lexicographic order as follows:

$$E = \{(p_0, q_0, t_0), (p_0 + 2, q_0 - 1, t_0 - 2), \dots, (p_0 + 2\lambda, q_0 - \lambda, t_0 - 2\lambda)\},$$

where $\lambda = \min\{q_0, \lfloor t_0/2 \rfloor\}$ ($\lfloor x \rfloor$ denotes integer part of x) and $p_0 = 0$ or 1 . The smallest (in the lexicographic order) element of the equivalence class will be called the *class leader*.

Example. The equivalence classes in σ_6 (with class leaders underlined) are:

$$E_1 = \{(\underline{0}, 0, 21)\},$$

$$E_2 = \{(\underline{0}, 3, 17), (2, 2, 15), (4, 1, 13), (6, 0, 11)\},$$

$$E_3 = \{(\underline{0}, 6, 13), (2, 5, 11), (4, 4, 9), (6, 3, 7), (8, 2, 5), (10, 1, 3), (12, 0, 1)\},$$

$$E_4 = \{(\underline{0}, 9, 9), (2, 8, 7), (4, 7, 5), (6, 6, 3), (8, 5, 1)\},$$

$$E_5 = \{(\underline{0}, 12, 5), (2, 11, 3), (4, 10, 1)\},$$

$$E_6 = \{(\underline{1}, 1, 18), (3, 0, 16)\},$$

$$E_7 = \{(\underline{1}, 4, 14), (3, 3, 12), (5, 2, 10), (7, 1, 8), (9, 0, 6)\},$$

$$E_8 = \{(\underline{1}, 7, 10), (3, 6, 8), (5, 5, 6), (7, 4, 4), (9, 3, 2), (11, 2, 0)\},$$

$$E_9 = \{(\underline{1}, 10, 6), (3, 9, 4), (5, 8, 2), (7, 7, 0)\},$$

$$E_{10} = \{(\underline{1}, 13, 2), (3, 12, 0)\}.$$

We are now ready to proceed with the proof of Theorem 2.2.

Proof of Theorem 2.2. Let (p_0, q_0, t_0) be the class leader in an equivalence class E of σ_n . We will construct a zero-sum partition π of $(Z_2^n)^*$ of type (p_0, q_0, t_0) and having $\lambda = \min\{q_0, \lfloor t_0/2 \rfloor\}$ disjoint pivotal configurations. By Remark 1, performing λ successive pivots will yield zero-sum partitions corresponding to each of the remaining λ types in E .

Since (p_0, q_0, t_0) is a class leader, we have $p_0 = 0$ or 1 (if $p_0 \geq 2$, $(p_0 - 2, q_0 + 1, t_0 + 2)$ would precede (p_0, q_0, t_0) in the equivalence class). We will consider four cases.

Case 1A. $p_0 = 0, n$ even. Here $4q_0 + 3t_0 = 2^n - 1$. Since n is even, $2^n - 1 \equiv 0 \pmod{3}$ and $2^n - 1 \equiv 3 \pmod{4}$. This implies $q_0 \equiv 0 \pmod{3}$ and $t_0 \equiv 1 \pmod{4}$. Let $q_0 = 3r$ and $t_0 = 4s + 1$. Then $2^n - 1 = 12r + 12s + 3$ and $2^{n-2} - 1 = 3(r + s)$. We now use Lemma 2.3 to obtain a zero-sum partition of $(Z_2^{n-2})^*$ into $s + r$ zero-sum triples $\bar{T}_1, \dots, \bar{T}_s, \bar{T}_{s+1}, \dots, \bar{T}_{s+r}$.

For each \bar{T}_i ($i = 1, \dots, s$) we construct four zero-sum triples $T_{i1}, T_{i2}, T_{i3}, T_{i4}$ in $(Z_2^n)^*$ using the same construction as in Lemma 2.3 (see Fig. 3). For every other element $\bar{\alpha}_j$ in $\cup_{i=s+1}^{s+r} \bar{T}_i$ we form the zero-sum quadruple $Q_j = \{00\bar{\alpha}_j, 01\bar{\alpha}_j, 10\bar{\alpha}_j, 11\bar{\alpha}_j\}$. It is easy to see that any two triples T_{ik}, T_{il} ($k, l = 1, 2, 3, 4, k \neq l$) can be combined with any quadruple Q_j to form a pivotal configuration and that this procedure can be repeated as many as $\min\{3r, 2s\} = \min\{q_0, \lfloor t_0/2 \rfloor\} = \lambda$ times.

It is worth noting that the triple $T_0 = \{01\bar{0}, 10\bar{0}, 11\bar{0}\}$ has yet to be accounted for in the construction, but since $t_0 = 4s + 1$ is odd, T_0 will never need to appear as part of a pivotal configuration.

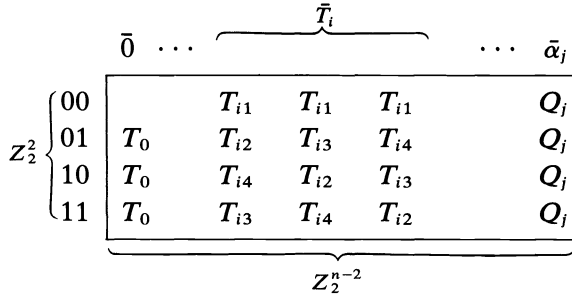


FIG. 3

Case 1B. $p_0 = 0, n$ odd, $n > 3$. (If $n = 3$ then the theorem holds vacuously.) Again, $4q_0 + 3t_0 = 2^n - 1$. Since n is odd, $2^n - 1 \equiv 1 \pmod{3}$ and $2^n - 1 \equiv 3 \pmod{4}$. This implies $q_0 \equiv 1 \pmod{3}$ and $t_0 \equiv 1 \pmod{4}$. Let $q_0 = 3r + 1$ and $t_0 = 4s + 1$. We consider two subcases.

(i) $q_0 = 1$. We use the partition of $(Z_2^n)^*$ constructed in Lemma 2.4 and observe that T_1, T_2 and Q_0 form a pivotal configuration (choose $x_1 = 00\bar{c}_0, y_1 = 01\bar{b}_0, z_1 = 00\bar{b}_0, z_2 = 01\bar{c}_0$ to satisfy (*)).

(ii) $q_0 \geq 4$. Using Lemma 2.4, partition $(Z_2^{n-2})^*$ into one quadruple \bar{Q}_0 and triples \bar{T}_i . We now use an identical construction as in Case 1A, choosing s of the triples \bar{T}_i to obtain $4s$ triples $T_{i1}, T_{i2}, T_{i3}, T_{i4}$ in $(Z_2^n)^*$ and the remaining $3r + 1 \geq 4$ elements $\bar{\alpha}_j$ of $(Z_2^n)^*$ to obtain quadruples $Q_j = \{00\bar{\alpha}_j, 01\bar{\alpha}_j, 10\bar{\alpha}_j, 11\bar{\alpha}_j\}$. As in Case 1A, this construction yields as many as $\min\{3r + 1, 2s\} = \min\{q_0, \lfloor t_0/2 \rfloor\} = \lambda$ pivotal configurations.

Case 2A. $p_0 = 1, n$ even. Here $4q_0 + 3t_0 = 2^n - 6$. Since n is even, $2^n - 6 \equiv 1 \pmod{3}$ and $2^n - 6 \equiv 2 \pmod{4}$. This implies $q_0 \equiv 1 \pmod{3}$ and $t_0 \equiv 2 \pmod{4}$. Let $q_0 = 3r + 1$ and $t_0 = 4s + 2$. Using Lemma 2.3 we partition $(Z_2^{n-2})^*$ into triples \bar{T}_i ($i = 0, 1, \dots, s + r$). We choose one of these triples (say $\bar{T}_0 = \{\bar{a}, \bar{b}, \bar{c}\}$) and form the following zero-sum sets in $(Z_2^n)^*$ (see Fig. 4):

$$T_{01} = \{10\bar{a}, 01\bar{a}, 11\bar{0}\}, \quad T_{02} = \{10\bar{c}, 01\bar{b}, 11\bar{a}\},$$

$$Q_0 = \{00\bar{c}, 01\bar{c}, 10\bar{b}, 11\bar{b}\}, \quad P_0 = \{00\bar{a}, 00\bar{b}, 01\bar{0}, 10\bar{0}, 11\bar{c}\}.$$

T_{01}, T_{02} and Q_0 form a pivotal configuration ($10\bar{a} + 11\bar{a} = 10\bar{b} + 11\bar{b}$). The remaining $q_0 - 1 = 3r$ quadruples and $t_0 - 2 = 4s$ triples are obtained as in Case 1A.

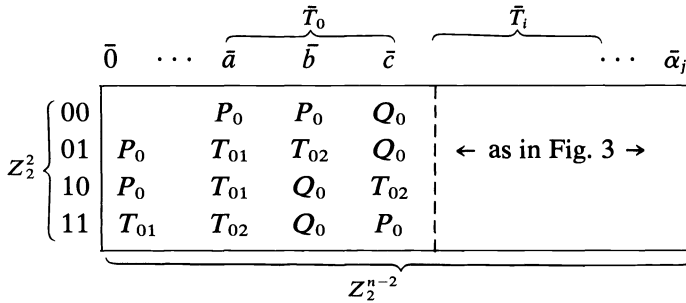


FIG. 4

Case 2B. $p_0 = 1, n$ odd, $n > 3$. Here $4q_0 + 3t_0 = 2^n - 6$. Since n is odd, $2^n - 6 \equiv 2 \pmod{3}$ and $2^n - 6 \equiv 2 \pmod{4}$. This implies $q_0 \equiv 2 \pmod{3}$ and $t_0 \equiv 2 \pmod{4}$. Let $q_0 = 3r + 2$ and $t_0 = 4s + 2$. Using Lemma 2.4, partition $(Z_2^{n-2})^*$ into \bar{Q}_0 and \bar{T}_i

($i = 1, \dots, r+s$). Suppose that $\bar{Q}_0 = \{\bar{d}, \bar{e}, \bar{f}, \bar{g}\}$, ($\bar{d} + \bar{e} + \bar{f} + \bar{g} = 0$). We form the following zero sum sets in $(Z_2^n)^*$ (see Fig. 5):

$$\begin{aligned} Q_{01} &= \{00\bar{d}, 01\bar{d}, 10\bar{d}, 11\bar{d}\}, & Q_{02} &= \{00\bar{e}, 01\bar{e}, 10\bar{e}, 11\bar{e}\}, \\ T_{01} &= \{01\bar{f}, 10\bar{0}, 11\bar{f}\}, & T_{02} &= \{01\bar{g}, 10\bar{g}, 11\bar{0}\}, \\ P_0 &= \{00\bar{f}, 00\bar{g}, 01\bar{0}, 10\bar{f}, 11\bar{g}\}. \end{aligned}$$

	$\bar{0}$	\bar{Q}_0				\dots	\bar{T}_i		\dots	α_j
Z_2^n	<div style="display: flex; flex-direction: column; align-items: center;"> { <div style="display: flex; flex-direction: column; align-items: center;"> 00 01 10 11 </div> } </div>	<div style="display: flex; flex-direction: column; align-items: center;"> Q_{01} P_0 T_{01} T_{02} </div>	<div style="display: flex; flex-direction: column; align-items: center;"> Q_{02} Q_{01} Q_{01} Q_{01} </div>	<div style="display: flex; flex-direction: column; align-items: center;"> P_0 Q_{02} Q_{02} Q_{02} </div>	<div style="display: flex; flex-direction: column; align-items: center;"> P_0 T_{01} P_0 T_{01} </div>	<div style="display: flex; flex-direction: column; align-items: center;"> T_{02} T_{02} T_{02} P_0 </div>	← as in Fig. 3 →			
	Z_2^{n-2}									

FIG. 5

T_{01} , T_{02} and Q_{01} form a pivotal configuration ($10\bar{0} + 11\bar{0} = 00\bar{d} + 01\bar{d}$), and Q_{02} can form a pivotal configuration with any pair of triples T_{ik} , T_{il} ($i = 1, \dots, s; k, l = 1, 2, 3, 4, k \neq l$).

This completes the proof of Theorem 2.1.

3. Partitions of Z_2^n and single error correcting perfect codes. Let G_1, \dots, G_l be finite abelian groups and let W be their direct product $W = G_1 \times \dots \times G_l$. A *mixed group code* C is simply a subgroup of W . (If the G_i 's ($i = 1, \dots, l$) are known to be isomorphic then C is usually referred to as a *group code*.) In the remainder of the section, C will always denote a mixed group code in $W = G_1 \times \dots \times G_l$. If there exists a positive integer e such that W is the disjoint union of all spheres with centers in C and radii e , then C is called an *e-error correcting perfect code*. (Here the distance function is the traditional Hamming distance $d(\bar{x}, \bar{y}) = |\{i | x_i \neq y_i\}|$ where $\bar{x} = (x_1, \dots, x_l)$, $\bar{y} = (y_1, \dots, y_l)$.)

We now restrict our attention to single error correcting perfect codes which we will refer to as SECP codes. The following theorem establishes the connection between SECP codes and partitions of a group.

THEOREM 3.1. *Let G_1, G_2, \dots, G_l be finite abelian groups, $W = G_1 \times \dots \times G_l$. There exists an SECP code in W if and only if there exist abelian groups $G; S_1, S_2, \dots, S_l$ satisfying the following conditions:*

- (i) $G = S_1 \cup S_2 \dots \cup S_l$,
- (ii) $S_i \cap S_j = \{0\}$,
- (iii) G_i isomorphic to S_i .

A nontrivial decomposition of G satisfying conditions (i), (ii), (iii) is called a *group partition* of G of type $\{G_1, G_2, \dots, G_l\}$.

The "if" part of Theorem 3.1 was proved by Herzog and Schonheim [3], the "only if" part by Lindstrom [5].

The following result was originally proved by G. A. Miller [6] and subsequently by J. W. Young [8] and Herzog and Schonheim [4]:

THEOREM 3.2. *If a finite abelian group G has a group partition $G = S_1 \cup S_2 \cup \cdots \cup S_l$, then $G \cong Z_p^n$, $S_i \cong Z_p^{m_i}$ ($i = 1, \dots, l$) for some prime p . (We continue using the notation $Z_p^n = Z_p \times \cdots \times Z_p$.)*

It is easy to see that the following two conditions are necessary for the existence of a group partition of Z_p^n of type $Z_p^{m_1}, \dots, Z_p^{m_l}$:

- (A) $p^n - 1 = \sum_{i=1}^l (p^{m_i} - 1)$,
 (B) $m_i + m_j \leq n$, ($i \neq j, i, j = 1, \dots, l$).

Condition (A) follows from a simple counting argument, condition (B) from the fact that since $S_i \cong Z_p^{m_i}$, $S_j \cong Z_p^{m_j}$ and $S_i \cap S_j = \{0\}$ the subgroup generated by S_i and S_j is of order $p^{m_i+m_j}$.

At the end of this section we give an example to show that conditions (A) and (B) are not sufficient. The problem of finding necessary and sufficient conditions for the existence of a group partition of Z_p^n is open.

There is a natural connection between group partitions of G and zero-sum partitions of G^* . If $G = S_1 \cup \cdots \cup S_l$ is a group partition of G then, by Theorem 3.2, $G \cong Z_p^n$, $S_i \cong Z_p^{m_i}$ and therefore S_i^* is a zero-sum set unless $S_i \cong Z_2$. In this latter case $S_i = \{0, a\}$ for some $a \in G^*$ and we must have that for all such S_i 's the union of the corresponding S_i^* 's is itself a zero-sum set.

We will now restrict our attention to the binary case ($p = 2$) and consider group partitions of Z_2^n of type $Z_2^{m_1}, \dots, Z_2^{m_l}$ such that $m_i > 1$ ($i = 1, \dots, l$). By the previous observation, if $Z_2^n = S_1 \cup \cdots \cup S_l$ is one such partition then $(Z_2^n)^* = S_1^* \cup \cdots \cup S_l^*$ is a partition of $(Z_2^n)^*$ into zero-sum parts. The following two statements are the "group partition" versions of Lemmas 2.3, 2.4 respectively.

COROLLARY 3.3. *If n is even then Z_2^n has a group partition of type Z_2^2, \dots, Z_2^2 ($(2^n - 1)/3$ terms).*

COROLLARY 3.4. *If n is odd then Z_2^n has a group partition of type $Z_2^3; Z_2^2, \dots, Z_2^2$ ($(2^n - 8)/3$ terms).*

Both of these corollaries follow from the observation that every zero sum triple $T = \{a, b, a + b\}$ in $(Z_2^n)^*$ is isomorphic to $(Z_2^2)^*$, and that a set S in $(Z_2^n)^*$ is isomorphic to $(Z_2^3)^*$ if and only if S is the union of a zero sum triple $T = \{a, b, a + b\}$ and a zero sum quadruple of the form $Q = \{d, a + d, b + d, a + b + d\}$.

Using the methods of § 2, we have determined all possible group partitions of Z_2^n for $2 \leq n \leq 6$ when $m_i > 1$.

For $n = 2$ and 3, Z_2^n has no nontrivial group partition. For $n = 4$ and 5, Z_2^n has a unique nontrivial group partition and it is given by Corollaries 3.3 and 3.4 respectively. For $n = 6$, Z_2^n has nontrivial group partitions for the following five types:

- (1) Z_2^2, \dots, Z_2^2 (21 terms),
- (2) $Z_2^3, Z_2^3, Z_2^3, Z_2^2, \dots, Z_2^2$ (14 terms),
- (3) Z_2^3, \dots, Z_2^3 (6 terms), Z_2^2, \dots, Z_2^2 (7 terms),
- (4) Z_2^3, \dots, Z_2^3 (9 terms),
- (5) $Z_2^4; Z_2^2, \dots, Z_2^2$ (16 terms).

We conclude this section with the following example: From conditions (A) and (B) it would be possible for Z_2^8 to have a nontrivial group partition of type $Z_2^2; Z_2^3, \dots, Z_2^3$ (36 terms). With a simple counting argument we will show, using some of the concepts developed in this section, that this is impossible. Let $\bar{a} = a_1 \cdots a_8$ ($a_i = 0$ or 1) denote a typical element in Z_2^8 and let $X = \{\bar{a} \in (Z_2^8)^* | a_1 = 0\}$. If $S_0 \cong Z_2^2$ then we must have $S_0^* = \{\bar{a}, \bar{b}, \bar{c} = \bar{a} + \bar{b}\}$. For some i ($i = 1, \dots, 8$) we must have

that a_i , b_i and c_i are not all 0. Without loss of generality assume $i = 1$. Since \bar{a} , \bar{b} and \bar{c} do not all belong to X , only one of them does, say \bar{a} . For $S_i \cong Z_2^3$ ($i = 1, \dots, 36$), it is easy to see that $|S_i^* \cap X| = 3$ or 7 . Thus, the remaining 126 elements of X (other than \bar{a}) must be partitioned into h sets of cardinality 3 and $(36 - h)$ sets of cardinality 7. Clearly this is impossible.

REFERENCES

- [1] T. BU, *Partitions of a vector space*, Discrete Math., 31 (1980), pp. 79–83.
- [2] R. J. FRIEDLANDER, B. GORDON AND P. TANNENBAUM, *Partitions of groups and complete mappings*, Pacific J. Math., 92 (1981), pp. 283–293.
- [3] M. HERZOG AND J. SCHONHEIM, *Linear and nonlinear single error-correcting perfect mixed codes*, Inform. and Control, 18 (1971), pp. 364–368.
- [4] ———, *Group partition, factorization and the vector covering problem*, Canad. Math. Bull., 15 (1972), pp. 207–214.
- [5] B. LINDSTROM, *Group partitions and mixed perfect codes*, Canad. Math. Bull., 18 (1975), pp. 57–60.
- [6] G. A. MILLER, *Groups in which all the operators are contained in a series of subgroups such that any two have only identity in common*, Bull. Amer. Math. Soc., 12 (1906), pp. 446–449.
- [7] P. TANNENBAUM, *Partitions of abelian groups into sets with zero sums*, Congressus Numerantium, 33 (1981), pp. 341–348.
- [8] J. W. YOUNG, *On the partitions of a group and the resulting classification*, Bull. Amer. Math. Soc., 33 (1927), pp. 453–461.

THE OPTIMAL LATTICE QUANTIZER IN THREE DIMENSIONS*

E. S. BARNES† AND N. J. A. SLOANE‡

Abstract. The body-centered cubic lattice is shown to have the smallest mean squared error of any lattice quantizer in three dimensions, assuming that the input to the quantizer has a uniform distribution.

1. Introduction. Let Λ be a lattice in real three-dimensional Euclidean space \mathbb{R}^3 . Around each lattice point $t \in \Lambda$ is its *Voronoi* (or nearest neighbor) region $S(t)$, consisting of all points of the space that are at least as close to t as to any other lattice point ([1], [2], [5], [14]). The Voronoi regions $S(t)$ are all congruent, and have volume \sqrt{D} , where D is the determinant of Λ , i.e. the square of the volume of a fundamental cell of Λ . If Λ is used as a quantizer, for quantizing data that is uniformly distributed over a large region of \mathbb{R}^3 , its average mean squared error per symbol is given by

$$(1) \quad G = G(\Lambda) = \frac{1}{3} \frac{\int_{S(0)} \tau \cdot \tau d\tau}{D^{5/6}}$$

—see [5]–[7], [9], [15]. (This formula ignores the fact that points near the boundary of the input region have irregular Voronoi regions, and so applies to the case when the number of output levels of the quantizer is very large.) $G(\Lambda)$ is a normalized second moment of $S(0)$, the Voronoi region around the origin, the denominator being determined by the condition that $G(\Lambda)$ should be dimensionless.

It was conjectured by Gersho in [9] that the body-centered cubic lattice D_3^* has the smallest value of $G(\Lambda)$ of any three-dimensional lattice. It is the goal of this paper to establish that conjecture. Furthermore, we shall see that there is no other lattice for which $G(\Lambda)$ is even a local minimum. Thus our main result is the following:

THEOREM 1. *For any three-dimensional lattice Λ , $G(\Lambda) \geq 19/(192 \cdot 2^{1/3}) = 0.0785433 \dots$, with equality if and only if Λ is equivalent to the body-centered cubic lattice D_3^* . Furthermore, for no other lattice is $G(\Lambda)$ a local minimum.*

Three (of the infinitely many) lattices which compete with the body-centered cubic lattice are the face-centered cubic lattice D_3 , for which $G = 2^{-11/3} = 0.0787451 \dots$; the lattice $\sqrt{3}A_2 \oplus \sqrt{5}\mathbb{Z}$, where A_2 is the hexagonal lattice in the plane with minimum norm 2, for which $G = 5^{2/3}/36 = 0.0812227 \dots$; and the cubic lattice \mathbb{Z}^3 , with $G = 1/12 = 0.0833333 \dots$. However, as we shall see, these three values of G can all be reduced by perturbing the lattices slightly. The Voronoi regions for D_3^* , D_3 , $\sqrt{3}A_2 \oplus \sqrt{5}\mathbb{Z}$, and \mathbb{Z}^3 are respectively truncated octahedra, rhombic dodecahedra, hexagonal prisms, and cubes (see [5], [9]).

Finally, it is worth pointing out that our results have wider application than to just uniformly distributed data, because (i) Zador (see [8], [9], [15]) has reduced the problem of finding the minimal quantization error for data with any integrable density function to that of solving the uniformly distributed case, and (ii) the so-called *companding* techniques for quantizing (see [9], [10]) handle nonuniform data by first applying a nonlinear transformation, then a uniform quantizer, and finally the inverse transformation.

The proof of Theorem 1 will be given in §§ 2 and 3. In § 2 we first recall some properties of lattices in \mathbb{R}^3 , in particular the fact that a lattice Λ can be represented

* Received by the editors December 29, 1981, and in revised form March 22, 1982.

† Department of Pure Mathematics, The University of Adelaide, Adelaide, S.A. 5001, Australia.

‡ Mathematics and Statistics Research Center, Bell Laboratories, Murray Hill, New Jersey 07974.

by a vector $[\rho_{01}, \rho_{02}, \rho_{03}, \rho_{12}, \rho_{13}, \rho_{23}]$ with six nonnegative real components. We then derive a fundamental formula (Theorem 2) which expresses $G(\Lambda)$ in terms of the ρ_{ij} . In § 3 we complete the proof by showing that the only local minimum of this expression for $G(\Lambda)$ occurs when the ρ_{ij} are all equal, which is precisely the case when Λ is a body-centered cubic lattice.

2. A formula for $G(\Lambda)$. In this paper all vectors are column vectors, written for example as $\mathbf{t} = (t_1, t_2, t_3)^{\text{tr}}$, where tr denotes transpose. As is customary, we shall represent lattices by their associated quadratic forms (see [4], [11]). If a lattice Λ in \mathbb{R}^3 is spanned by three vectors $\mathbf{t}^{(1)} = (t_1^1, t_2^1, t_3^1)^{\text{tr}}, \dots, \mathbf{t}^{(3)} = (t_1^3, t_2^3, t_3^3)^{\text{tr}}$, then $f(\mathbf{x}) = f(x_1, x_2, x_3) = \mathbf{x}^{\text{tr}} \mathbf{A} \mathbf{x}$ is a quadratic form associated with Λ , where $\mathbf{A} = \mathbf{T}^{\text{tr}} \mathbf{T}$ and \mathbf{T} has columns $\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \mathbf{t}^{(3)}$. A typical lattice point can be described in three ways, either by its Euclidean coordinates $\mathbf{t} = (t_1, t_2, t_3)^{\text{tr}}$, its \mathbf{x} -coordinates $\mathbf{x} = (x_1, x_2, x_3)^{\text{tr}}$, where x_1, x_2, x_3 are integers satisfying $\mathbf{t} = \mathbf{T} \mathbf{x}$, or by its \mathbf{y} -coordinates $\mathbf{y} = (y_1, y_2, y_3)^{\text{tr}}$, given by $\mathbf{y} = \mathbf{A} \mathbf{x}$. The *norm*, or squared distance from the origin, of this point is

$$(2) \quad \mathbf{t} \cdot \mathbf{t} = \mathbf{t}^{\text{tr}} \mathbf{t} = \mathbf{x}^{\text{tr}} \mathbf{A} \mathbf{x} = f(\mathbf{x}) = f^{-1}(\mathbf{y}),$$

where $f^{-1}(\mathbf{x}) = \mathbf{x}^{\text{tr}} \mathbf{A}^{-1} \mathbf{x}$ is the inverse form of f .

Two lattices Λ and M are equivalent, written $\Lambda \cong M$, if one can be obtained from the other by a rotation and change of scale. Two forms $f(\mathbf{x}) = \mathbf{x}^{\text{tr}} \mathbf{A} \mathbf{x}$ and $g(\mathbf{x}) = \mathbf{x}^{\text{tr}} \mathbf{B} \mathbf{x}$ are equivalent if $\mathbf{B} = \mathbf{U}^{\text{tr}} \mathbf{A} \mathbf{U}$, where \mathbf{U} is integral and $\det \mathbf{U} = \pm 1$ [4], [11].

If Λ is a lattice in \mathbb{R}^3 , Voronoi (see [1], [11], [13, p. 150]) has shown that Λ has a quadratic form of the shape

$$f(x_1, x_2, x_3) = \rho_{01}x_1^2 + \rho_{02}x_2^2 + \rho_{03}x_3^2 + \rho_{12}(x_1 - x_2)^2 + \rho_{13}(x_1 - x_3)^2 + \rho_{23}(x_2 - x_3)^2$$

associated with it, where the ρ_{ij} are nonnegative. If we define $x_0 = 0$, $\rho_{ii} = 0$, and $\rho_{ij} = \rho_{ji}$ for $i > j$, this may be written more symmetrically as

$$(3) \quad f(x_1, x_2, x_3) = \frac{1}{2} \sum_{i=0}^3 \sum_{j=0}^3 \rho_{ij} (x_i - x_j)^2.$$

Thus Λ is represented by the six nonnegative parameters $[\rho_{01}, \rho_{02}, \rho_{03}, \rho_{12}, \rho_{13}, \rho_{23}]$. In general Λ has 24 such representations, corresponding to the $4!$ permutations of the subscripts of the ρ_{ij} [1, Lem. 2.1]; multiplying f by a scalar leads to an equivalent lattice. For example, applying the permutation (01), we find that $[\rho_{01}, \rho_{12}, \rho_{13}, \rho_{02}, \rho_{03}, \rho_{23}]$ also represents Λ .

For later reference we mention that the body-centered cubic lattice D_3^* may be represented by the parameters $[1, 1, 1, 1, 1, 1]$, D_3 by $[0, 1, 1, 1, 1, 0]$ (more generally with any pair $\rho_{ij} = \rho_{kl} = 0$, where i, j and k, l are disjoint subscripts, and all other ρ_{ij} equal), \mathbb{Z}^3 by $[1, 1, 1, 0, 0, 0]$ for example, and $\sqrt{3}A_2 \oplus \sqrt{5}Z$ by $[3, 3, 5, 3, 0, 0]$.

Our main result in this section is the following:

THEOREM 2. *The average mean squared error of Λ (or the normalized second moment of the Voronoi region $S(\mathbf{0})$) is given by*

$$(4) \quad G(\Lambda) = \frac{D \cdot S_1 + 2S_2 + K}{36 D^{4/3}},$$

where

$$(5) \quad \begin{aligned} D &= \det \Lambda = \sum^{(4)} \rho_{01} \rho_{02} \rho_{03} + \sum^{(3)} \rho_{01} \rho_{23} (\rho_{02} + \rho_{03} + \rho_{12} + \rho_{13}), \\ S_1 &= \rho_{01} + \rho_{02} + \rho_{03} + \rho_{12} + \rho_{13} + \rho_{23}, \\ S_2 &= \rho_{01} \rho_{02} \rho_{13} \rho_{23} + \rho_{01} \rho_{03} \rho_{12} \rho_{23} + \rho_{02} \rho_{03} \rho_{12} \rho_{13}, \end{aligned}$$

and

$$K = \sum^{(4)} \rho_{01}\rho_{02}\rho_{03}(\rho_{12} + \rho_{13} + \rho_{23}).$$

We are using the standard notation for symmetric functions, so that $\sum^{(4)} \rho_{01}\rho_{02}\rho_{03}$, for example, is an abbreviation for $\rho_{01}\rho_{02}\rho_{03} + \rho_{01}\rho_{12}\rho_{13} + \rho_{02}\rho_{12}\rho_{23} + \rho_{03}\rho_{13}\rho_{23}$, and the superscript indicates the number of distinct summands. The formula for $\det \Lambda$ was given in [1], and the quantity K also occurs there. In the proof of Theorem 2 we shall make considerable use of the information about the Voronoi regions of Λ given in [1], and in § 3 the proof of Theorem 1 is modeled on the proof of the main result in [1] (although the techniques used are quite different).

Proof of Theorem 2. The matrix A associated with the quadratic form (3) is

$$A = \begin{bmatrix} \rho_{01} + \rho_{12} + \rho_{13} & -\rho_{12} & -\rho_{13} \\ -\rho_{12} & \rho_{02} + \rho_{12} + \rho_{23} & -\rho_{23} \\ -\rho_{13} & -\rho_{23} & \rho_{03} + \rho_{13} + \rho_{23} \end{bmatrix},$$

and $\det A = \det \Lambda = D$ gives (5). The norm of a vector (see (2)) is best expressed in terms of its y -coordinates, and is given by

$$(6) \quad f^{-1}(\mathbf{y}) = \frac{1}{D} \{ (\rho_{12}\rho_{13} + \rho_{12}\rho_{23} + \rho_{13}\rho_{23})y_0^2 + (\rho_{02}\rho_{03} + \rho_{02}\rho_{23} + \rho_{03}\rho_{23})y_1^2 \\ + (\rho_{01}\rho_{03} + \rho_{01}\rho_{13} + \rho_{03}\rho_{13})y_2^2 + (\rho_{01}\rho_{02} + \rho_{01}\rho_{12} + \rho_{02}\rho_{12})y_3^2 \\ + \rho_{03}\rho_{12}(y_1 + y_2)^2 + \rho_{02}\rho_{13}(y_1 + y_3)^2 + \rho_{01}\rho_{23}(y_2 + y_3)^2 \},$$

where we have set $y_0 = -y_1 - y_2 - y_3$.

The Voronoi region $S(\mathbf{0})$ is described in [1]. It is a (possibly degenerate) truncated octahedron, with in general 14 faces, given by

$$F_i : 2y_i = \sum_{l \neq i}^{(3)} \rho_{il},$$

$$F_{ij} : 2(y_i + y_j) = \sum_{l \neq i, j}^{(2)} (\rho_{il} + \rho_{jl}),$$

$$F_{ijk} : 2(y_i + y_j + y_k) = \sum_{l \neq i, j, k}^{(1)} (\rho_{il} + \rho_{jl} + \rho_{kl}),$$

where all subscripts and summations run from 0 to 3, and the subscripts on F are unordered. There are in general 24 vertices v_{ijk} , where the subscripts are an ordered 3-subset of $\{0, 1, 2, 3\}$. For example, the vertex v_{123} lies at the intersection of the faces F_1, F_{12} and F_{123} (see [1, Eq. (2.4)]). The Voronoi region is sketched in Fig. 1.

If P is any polyhedron in \mathbb{R}^3 , we define its unnormalized second moment $U(P)$, its moment of inertia $I(P)$, and its normalized second moment $G(P)$ (all about the origin) by

$$U(P) = \int_P \boldsymbol{\tau} \cdot \boldsymbol{\tau} d\boldsymbol{\tau},$$

$$I(P) = \frac{U(P)}{\text{Volume}(P)},$$

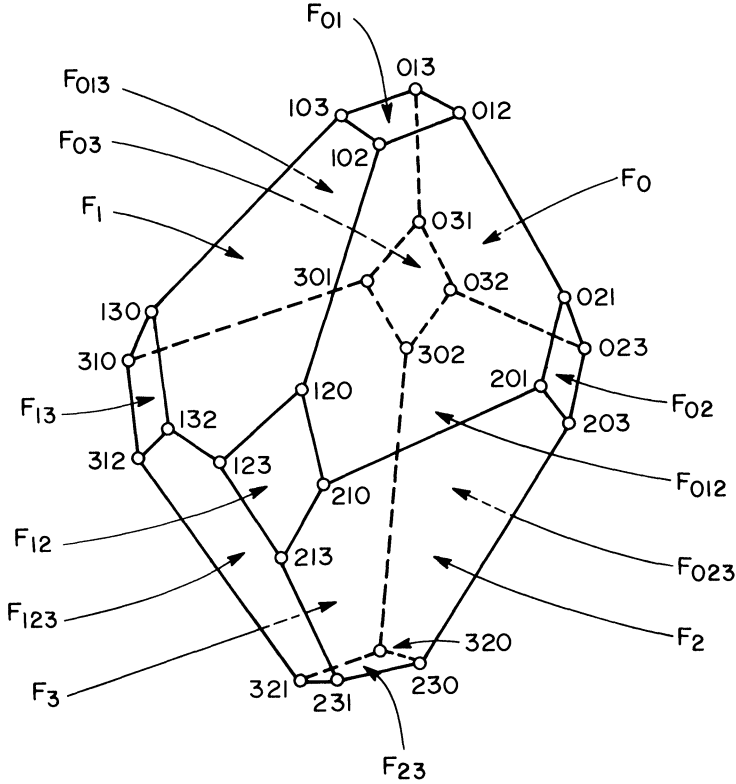


FIG. 1. Voronoi region $S(\mathbf{0})$ for a three-dimensional lattice, showing the 14 faces F_i, F_{ij}, F_{ijk} and the 24 vertices v_{ijk} (only the subscripts are given). The faces F_i, F_{jkl} and the faces F_{ij}, F_{kl} are parallel, where i, j, k, l is any permutation of 0, 1, 2, 3.

and

$$G(\mathbf{P}) = \frac{1}{3} \frac{U(\mathbf{P})}{\text{Volume}(\mathbf{P})^{5/3}}.$$

Then the theorem asserts that $G(S(\mathbf{0}))$ is given by (4). To compute $G(S(\mathbf{0}))$ we shall dissect $S(\mathbf{0})$ into 60 tetrahedra, and use the fact that there is an explicit formula (see for example [5]) for the moment of inertia of a tetrahedron. In fact, if T is a tetrahedron with vertices $\mathbf{0}, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ then its barycenter is $\mathbf{q} = (\mathbf{p}_1 + \mathbf{p}_2 + \mathbf{p}_3)/4$, and

$$(7) \quad I(T) = \frac{4}{5} \mathbf{q} \cdot \mathbf{q} + \frac{1}{20} (\mathbf{p}_1 \cdot \mathbf{p}_1 + \mathbf{p}_2 \cdot \mathbf{p}_2 + \mathbf{p}_3 \cdot \mathbf{p}_3).$$

If $\mathbf{0}, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ are the y -coordinates of the vertices, then T has volume $(6\sqrt{D})^{-1} \det(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3)$.

We first consider a hexagonal face, say F_1 , and divide it into six triangles meeting at the center (\mathbf{m}_1) of the face. By joining these triangles to $\mathbf{0}$ we form six tetrahedra. Let us analyze one of these tetrahedra, say the tetrahedron T_{12} with vertices

$\mathbf{0}$, \mathbf{m}_1 , \mathbf{v}_{120} , \mathbf{v}_{123} . The y -coordinates for its vertices are easily found to be

$$\mathbf{0}: (0, 0, 0),$$

$$\mathbf{m}_1: \frac{1}{2}(\rho_{01} + \rho_{12} + \rho_{13}, -\rho_{12}, -\rho_{13}),$$

$$\mathbf{v}_{120}: \frac{1}{2}(\rho_{01} + \rho_{12} + \rho_{13}, \rho_{02} - \rho_{12} + \rho_{23}, -\rho_{03} - \rho_{13} - \rho_{23}),$$

$$\mathbf{v}_{123}: \frac{1}{2}(\rho_{01} + \rho_{12} + \rho_{13}, \rho_{02} - \rho_{12} + \rho_{23}, \rho_{03} - \rho_{13} - \rho_{23}).$$

(The x -coordinates for \mathbf{m}_1 are $(\frac{1}{2}, 0, 0)$.) The norms of these vectors can be obtained from (6). It turns out that

$$(8) \quad \begin{aligned} \mathbf{m}_1 \cdot \mathbf{m}_1 &= \frac{1}{4}(\rho_{01} + \rho_{12} + \rho_{13}), \\ \mathbf{v}_{120} \cdot \mathbf{v}_{120} &= \frac{1}{4D} \{D \cdot S_1 - K - 4\lambda_2\lambda_3\}, \end{aligned}$$

$$(9) \quad \mathbf{v}_{123} \cdot \mathbf{v}_{123} = \frac{1}{4D} \{D \cdot S_1 - K - 4\lambda_1\lambda_3\},$$

where D , S_1 and K have already been defined (see Theorem 1) and

$$\lambda_1 = \rho_{01}\rho_{23}, \lambda_2 = \rho_{02}\rho_{13}, \lambda_3 = \rho_{03}\rho_{12}.$$

(Formulas (8) and (9) are given in [1].) To find the moment of inertia $I(T_{12})$, we compute the barycenter $\mathbf{q} = \frac{1}{4}(\mathbf{m}_1 + \mathbf{v}_{120} + \mathbf{v}_{123})$ and use (7), eventually obtaining

$$\begin{aligned} I(T_{12}) &= \frac{1}{40D} \{D(6\rho_{01} + 3\rho_{02} + \rho_{03} + 6\rho_{12} + 6\rho_{13} + 3\rho_{23}) \\ &\quad - \rho_{01}\rho_{02}\rho_{03}(3\rho_{12} + \rho_{13} + \rho_{23}) - 3\rho_{01}\rho_{02}\rho_{12}(\rho_{13} + \rho_{23}) \\ &\quad - \rho_{01}\rho_{03}(\rho_{12}\rho_{13} + 4\rho_{12}\rho_{23} + \rho_{13}\rho_{23}) - \rho_{02}\rho_{03}(4\rho_{12}\rho_{13} + \rho_{12}\rho_{23} + \rho_{13}\rho_{23}) \\ &\quad - 3\rho_{12}\rho_{13}\rho_{23}(\rho_{01} + \rho_{02} + \rho_{03})\}. \end{aligned}$$

Also the volume of T_{12} is

$$\frac{1}{24\sqrt{D}} \rho_{03}(\rho_{02} + \rho_{23})(\rho_{01} + \rho_{12} + \rho_{13}).$$

The six tetrahedra meeting at the face F_1 are:

Name	Vertices
T_{12}	$(\mathbf{0}, \mathbf{m}_1, \mathbf{v}_{120}, \mathbf{v}_{123})$
$T_{1\bar{3}}$	$(\mathbf{0}, \mathbf{m}_1, \mathbf{v}_{102}, \mathbf{v}_{120})$
T_{10}	$(\mathbf{0}, \mathbf{m}_1, \mathbf{v}_{103}, \mathbf{v}_{102})$
$T_{1\bar{2}}$	$(\mathbf{0}, \mathbf{m}_1, \mathbf{v}_{130}, \mathbf{v}_{103})$
T_{13}	$(\mathbf{0}, \mathbf{m}_1, \mathbf{v}_{132}, \mathbf{v}_{130})$
$T_{1\bar{0}}$	$(\mathbf{0}, \mathbf{m}_1, \mathbf{v}_{123}, \mathbf{v}_{132})$

We find that $I(T_{1\bar{2}}) = I(T_{12})$, $I(T_{1\bar{3}}) = I(T_{13}) = \pi_{23}I(T_{12})$, $I(T_{1\bar{0}}) = I(T_{10}) = \pi_{02}I(T_{12})$, where π_{ab} denotes the transposition (ab) applied to the subscripts of the ρ_{ij} (e.g. $\pi_{02}(\rho_{23}) = \rho_{03}$). Up to this point the calculations may be performed by hand. But to proceed further a computer is desirable. We used the symbolic manipulation program Altran [3]. The contribution to the unnormalized second moment $U(S(\mathbf{0}))$ from all

six tetrahedra meeting the face F_1 is

$$\sum^{(6)} \text{Volume}(T_{ij})I(T_{ij}) = U_1 \quad (\text{say}).$$

The total contribution from all eight hexagonal faces is then

$$U_{\text{hex}} = 2(U_1 + \pi_{01}U_1 + \pi_{12}U_1 + \pi_{13}U_1).$$

It remains to consider the quadrilateral faces. It is simplest to divide each of them into two triangles. For example, we divide the quadrilateral face F_{12} into the triangles $\mathbf{v}_{123}, \mathbf{v}_{120}, \mathbf{v}_{213}$ and $\mathbf{v}_{120}, \mathbf{v}_{213}, \mathbf{v}_{210}$. Proceeding as before, we find that the moments of inertia of the corresponding tetrahedra $\mathbf{0}, \mathbf{v}_{123}, \mathbf{v}_{120}, \mathbf{v}_{213}$ and $\mathbf{0}, \mathbf{v}_{120}, \mathbf{v}_{213}, \mathbf{v}_{210}$ are both equal to

$$\frac{1}{20D} \{D(3\rho_{01} + 3\rho_{02} + \rho_{03} + \rho_{12} + 3\rho_{13} + 3\rho_{23}) - K - 2\lambda_1\lambda_3 - 2\lambda_2\lambda_3\},$$

and that both tetrahedra have volume

$$\frac{1}{12\sqrt{D}} \rho_{03}\rho_{12}(\rho_{01} + \rho_{02} + \rho_{13} + \rho_{23}).$$

Twice the product of these two expressions gives the contribution U_{12} to $U(S(\mathbf{0}))$ from the face F_{12} . The total contribution from all six quadrilateral faces is then

$$U_{\text{quad}} = 2(U_{12} + U_{23} + U_{13}) = 2(U_{12} + \pi_{13}U_{12} + \pi_{23}U_{12}).$$

Finally we use Altran to compute $U(S(\mathbf{0})) = U_{\text{hex}} + U_{\text{quad}}$, and $G(S(\mathbf{0})) = U(S(\mathbf{0}))/3D^{5/6}$. After a factor D is removed from the numerator, the result is the right-hand side of (4), which proves Theorem 2. \square

3. Minimizing $G(\Lambda)$. We complete the proof of Theorem 1 by establishing the following result:

THEOREM 3. *The only local minimum of the right-hand side of (4) subject to the constraints $\rho_{ij} \geq 0$ (for all i, j) and $D \neq 0$ occurs when all the ρ_{ij} are equal.*

Proof. Our method is the one used in [1], namely, to exhibit small variations in the ρ_{ij} which will reduce the right-hand side of (4) unless all the ρ_{ij} are equal. For convenience we define

$$\begin{aligned} \boldsymbol{\rho} &= [\rho_{01}, \rho_{02}, \rho_{03}, \rho_{12}, \rho_{13}, \rho_{23}], \\ G(\boldsymbol{\rho}) &= (D \cdot S_1 + 2S_2 + K)/36D^{4/3}, \\ N &= D \cdot S_1 + 2S_2 + K. \end{aligned}$$

The proof will be divided into several steps.

Step 3.1. $D_3, \sqrt{3}A_2 \oplus \sqrt{5}\mathbb{Z}$ and \mathbb{R}^3 are not local minima. In fact, with ε small and positive, when

$$\begin{aligned} \boldsymbol{\rho} &= [\varepsilon, 1, 1, 1, 1, \varepsilon], & G(\boldsymbol{\rho}) &= \frac{1}{2^{11/3}} \left(1 - \frac{4\varepsilon^3}{81} + \dots \right), \\ \boldsymbol{\rho} &= [3, 3, 5, 3, \varepsilon, 0], & G(\boldsymbol{\rho}) &= \frac{5^{2/3}}{36} \left(1 - \frac{209\varepsilon^2}{3^6 \cdot 5^2} + \dots \right), \\ \boldsymbol{\rho} &= [1, 1, 1, \varepsilon, 0, 0], & G(\boldsymbol{\rho}) &= \frac{1}{12} \left(1 - \frac{2\varepsilon^2}{9} + \dots \right). \end{aligned}$$

So in each case a small variation in ρ will reduce $G(\rho)$.

Step 3.2. D_3^ is a local minimum.* Since the ρ_{ij} are homogeneous coordinates for Λ , the effect of any variation of the ρ_{ij} on G is the same as the effect of a variation in which one of the ρ_{ij} , say ρ_{01} , is held constant. Temporarily setting $\rho_{02} = x_2$, $\rho_{03} = x_3, \dots, \rho_{23} = x_6$, we find that the first partial derivatives $\partial G/\partial x_i$ ($i = 2, \dots, 6$) vanish at $[1, 1, 1, 1, 1, 1]$, while the matrix of second partial derivatives, $(\partial^2 G/\partial x_i \partial x_j)(i, j = 2, \dots, 6)$, is equal to a constant times

$$\begin{bmatrix} 20 & -1 & -1 & -16 & -1 \\ -1 & 20 & -16 & -1 & -1 \\ -1 & -16 & 20 & -1 & -1 \\ -16 & -1 & -1 & 20 & -1 \\ -1 & -1 & -1 & -1 & 20 \end{bmatrix}.$$

Since the eigenvalues of this matrix are positive, the matrix itself is positive definite, and $[1, 1, 1, 1, 1, 1]$ is a local minimum of G .

It remains to show that there is no other local minimum. From now on we assume that $\bar{\rho} = [\bar{\rho}_{01}, \dots, \bar{\rho}_{23}]$ is a local minimum, and eventually deduce that all the $\bar{\rho}_{ij}$ must be equal.

Step 3.3. Not more than two $\bar{\rho}_{ij}$ may be zero. There are essentially only two cases in which three or more of the $\bar{\rho}_{ij}$ may be zero while D is nonzero, namely (a) $\bar{\rho}_{01} = \bar{\rho}_{02} = \bar{\rho}_{12} = 0$, and (b) $\bar{\rho}_{01} = \bar{\rho}_{02} = \bar{\rho}_{23} = 0$.

Case (a) Suppose $\bar{\rho} = [0, 0, 1, 0, y, z]$ with $y > 0, z > 0$. At $\bar{\rho}$ we must have

$$\frac{\partial G}{\partial \rho_{03}} = \frac{\partial G}{\partial \rho_{13}} = \frac{\partial G}{\partial \rho_{23}} = 0.$$

Now $\partial G/\partial \rho_{13} - \partial G/\partial \rho_{23} = yz(y - z)(y + z + 1)$ at $\bar{\rho}$, so $y = z$. Then $\partial G/\partial \rho_{03} = -2(y - 1)y^4$, so $y = z = 1$, and therefore $\bar{\rho} = [0, 0, 1, 0, 1, 1]$. But this is Z^3 , which we have already seen is not a local minimum. Case (b) is almost identical and is omitted.

Step 3.4. Some variations of the ρ_{ij} that fix D . We shall generally use δR to denote the first order variation in a function $R(\rho)$ resulting from small variations $\delta \rho_{ij}$. Let V_0 denote the following variation of the ρ_{ij} :

$$\begin{aligned} \rho_{01} &\rightarrow \rho_{01}, \\ \rho_{02} &\rightarrow \rho_{02} - \varepsilon \rho_{01}, \\ \rho_{03} &\rightarrow \rho_{03} + \varepsilon \rho_{01}, \\ \rho_{12} &\rightarrow \rho_{12} + \varepsilon(\rho_{01} + \rho_{12} + \rho_{13}), \\ \rho_{13} &\rightarrow \rho_{13} - \varepsilon(\rho_{01} + \rho_{12} + \rho_{13}), \\ \rho_{23} &\rightarrow \rho_{23} + \varepsilon(\rho_{12} - \rho_{13}), \end{aligned}$$

where ε is small. When applying V_0 , we must be careful to ensure that the ρ_{ij} remain nonnegative. For example, we may not apply V_0 with ε positive if $\rho_{02} = 0$ and $\rho_{01} > 0$, since the new value of ρ_{02} would be negative. The variation V_0 has the useful property that it fixes D to the first order in ε . To see this, we note that

$$\frac{\partial D}{\partial \rho_{01}} = \sigma_0 + \sigma_1 + \lambda_2 + \lambda_3,$$

etc., where

$$\sigma_i = \rho_{jk}\rho_{jl} + \rho_{jk}\rho_{kl} + \rho_{jl}\rho_{kl}$$

i, j, k, l being a permutation of 0, 1, 2, 3 (see [1, p. 297]). Then the fact that

$$\delta D = \sum \frac{\partial D}{\partial \rho_{ij}} \delta \rho_{ij} = 0$$

is an immediate consequence of the identity

$$\sigma_2 \rho_{12} + \lambda_2 (\rho_{01} + \rho_{12}) = \sigma_3 \rho_{13} + \lambda_3 (\rho_{01} + \rho_{13}).$$

Thus the denominator of G is fixed by V_0 to the first order. The numerator is increased by

$$(10) \quad \delta(N) = -2\varepsilon J_0,$$

where

$$(11) \quad J_0 = \rho_{03} \rho_{12} \{(\rho_{01} + \rho_{13})^2 - \rho_{12}(\rho_{01} + \rho_{13})\} \\ - \rho_{02} \rho_{13} \{(\rho_{01} + \rho_{12})^2 - \rho_{13}(\rho_{01} + \rho_{12})\} + \rho_{01} \rho_{12} \rho_{13} (\rho_{13} - \rho_{12}).$$

Although this formula (and others such as (4), (12) and (19)) could have been obtained by hand, it was actually derived with the aid of the interactive symbolic manipulation program Macsyma [12]. Nevertheless, the computer did not produce (11) in its present form. Considerable manipulation by hand is almost always required to transform the computer's output into the most appropriate form. This is especially true of (4), (12) and (19).

The expression J_0 is linear in ρ_{02} and ρ_{03} , does not involve ρ_{23} , and goes into $-J_0$ under π_{23} . Other variations of the ρ_{ij} can be obtained from V_0 by applying suitable permutations of the subscripts. We shall require the transformations

$$V_1, V_2, V_3, \quad V_4, \quad V_5, \quad V_6,$$

which are obtained by applying the permutations

$$\pi_{12}, \pi_{01}, \pi_{02}, \pi_{02}\pi_{01} = (012), (021), (132),$$

respectively to V_0 . Under $V_i (i = 1, \dots, 6)$ we have $\delta(N) = -2\varepsilon J_i$, where J_i is obtained from J_0 by applying the permutation that produced V_i from V_0 . To be quite explicit we write out V_1 and J_1 in full:

$$V_1 : \rho_{01} \rightarrow \rho_{01} - \varepsilon \rho_{02}, \\ \rho_{02} \rightarrow \rho_{02}, \\ \rho_{03} \rightarrow \rho_{03} + \varepsilon \rho_{02}, \\ \rho_{12} \rightarrow \rho_{12} + \varepsilon (\rho_{02} + \rho_{12} + \rho_{23}), \\ \rho_{13} \rightarrow \rho_{13} + \varepsilon (\rho_{12} - \rho_{23}), \\ \rho_{23} \rightarrow \rho_{23} - \varepsilon (\rho_{02} + \rho_{12} + \rho_{23}),$$

$$J_1 = \rho_{03} \rho_{12} \{(\rho_{02} + \rho_{23})^2 - \rho_{12}(\rho_{02} + \rho_{23})\} \\ - \rho_{01} \rho_{23} \{(\rho_{02} + \rho_{12})^2 - \rho_{23}(\rho_{02} + \rho_{12})\} + \rho_{02} \rho_{12} \rho_{23} (\rho_{23} - \rho_{12}).$$

Step 3.5. Two $\bar{\rho}_{ij}$ cannot be simultaneously zero. Again there are essentially only two cases: (a) $\bar{\rho}_{01} = \bar{\rho}_{23} = 0$, with disjoint subscripts, or (b) $\bar{\rho}_{01} = \bar{\rho}_{02} = 0$, with overlapping subscripts.

Case (a). We assume $\bar{\rho}_{01} = \bar{\rho}_{23} = 0$, with the remaining $\bar{\rho}_{ij} > 0$. V_1 is a valid variation if $\varepsilon < 0$, and then

$$\delta(N) = (-2\varepsilon)\bar{\rho}_{02}\bar{\rho}_{03}\bar{\rho}_{12}(\bar{\rho}_{02} - \bar{\rho}_{12}).$$

V_4 is also valid if $\varepsilon < 0$, and

$$\delta(N) = (-2\varepsilon)\bar{\rho}_{02}\bar{\rho}_{12}\bar{\rho}_{13}(\bar{\rho}_{12} - \bar{\rho}_{02}).$$

Since these variations have opposite signs, $\bar{\rho}$ is not a local minimum unless $\bar{\rho}_{02} = \bar{\rho}_{12}$. Applying the permutations (23) and (03)(12) (which leave the assumption $\bar{\rho}_{01} = \bar{\rho}_{23} = 0$ invariant) we obtain the further necessary conditions $\bar{\rho}_{03} = \bar{\rho}_{13}$ and $\bar{\rho}_{13} = \bar{\rho}_{12}$. Thus $\bar{\rho}$ is a multiple of $[0, 1, 1, 1, 1, 0]$, which we have seen is not a local minimum.

Case (b). With $\bar{\rho}_{01} = \bar{\rho}_{02} = 0$ and the other $\bar{\rho}_{ij} > 0$, we may apply the variation V_0 with an ε of either sign, and so $\delta(N) = -2\varepsilon J_0 = -2\varepsilon\bar{\rho}_{03}\bar{\rho}_{12}\bar{\rho}_{13}(\bar{\rho}_{13} - \bar{\rho}_{12}) = 0$, hence $\bar{\rho}_{12} = \bar{\rho}_{13}$. Similarly V_1 leads to $\bar{\rho}_{12} = \bar{\rho}_{23}$. Then $\partial G/\partial\rho_{03} = 0$ gives $\bar{\rho}_{12} = \frac{3}{5}$, so $\bar{\rho}$ is a multiple of $[0, 0, 1, \frac{3}{5}, \frac{3}{5}, \frac{3}{5}]$. But this is $\sqrt{3}A_2 \oplus \sqrt{5}\mathbb{Z}$, which is also not a local minimum.

Step 3.6. No single $\bar{\rho}_{ij}$ may be zero. We may assume $\bar{\rho}_{01} = 0$, the other $\bar{\rho}_{ij} > 0$. Using V_0 we obtain $J_0 = 0$, or $\bar{\rho}_{12} = \bar{\rho}_{13}$, and similarly $\bar{\rho}_{02} = \bar{\rho}_{03}$ from V_2 . But $\bar{\rho} = [0, \bar{\rho}_{02}, \bar{\rho}_{02}, \bar{\rho}_{12}, \bar{\rho}_{12}, \bar{\rho}_{23}]$ is not a local minimum. For if we evaluate G at $[\varepsilon, v, v, 1, 1, z]$ we find $G = G_0 - \varepsilon G_1 + \text{higher order terms}$, where

$$G_1 = \frac{(v+1)z^2 + (v-1)^2(v+3z+1)}{108 \cdot 2^{1/3} v^{4/3} (v+2z+1)^{4/3}} > 0.$$

Step 3.7. The path lemma and its consequences. We may now assume that all $\bar{\rho}_{ij}$ are greater than zero. Then all the variations V_0, \dots, V_6 may be used without restriction. Certainly we must have $E \triangleq -(J_3 + J_4) = 0$ at $\bar{\rho}$. But E may be written as

$$(12) \quad E = \rho_{12}(\rho_{01} + \rho_{02})(\rho_{13} + \rho_{23})(\rho_{01} + \rho_{02} - \rho_{13} - \rho_{23}) \\ + \rho_{01}\rho_{13}(\rho_{02} + \rho_{23})(\rho_{01} - \rho_{13}) + \rho_{02}\rho_{23}(\rho_{01} + \rho_{13})(\rho_{02} - \rho_{23}).$$

Therefore if $\bar{\rho}_{01} \cong \bar{\rho}_{13}$ we must have $\bar{\rho}_{02} \cong \bar{\rho}_{23}$, while $\bar{\rho}_{01} \cong \bar{\rho}_{13}$ implies $\bar{\rho}_{02} \cong \bar{\rho}_{13}$. We express this in words by saying that, of the two paths 0-1-3 and 0-2-3 in Fig. 2, one must rise and the other must fall (where rise means \cong , and fall means \cong). This holds between any pair of nodes, and so we may deduce:

LEMMA 1. (the path lemma). In Fig. 2, of the two paths $i-k-j$ and $i-l-j$ between any pair of nodes i, j , one must rise and the other must fall.

Let

$$\left\{ \begin{array}{l} \bar{\rho}_{13} \\ \bar{\rho}_{02} \end{array} \right\} \cong \left\{ \begin{array}{l} \bar{\rho}_{12} \\ \bar{\rho}_{03} \end{array} \right\}$$

be an abbreviation for the inequalities $\bar{\rho}_{13} \cong \bar{\rho}_{12}$, $\bar{\rho}_{13} \cong \bar{\rho}_{03}$, $\bar{\rho}_{02} \cong \bar{\rho}_{12}$ and $\bar{\rho}_{02} \cong \bar{\rho}_{03}$.

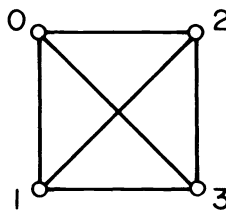


FIG. 2

LEMMA 2. *Without loss of generality we may assume that*

$$\left\{ \begin{array}{l} \bar{\rho}_{13} \\ \bar{\rho}_{02} \end{array} \right\} \cong \left\{ \begin{array}{l} \bar{\rho}_{12} \\ \bar{\rho}_{03} \end{array} \right\} \cong \left\{ \begin{array}{l} \bar{\rho}_{23} \\ \bar{\rho}_{01} \end{array} \right\}.$$

Proof. Without loss of generality, no $\bar{\rho}_{ij}$ is larger than $\bar{\rho}_{13}$. By applying Lemma 1 to the paths between nodes 0&1, 0&3, 1&2, and 2&3 we deduce that

$$\left\{ \begin{array}{l} \bar{\rho}_{13} \\ \bar{\rho}_{02} \end{array} \right\} \cong \left\{ \begin{array}{l} \bar{\rho}_{12} \\ \bar{\rho}_{03} \end{array} \right\} \quad \text{and} \quad \left\{ \begin{array}{l} \bar{\rho}_{13} \\ \bar{\rho}_{02} \end{array} \right\} \cong \left\{ \begin{array}{l} \bar{\rho}_{23} \\ \bar{\rho}_{01} \end{array} \right\}.$$

Consideration of the remaining pairs 0&2 and 1&3, and applying the transposition (13) if necessary, then leads to the desired conclusion. \square

LEMMA 3. *By suitably labeling the $\bar{\rho}_{ij}$, we may assume that either*

$$\bar{\rho}_{13} \cong \bar{\rho}_{02} \cong \bar{\rho}_{12} \cong \bar{\rho}_{03} \cong \bar{\rho}_{23} \cong \bar{\rho}_{01}$$

or

$$\bar{\rho}_{13} \cong \bar{\rho}_{02} \cong \bar{\rho}_{03} \cong \bar{\rho}_{12} \cong \bar{\rho}_{23} \cong \bar{\rho}_{01}.$$

Proof. We may assume that $\bar{\rho}_{13}$ is the largest $\bar{\rho}_{ij}$, and then by Lemma 2 the second largest is $\bar{\rho}_{02}$, corresponding to the edge in Fig. 2 that does not meet 13. Using (02) (13) we may assume that $\bar{\rho}_{23} \cong \bar{\rho}_{01}$, and finally there are two possibilities for the ranking of $\bar{\rho}_{12}$ and $\bar{\rho}_{03}$. \square

Step 3.8. Two $\bar{\rho}_{ij}$ cannot be equal unless Λ is equivalent to D_3^* . The subscripts either overlap or are disjoint.

Case (a). If they overlap, we may assume that $\bar{\rho}_{01} = \bar{\rho}_{13} = 1$. Let $\bar{\rho} = [1, v, w, x, 1, z]$. Then $E = 0$ becomes $(z - v)(vxz + vx + 2vz + xz + x) = 0$, so $v = z$; $J_6 = v(v - 1)(v + 4w + 1) = 0$ implies $v = 1$; $J_2 = 0$ implies $wx + w - 2x = 0$; $J_1 = 0$ implies $wx - 2w + x = 0$; and hence $v = w = x = z = 1$, which is D_3^* .

If the equal ρ_{ij} have disjoint subscripts, we shall invoke Lemma 3, and therefore there are three possibilities to consider.

Case (b), $\bar{\rho}_{01} = \bar{\rho}_{23}$. Let $\bar{\rho} = [1, v, w, x, y, 1]$, where by Lemma 3 we have $y \cong v \cong \{x, w\} \cong 1$. If any of these are equal we can apply Case (a), and so we may assume that $y > v > \{x, w\} > 1$. By solving $J_3 = 0$ we express v in terms of x and y , and substituting this into $J_4 = 0$ leads to the equation

$$(13) \quad (y - 1)((y - x)(xy + y + x) + y + x)(x^2y^2 + 3xy^2 + 2x^2y + 2y^2 - 2y + x^2 - x) = 0$$

(found by Macsyma). However, all three factors in (13) are visibly positive, so this case cannot occur.

Case (c), $\bar{\rho}_{02} = \bar{\rho}_{13}$. Let $\bar{\rho} = [u, 1, w, x, 1, z]$, where by Lemma 3 and Case (a) we may assume $1 > \{x, w\} > z > u > 0$. Applying the permutation (032) to (13) (or alternatively eliminating w and z from $J_0 = 0$, $J_3 = 0$ and $J_2 = 0$) leads us to

$$(14) \quad \begin{aligned} & -(1 - x)(ux(x - u) + u(1 - u) + x^2 + x) \\ & \times (u^2x^2 + 3ux^2 + 2u^2x + 2x^2 - 2x + u^2 - u) = 0. \end{aligned}$$

The first two factors in (14) are visibly positive, so

$$(15) \quad u^2x^2 + 3ux^2 + 2u^2x + 2x^2 - 2x + u^2 - u = 0.$$

Similarly applying the transposition (03) to (13) (or eliminating w and z from $J_0 = 0$, $J_3 = 0$ and $J_4 = 0$) leads us to

$$(16) \quad u^2x^2 + 2ux^2 + 3u^2x + x^2 - x + 2u^2 - 2u = 0.$$

Then (15)–(16) gives $(x-u)(ux+x+u-1)=0$, hence $x=(1-u)/(1+u)$, and from (16) we obtain $u=0$ or $u=1$, a contradiction.

Case (d), $\bar{\rho}_{03}=\bar{\rho}_{12}$. Let $\bar{\rho}=[u, v, 1, 1, y, z]$, where we may assume $y>v>1>z>u>0$. Applying (0321) to (13) (or eliminating y and z from $J_2=0, J_5=0, J_0=0$) leads us to

$$(v-1)(uv(v-u)+(v^2-u^2)+v+u)(u^2v^2+u(3v^2-1)+2u^2v+2v(v-1)+u^2)=0,$$

which is impossible since all three factors are positive.

Step 3.8. The remaining cases,

$$(17) \quad \bar{\rho}_{13} > \bar{\rho}_{02} > \bar{\rho}_{12} > \bar{\rho}_{03} > \bar{\rho}_{23} > \bar{\rho}_{01} > 0$$

and

$$(18) \quad \bar{\rho}_{13} > \bar{\rho}_{02} > \bar{\rho}_{03} > \bar{\rho}_{12} > \bar{\rho}_{23} > \bar{\rho}_{01} > 0,$$

are impossible. By Lemma 3 these are the only remaining cases. We set $\bar{\rho}=[u, v, w, x, y, 1]$ where $y>x>1>u>0$. By showing that this is impossible, we rule out both (17) and (18). As in Case (b) of the previous step, we solve $J_3=0$ for v and substitute into $J_4=0$. The numerator of the resulting expression is equal to $-u(y-u)$ times

$$\begin{aligned} & \{(y^4x^4-y^3x^5)+(2y^4x^3u-2y^3x^4u)+(3y^4x^3-3y^2x^5) \\ & + (6y^4x^2u-2y^3x^3u^2-3y^2x^4u)+(y^3x^3-y^2x^4) \\ & + (3y^4x^2-3yx^5)+(6y^4xu-6y^3x^2u^2)+(6y^3x^2u-4y^2x^3u^2)+(8y^2x^3u-6y^3xu^2) \\ & + (2y^3x^2-2yx^4)+(y^4x-x^5)+(9y^2x^2u-x^2u^3-3y^2x^2u^2)+(7y^3xu-2y^2u^3-3yxu^3) \\ (19) \quad & + (2y^4u-y^2xu^2)+(y^3x^3u+2yx^4u+yx^3u^2+x^4u+5yx^3u+x^3u^2)\}. \end{aligned}$$

The terms in parentheses are all visibly positive, and so $\delta(N)$ cannot vanish, a contradiction.

This completes the proof of Theorem 3 and therefore of Theorem 1. \square

Acknowledgments. Some (but not all!) of the extensive algebraic manipulations needed for the proof were carried out using Altran [3] and Macsyma [12]. We are very grateful to the M.I.T. Laboratory for Computer Science for allowing us to use Macsyma. E.S.B. also wishes to thank the Mathematics Department of Ohio State University at Columbus and Bell Laboratories for their hospitality and support during the course of this work.

REFERENCES

- [1] E. S. BARNES, *The covering of space by spheres*, *Canad. J. Math.*, 8 (1956), pp. 293–304.
- [2] E. S. BARNES AND T. J. DICKSON, *Extreme coverings of n-space by spheres*, *J. Austral. Math. Soc.*, 7 (1967), pp. 115–127; 8 (1968), pp. 638–640.
- [3] W. S. BROWN, *ALTRAN User's Manual*, Bell Laboratories, Murray Hill, NJ, 4th ed., 1977.
- [4] J. W. S. CASSELS, *An Introduction to the Geometry of Numbers*, Springer-Verlag, New York, 1971.
- [5] J. H. CONWAY AND N. J. A. SLOANE, *Voronoi regions of lattices, second moments of polytopes, and quantization*, *IEEE Trans. Inform. Theory*, IT-28 (1982), pp. 211–226.
- [6] ———, *Fast quantizing and decoding algorithms for lattice quantizers and codes*, *IEEE Trans. Inform. Theory*, IT-28 (1982), pp. 227–232.
- [7] ———, *Fast 4- and 8-dimensional quantizers and decoders*, *National Telecommunications Record-1981*, Vol. 3, IEEE Press, New York, pp. F4.2.1–F4.2.4.
- [8] N. C. GALLAGHER, JR., AND J. A. BUCKLEW, *Some recent developments in quantization theory*, *Proc. 12th Annual Symposium on System Theory*, Virginia Beach, VA, May 19–20, 1980.

- [9] A. GERSHO, *Asymptotically optimal block quantization*, IEEE Trans. Inform. Theory, IT-25 (1979), pp. 373–380.
- [10] ———, *On the structure of vector quantizers*, IEEE Trans. Inform. Theory, IT-28 (1982), pp. 157–166.
- [11] C. G. LEKKERKERKER, *Geometry of Numbers*, Wolters-Noordhoff, Groningen, and North-Holland, Amsterdam, 1969.
- [12] *MACSYMA Reference Manual*, Version 9, Mathlab Group, Laboratory for Computer Science, MIT, Cambridge, MA, 1977.
- [13] G. VORONOI, *Sur quelques propriétés des formes quadratiques positives parfaites*, J. Reine Angew. Math., 133 (1907), pp. 97–178.
- [14] ———, *Recherches sur les paralléloèdres primitifs* (Part 1), J. Reine Angew. Math., 134 (1908), pp. 198–287.
- [15] P. ZADOR, *Asymptotic quantization error of continuous signals and the quantization dimension*, IEEE Trans. Inform. Theory, IT-28 (1982), pp. 139–149.

AN OPTIMAL DIAGONAL TREE CODE*

S. CHAIKEN,† A. K. DEWDNEY‡ AND P. J. SLATER§

Abstract. It has been observed that not all the entries of $M(T)$, the distance matrix for a tree T , are necessary to determine T uniquely. For example, the submatrix of distances between end-vertices uniquely determines T . In this paper it is shown when T is canonically ordered, $2n - 3$ of the distances between the n end-vertices suffice for this determination. It is shown that $2n - 3$ is the minimum number of distances with this property.

Linear algorithms for finding such a set of distances (encoding), given the tree, and for finding a tree (decoding), given the distances, are described.

1. Introduction. The matrix M of intervertex distances of a graph has a number of interesting properties. As Smolenskii [6] and Zaretskii [8] observed, if the graph is a tree then the submatrix of M consisting only of the distances between end-vertices suffices to determine the tree uniquely. In [2] it was observed by one of us that if n is the number of end-vertices in a canonically ordered tree, then all but $2n$ of the n^2 entries can be removed from the latter matrix, and the ordered tree would still be uniquely determined. A natural question to ask is the following. What is the fewest number of entries from M upon which one can base a code to uniquely represent T ? Here it is shown that $2n - 3$ entries are sufficient and, in a sense, necessary.

In the next section, an optimal diagonal code for a tree is defined and proved to be minimal. In the following section we describe two algorithms. The first algorithm, called CODE, labels the end-vertices of an arbitrary tree T in a canonical way. The resulting labelling, by the integers $1, 2, \dots, n$ results in the set of distances

$$d(1, 2), d(2, 3), d(1, 3), d(3, 4), d(1, 4), \dots, d(i, i+1), d(1, i+1), \dots, \\ d(1, n-1), d(n-1, n), d(1, n).$$

These $(2n - 3)$ distances comprise a "diagonal" code [2] since they form one super-diagonal and the top row of the distance matrix based on this labelling. The second algorithm, called DECODE, converts a set of distances of this type into a tree. Both algorithms have linear worst-case time complexity.

In the concluding section of the paper, we compare the improved diagonal tree code with the other codes discussed in Read's survey article [5]. A case is made for the superiority in compactness of the improved diagonal code over all of these.

2. Definitions and theory. Let the end-vertices of a tree T be labelled with the integers $1, 2, \dots, n$, where n is the number of end-vertices. Where convenient, we shall refer to these vertices by their labels. Otherwise, these and other vertices will be given names such as u, v, w , and so on.

Any two vertices u and v of T are joined by a unique path, denoted by $P(u, v)$. The distance, $d(u, v)$, between u and v in T is the number of edges in $P(u, v)$. The matrix of distances

$$D(T) = (d_{ij}),$$

* Received by the editors January 31, 1979, and in revised form March 8, 1982. This research was supported in part by the National Science and Engineering Research Council of Canada grant no. A9271 and by the United States Department of Energy contract no. AT(29-1)-789.

† Department of Computer Science, State University of New York at Albany, Albany, New York 12222.

‡ Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B9.

§ Applied Mathematics Division 5641, Sandia Laboratories (A United States Department of Energy Facility), Albuquerque, New Mexico, 87185.

where $d_{ij} = d(i, j)$, $1 \leq i, j \leq n$, will be called the *distal matrix* to distinguish it from the (larger) *distance matrix* consisting of all the distances in T . Note that $d_{ij} = d_{ji}$ and $d_{ii} = 0$.

Given any three end-vertices i, j, k of a tree labelled as above, there is a single vertex which lies on all three paths $P(i, j)$, $P(i, k)$, $P(j, k)$. This vertex is called the *hub* of i, j, k . Given a single end-vertex i of T , there is, moreover, at most one vertex closest to i and having degree 3 or more. This vertex, if it exists, is called the *vertex of attachment of i* .

In Fig. 1 below is shown an end-labelled tree T and its distal matrix. Only the upper triangular portion of $D(T)$ is shown, the matrix being symmetric and the main diagonal consisting only of zeros.

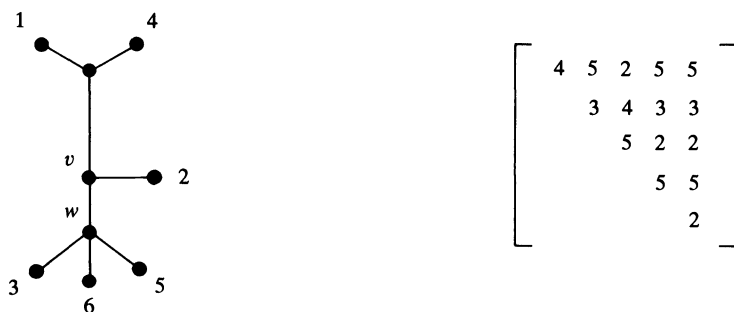


FIG. 1. An end-labelled tree and its distal matrix.

In [2], a *diagonal ordering* for the vertices $1, 2, \dots, n$ was defined as a permutation P_1, P_2, \dots, P_n of $1, 2, \dots, n$ such that the hub of each triple P_{i-1}, P_i, P_{i+1} is the vertex of attachment of P_i for $i = 1, 2, \dots, n$, subscripts being taken modulo n . Evidently, the ordering $1, 2, 3, 4, 5, 6$ above is not a diagonal ordering since the hub v of $2, 3, 4$ is not the vertex of attachment w of 3 . However, if the end-vertices are listed in the order

$$1, 3, 6, 5, 2, 4$$

then it is easy to check that this is, in fact, a diagonal ordering of the end-vertices of T . The diagonal code to which this ordering gives rise is

$$4, 2, 4, 2, 4, 3, 3, 3, 2.$$

As shown in [2], if a tree is embedded in the plane and its vertices numbered in, say, clockwise order, the resulting ordering is a diagonal ordering. In [2], however, the corresponding diagonal code had $2n$ entries. The one we are about to define has $2n - 3$ entries.

In either case, the resulting sequence of numbers is not a real "code" unless the correspondence between (nonisomorphic) trees and their codes is bi-unique. Thus, some provision is made for first embedding a given tree in the plane canonically, as was done in [2], and then selecting a starting vertex. Such a "starting point" is more critical to the formation of the optimal diagonal code discussed here since cyclic permutations of the resulting clockwise sequence may not even be a diagonal code at all! However, the process of canonically embedding a tree in the plane and of selecting a starting vertex is easily carried out and is described in the next section.

Let T be a plane tree with one of its end-vertices v designated as “1”. Numbering the vertices in clockwise fashion starting at v results in a diagonal ordering called the *clockwise numbering at v* and the sequence

$$d(1, 3), d(2, 3), d(1, 3), d(3, 4), d(1, 4), \dots, d(i, i+1), d(1, i+1), \dots, \\ d(1, n-1), d(n-1, n), d(1, n)$$

of $2n-3$ distances is called the *optimal diagonal code* of T . We now prove two theorems: the first shows that the optimal diagonal code corresponds to only one plane tree/vertex pair while the second theorem shows that fewer than $2n-3$ distances will not suffice to do this.

THEOREM 1. *If (T, v) and (T', v') are two plane tree/vertex pairs such that T and T' both have the same optimal diagonal code, then there is an isomorphism between T and T' which preserves the clockwise numbering.*

Proof. The proof is by induction on the number n of end-vertices in T . If $n=2$ then T is a path and its optimal diagonal code is a single, positive integer, so that T' must be a path of the same length and the theorem follows immediately. If $n>2$, remove the paths from n to its vertices of attachment x and x' in T and T' , respectively. The optimal diagonal code for the resulting trees \tilde{T} and \tilde{T}' is now

$$d(1, 2), d(2, 3), \dots, d(i, i+1), d(1, i+1), \dots, d(1, n-1)$$

and, by the induction hypothesis, there is an isomorphism between \tilde{T} and \tilde{T}' which preserves the clockwise numbering. The distances $d(1, x)$, $d(n-1, x)$ and $d(n, x)$ may be expressed as

$$d(1, x) = \frac{1}{2}(d(1, n-1) + d(1, n) - d(n-1, n)), \\ d(n-1, x) = \frac{1}{2}(d(n-1, n) + d(1, n-1) - d(1, n)), \\ d(n, x) = \frac{1}{2}(d(n-1, n) + d(1, n) - d(1, n-1)),$$

so that in each case the distances are uniquely determined by the integers in the code. Since the same conclusion holds for both T and T' , it follows that the isomorphism between \tilde{T} and \tilde{T}' maps x into x' and the isomorphism is extended to one between T and T' by re-attaching paths of length $d(n, x) = d(n, x')$ at x and x' . Clearly, the resulting isomorphism continues to preserve the diagonal labelling. This completes the proof of the theorem.

THEOREM 2. *Let K be a fixed set of k pairs of subscripts from among $\{1, 2, \dots, n\}$. If $k < 2n-3$, then there are two plane tree/vertex pairs (T, v) and (T', v') such that if T and T' are given a clockwise numbering at v and v' , then*

- (a) *The distances $\{d(i, j) : (i, j) \in K\}$ will be the same in both T and T' , but*
- (b) *T and T' will be nonisomorphic.*

Proof. Let \tilde{T} be an arbitrarily chosen plane tree with n vertices of degree 1, the remaining vertices having degree 3. It is not hard to see that \tilde{T} has $2n-3$ edges. Let these be numbered $1, 2, \dots, 2n-3$ in some arbitrary but fixed manner and, selecting an end-vertex v of \tilde{T} , let \tilde{T} be numbered clockwise at v . It is now possible to form the $\binom{2n-3}{2} \times (2n-3)$ matrix $P(\tilde{T})$ whose ij th entry is a 1 if the i th distance among all $\binom{2n-3}{2}$ possible end-vertex distances includes the j th edge in its corresponding path. The rows of $P(\tilde{T})$ may be regarded as $(2n-3)$ -vectors over \mathbb{Z} . They generate a space \mathcal{M} of dimension $2n-3$ since, in fact, the vectors which correspond to the distances selected for the optimal diagonal code for (\tilde{T}, v) are linearly independent. On the other hand, the vectors corresponding to the set K generate a subspace of dimension less than

$2n - 3$, whence there exists a nonzero vector in \mathcal{M} with integer components that is orthogonal to the k vectors in K . Let this vector be $R = (r_1, r_2, \dots, r_{2n-3})$ and define $r = \max \{r_i : i = 1, 2, \dots, 2n - 3\}$. We may now construct the two trees T and T' of the theorem from \tilde{T} as follows. The tree T is obtained from \tilde{T} by replacing each edge by a path of length r . The tree T' is obtained from \tilde{T} by replacing the i th edge by a path of length $r + r_i$, ($i = 1, 2, \dots, 2n - 3$).

Let (i, j) be an arbitrary pair in K and suppose that the row vector of $P(\tilde{T})$ corresponding to this pair is $S = (s_1, s_2, \dots, s_{2n-3})$. Then the distance $d(i, j)$ from i to j in T is given by the expression

$$\sum_{l=1}^{2n-3} r s_l.$$

On the other hand, the distance $d'(i, j)$ from i to j in T' is given by

$$d'(i, j) = \sum_{l=1}^{2n-3} (r + r_l) s_l = \sum_{l=1}^{2n-3} r s_l + \sum_{l=1}^{2n-3} r_l s_l.$$

Since R is orthogonal to S , however, the second sum equals 0. It follows that $d'(i, j) = d(i, j)$ for every pair (i, j) in K . Since T and T' do not have the same numbers of vertices, they cannot be isomorphic. This establishes the result.

It can now be seen in precisely what sense the optimal diagonal code may be regarded as "optimal"; if one starts with a plane tree/vertex pair, it has been shown that one cannot recover the tree uniquely from fewer than $2n - 3$ of its end-vertex distances. Our argument also establishes that at least $2n - 3$ end-vertex distances are needed to recover uniquely an arbitrary (nonplane) tree. However, it can be shown that all $\binom{2n-3}{2}$ end-vertex distances are required to recover an arbitrary tree. This point will be taken up in a future publication.

3. Coding and decoding algorithms. The first algorithm discussed in this section is called CODE and transforms an arbitrary (nonplane) tree into an optimal diagonal code which is clearly unique to the tree. The algorithm thus amounts to a definition of the *optimal diagonal code* of a tree. It consists of two subalgorithms. The first of these is an efficient adaptation of Read's algorithm [5] which embeds a tree canonically in the plane. This embedding results automatically in a "leftmost" end-vertex which, in the second subalgorithm, receives the number "1".

Let a tree T be given by a system of linked lists. The formation of a canonically embedded plane version of T involves two steps (1 and 2 below), described in only enough detail to enable an interested reader to write the appropriate algorithm after reading [1, p. 176 ff.] using these steps as a guide. In what follows T has m vertices; n of these are end-vertices.

1. *Finding the center of T .* The center of T consists of those vertices the maximum of whose distances from end-vertices is a minimum. A well-known result [3, p. 35] asserts that the center of T consists of a single vertex or two adjacent vertices. A depth-first search of T (on m vertices) can be accomplished in time $O(m)$ [1, p. 178] and such a search pattern enables one to find the center of T by keeping track of the maximum distance to an end-vertex at each internal vertex. Every time the depth-first search backs up to a vertex v for the last time, the maximum of the distances along each branch is taken as the distance associated with the branch defined by vertex v at the next "higher" vertex. When this process is complete, all maximum distances outward from the starting vertex u will be known. The center of T will lie on the branch of maximum distance at u . Moving along this branch from u , two things are

done at each vertex v encountered: a) the maximum distance d along the branch at v containing u is calculated, and b) this distance is compared with the maximum distance along the other branches at v . If the two maxima are equal, then the center of the tree consists of v alone and the algorithm halts. If the two are not equal then continue to the next vertex v' lying on the branch of maximum distance at v . If v is revisited following an examination of v' , then the center of the tree consists of v and v' together and the algorithm halts. Clearly, it requires at most $O(m)$ additional steps to locate the center from u .

2. *Embedding the subtrees at the center of T .* If T has a one-vertex center at v , say, then T may be written as the union of subtrees, one defined by each vertex adjacent to v . If T has a two-vertex center, then T may be written as the union of two subtrees, one defined by each vertex in the center. These possibilities are illustrated in Fig. 2.

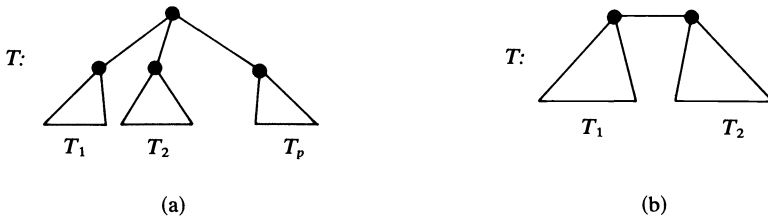


FIG. 2. The subtrees at (a) a one-vertex center and (b) a two-vertex center.

A depth-first search is again carried out on T , this time starting at the top node of each of the subtrees defined above. The *binary code* of a rooted tree (T', v) is defined recursively as the binary code $0C_1C_2 \cdots C_p1$, where the subtree codes C_i of the subtrees T'_i of T' at v are taken in nondescending order as binary integers. It is easy to compute such a code as the tree is being searched. At each vertex, however, information about the ordering of the branches imposed by the coding convention must be stored. This ordering determines the plane embedding of T . Thus, when the binary code of each subtree $T' = T_i$ has been computed, the subtrees T_i are themselves arranged in nondecreasing code order and the canonical embedding procedure of T itself is complete. Depth-first construction and lexicographic sort [1, p. 77 ff.] can ensure a completion time of $O(m)$ steps.

The second subalgorithm of CODE mentioned at the beginning of this section takes the canonical plane embedding of T obtained by the first subalgorithm, i.e., steps 1 and 2 above, and computes its optimal diagonal code.

3. *The optimal diagonal code of T .* Once again T is given a depth-first search but this time the order of visitation is determined by the left-to-right embedding order of the plane version of T produced by the first subalgorithm. As the depth-first search proceeds, d_1 , the distance from vertex 1, and d_2 , the distance from the last end-vertex, are kept track of. When the search begins at vertex 1, d_1 and d_2 are initialized to zero. Whenever an edge is traversed in the forward direction, d_1 and d_2 are incremented. Whenever an edge is traversed in the reverse direction, that is, when the search backs up, d_1 is decremented and d_2 is incremented. When end-vertex 2 is reached, d_1 is output and d_2 is set to zero. Whenever an end-vertex after vertex 2 is reached, d_2 and d_1 in order are output and then d_2 is set to zero. The search may terminate after end-vertex n is processed.

It is easy to verify by induction that whenever an edge is traversed, distances d_1 and d_2 are updated correctly. Notice that the transversal direction reverses each time an end-vertex is encountered and each time the hub of vertex 1, an end-vertex v , and the next end-vertex $v + 1$ is encountered.

As each new end-vertex is visited, two integers are outputted. The only exceptions to this involve the first end-vertex, where no integers are produced, and the second end-vertex, where only one is produced. The sequence of integers outputted by the subalgorithm is obviously the optimal diagonal code of the plane version of T . Moreover, since this subalgorithm, like the others, is based on depth-first search, the number of steps required for its completion is $O(m)$.

The second, main algorithm discussed in this section is called DECODE and transforms a valid optimal diagonal code into a tree represented by a linked list structure. This algorithm is based on our proof of Theorem 1. Although no attempt is made here to distinguish between valid and invalid optimal diagonal codes, i.e., between sequences of $2n - 3$ integers which are or are not optimal diagonal codes, such a study might well be based on the results of Patrinos and Hakimi [4] characterizing which distance matrices arise from trees.

If $d_{12}, d_{23}, d_{13}, d_{34}, d_{14}, \dots, d_{1n-1}, d_{n-1,n}, d_{1,n}$ is a valid optimal diagonal code, it is not difficult to see that the reduced sequence $d_{12}, d_{23}, d_{13}, d_{34}, \dots, d_{1,n-1}$ is also a valid optimal diagonal code. The algorithm DECODE is described inductively on the basis of this observation. First, if $n = 2$, the tree corresponding to the given code is easily constructed as a path of length d_{12} . Suppose now that the algorithm is able to construct a tree T' whose code has length $2n - 5$. Given the code for a tree with n end-vertices, remove the last two integers of the code and run DECODE on the resulting sequence, obtaining the tree T_{n-1} with $n - 1$ end-vertices. In order to obtain T_n , the tree corresponding to the given code, a path is attached to T_{n-1} . In order to do this, however, it is necessary to know two things:

- a) the length of path to be attached,
- b) the vertex in T_{n-1} where the path is to be attached.

Since the code $d_{12}, d_{23}, d_{13}, d_{34}, \dots, d_{1,n-1}$ is valid, let T_{n-1} be the unique tree for which it is the code. Let x be the vertex of attachment for the vertex n in T_{n-1} . Observe that the quantity sought in a) is $d(x, n)$ while x is actually located in T_{n-1} a distance $d(x, n - 1)$ from $n - 1$ upon $P(1, n - 1)$. The quantities $d(x, n)$ and $d(x, n - 1)$ are easily found to be

$$d(x, n) = \frac{1}{2}(d_{n-1,n} + d_{1,n} - d_{1,n-1}),$$

$$d(x, n - 1) = \frac{1}{2}(d_{n-1,n} + d_{1,n-1} - d_{1,n}).$$

The steps involved in the construction of T_n from T_{n-1} include the $d(x, n)$ steps in the construction of the linked list for $P(x, n)$ and the $d(x, n - 1)$ steps in the search for x along $P(n - 1, 1)$. Adding the two quantities we obtain $d_{n-1,n}$.

It follows that the total number of steps involved in the reconstruction of T from its optimal diagonal code is

$$O\left(\sum_{k=1}^{n-1} d_{k,k+1}\right).$$

The expression in parentheses counts every edge twice, so that the resulting complexity of DECODE is $O(m)$. It is easily proved by induction that DECODE works properly by using the same observation which inspired this algorithm, along with the ever useful distance equations above.

The complexity of our coding and decoding algorithms is $O(m)$ where m is the number of vertices in the tree, whereas the length of the code is $O(n)$. (The code length is $O(n \log m)$ if we count the digits to write the distances.) It is easy to modify our algorithms to use a tree representation in which each path of degree 2 vertices is represented by a single, integer-weighted edge. This way one can have trees, codes, and algorithms whose complexities, for practical purposes, are all linear in n , the number of end-vertices.

4. Summary and conclusions. There is an encoding of ordered or plane trees with n end-vertices by a sequence of just $2n - 3$ integers. This code, called here the “optimal diagonal code” of a tree has been defined and shown to be optimal in a certain sense. Furthermore, linear algorithms for converting a tree to its optimal diagonal code and back again have been described.

In [2], one of us described a “diagonal code” involving $2n$ distances. We note here in passing that the code is invalid as described; if, however, one labels the end-vertices (say clockwise) in the order $1, 3, 5, \dots, n, \dots, 6, 4, 2$ and then uses the same set of distances, the code is valid, and the claims made for it are correct. As such, the only serious competition for this code was shown in [2] to be the WAV (walk around valency) code as described by Read in [5]. It is obtained by visiting the m vertices of a canonically embedded plane tree in a prescribed order and recording their degrees as they are visited. The result is a code consisting of $m - 1$ numbers.

We can summarize the comparison of the “diagonal code” of [2] with the WAV code by stating that the diagonal code showed evidence of superiority (in terms of the proportion of trees with m vertices for which the diagonal code would be shorter) out to $m = 13$. Further comparison appears to require enumeration of the number of trees on m vertices with N end-vertices for $2 \leq N \leq m - 1$. In comparison with the WAV code the improved diagonal code described here must do even better.

The relative sizes of the numbers in the diagonal and WAV code have been neglected. We only point out that

- (a) they have the same worst-case order of magnitude, and
- (b) their average values tend to be reciprocally related: the greater the average degree of an m -vertex tree, the shorter the distance between end-vertices.

We conclude with an observation about labelled trees on m vertices as m gets large. As noted in [7, p. 52], for large m the probability of a given vertex being an end-vertex is approximately e^{-1} . Thus the expected number of endpoints would be m/e . Our diagonal code would thus have length $(2m/e) - 3$ which is less than three-fourths of the length $m - 1$ of the WAV code.

The results described raise at least two questions, one rather specific, the other general. The first question concerns the optimality of our code. We have shown that $2n - 3$ distances are necessary and sufficient for specifying a plane tree in any scheme in which the set of end-vertex pairs is fixed. Although we have referred to “the optimal diagonal code” here, it should be mentioned that a variety of other codes of this length exist. All appear, however, to have a superdiagonal as a subset of their distances. It is possible that there are distance coding schemes in which some but not all n endpoint trees are specified with fewer than $2n - 3$ distances. For example, if it is known that a tree has only one vertex with degree greater than 2, and the number of end-vertices n is odd, the n distances $d_{12}, d_{23}, \dots, d_{n1}$ suffice to specify the tree. The invention of such a scheme would combine the advantages of our scheme with those of the WAV code. The second question involves graphs other than trees: given a class of graphs, what subsets of their distance matrix entries uniquely determine them? If so, what is the minimum number of matrix entries sufficient for this purpose?

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] A. K. DEWDNEY, *Diagonal tree codes*, Information and Control, 40 (1979), pp. 234–239.
- [3] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.
- [4] A. N. PATRINOS AND S. L. HAKIMI, *The distance matrix of a graph and its tree realization*, Quart. Appl. Math., 30 (1972), pp. 255–269.
- [5] R. C. READ, *The coding of various kinds of unlabelled trees*, Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1972.
- [6] Y. A. SMOLENSKII, *A method for linear recording of graphs*, USSR Comput. Math. and Math. Phys., 2 (1963), pp. 396–397.
- [7] R. WILSON, *Introduction to Graph Theory*, Academic Press, New York and London, 1972.
- [8] K. ZARETSKII, *Constructing a tree on the basis of a set of distances between the hanging vertices*, Uspekhi Mat. Nauk, 20 (1965), pp. 90–92.

RECTANGULAR MATRICES AND SIGNED GRAPHS*

HARVEY J. GREENBERG[†], J. RICHARD LUNDGREN[‡] AND JOHN S. MAYBEE[§]

Abstract. This paper extends the theory of graphs associated with real rectangular matrices to include information about the signs of the elements. We show when signed row and column graphs can be defined for the matrix A . We also deduce conditions under which these graphs are balanced. This leads to a definition of the class of quasi-Morishima rectangular matrices A . It is shown that the Perron-Frobenius theorem applies to the matrices AA^T and A^TA when A is a quasi-Morishima matrix. Finally we examine the applications of our results to several classes of matrices occurring in energy economic models. All results in this paper are purely qualitative in character.

1. Introduction. This paper continues the development of the theory and use of graphs and digraphs associated with rectangular matrices which we initiated in [1] and [2]. Our aim here is to construct a theory that will adequately exploit the sign information contained in real matrices. We touched briefly upon this topic in the previous papers, but no systematic presentation was attempted. It turns out that, under certain conditions, a satisfactory theory of positivity can be devised for rectangular real matrices.

Our work in this paper has been motivated by our efforts to study the important properties of two special classes of matrices introduced by H. J. Greenberg in [3]. Greenberg has identified physical flows matrices (PFM) and physical flows with feedback matrices (PFFM) as important components of energy economic models.

Before introducing the PFM and PFFM we remind the reader how the basic graphs associated with the $m \times n$ matrix A are defined.

Given the $m \times n$ matrix A we define two sets of points, $R = \{r_1, \dots, r_m\}$ and $C = \{c_1, \dots, c_n\}$, to represent the rows and columns of A , respectively. We then have the following definitions.

Fundamental bipartite graph (bigraph): BG is a bigraph on the point sets R and C . The line $[r_i, c_j]$ belongs to BG if $a_{ij} \neq 0$.

Row graph. RG is defined on R . The line $[r_i, r_j]$ belongs to RG if there exists $c_k \in C$ such that $[r_i, c_k]$ and $[r_j, c_k]$ are in BG . Thus two rows are adjacent if they have a common column intersection in A .

Column graph. CG is defined on C . The line $[c_i, c_k]$ belongs to CG if there exists $r_k \in R$ such that $[c_i, r_k]$ and $[c_j, r_k]$ belong to BG . In other words, two columns are adjacent if they have a common row intersection in A .

A rectangular matrix A will be called *regular* if each row and column of A contains at least one nonzero element.

Now it is clear that the sign information in the real matrix A can be immediately incorporated into the bigraph BG . In fact, we label the line $[r_i, c_j]$ positive if $a_{ij} > 0$ and negative if $a_{ij} < 0$. The resulting signed graph will be denoted by BG^+ .

The usefulness of signed graphs and digraphs has been demonstrated by several authors (see Harary [4], [5], Maybee and Quirk [6], and Roberts [7], for example). Let us therefore show that for PFM and PFFM we can define the signed graphs RG^+ and CG^+ .

* Received by the editors April 21, 1981, and in revised form March 16, 1982.

[†] Energy Information Administration, Washington, D.C. 20461.

[‡] University of Colorado, Denver, Colorado 80202. The research of this author was supported by the National Science Foundation under grant SPI-7916608 while he was visiting the University of Colorado.

[§] University of Colorado, Boulder, Colorado 80309.

Physical flows matrix. The matrix A is a PFM if the rows can be partitioned into disjoint sets S and M and the columns into disjoint sets P , T and K such that:

- (1) Every element a_{ij} with $i \in S, j \in P$ is nonnegative.
- (2) Every element a_{ij} with $i \in M, j \in K$ is nonpositive.
- (3) Every element a_{ij} with $i \in S, j \in T$ is nonpositive.
- (4) Every element a_{ij} with $i \in M, j \in T$ is nonnegative.
- (5) All other elements of A are zero.

It is therefore true that A is a PFM if and only if there exist permutation matrices P and Q such that

$$(1.1) \quad PAQ = \begin{bmatrix} A_{11} & A_{12} & 0 \\ 0 & A_{22} & A_{23} \end{bmatrix},$$

where $A_{11} \geq 0, A_{12} \leq 0, A_{22} \geq 0, A_{23} \leq 0$.

The matrix A in PFM will be called regular if each of the blocks A_{11}, A_{12}, A_{22} and A_{23} is nonempty and regular.

Physical flows with feedback matrix. The matrix A is a PFFM if the rows can be partitioned into disjoint sets I, S, M and the columns into disjoint sets P, T, K such that (1) through (5) hold and:

- (6) Every a_{ij} with $i \in I, j \in P$ is nonpositive.
- (7) Every a_{ij} with $i \in I, j \in K$ is nonnegative.
- (8) Every a_{ij} with $i \in I, j \in T$ is zero.

Thus A is a PFFM if and only if there exist permutation matrices P and Q such that

$$(1.2) \quad PAQ = \begin{bmatrix} A_{11} & 0 & A_{13} \\ A_{21} & A_{22} & 0 \\ 0 & A_{32} & A_{33} \end{bmatrix},$$

where $A_{11} \leq 0, A_{13} \geq 0, A_{21} \geq 0, A_{22} \leq 0, A_{32} \geq 0, A_{33} \leq 0$.

The matrix A will be called a regular PFFM if each of the blocks $A_{11}, A_{13}, A_{21}, A_{22}, A_{32}, A_{33}$ is nonempty and regular.

Now let us observe that, when A is a PFM or a PFFM the scalar product of any two columns is positive, negative or zero independently of the magnitudes of the elements because all terms in the scalar product are weakly of the same sign. The same is true for the scalar product of two row vectors. Consequently to such matrices we can associate the signed graphs CG^+ and RG^+ in which the line $[c_i, c_j]$ ($[r_i, r_j]$) is positive if the corresponding column (row) vectors have a positive scalar product and negative if the scalar product is negative. The smallest regular PFM is shown in Fig. 1 together with CG^+ and RG^+ . The smallest regular PFFM is illustrated in Fig. 2. In drawing the graphs we have followed the convention of using dashed lines to represent negative lines as introduced in [8].

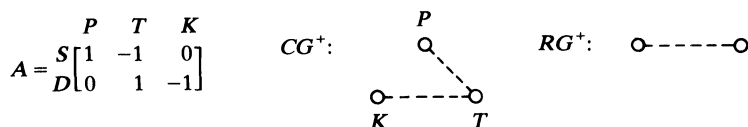


FIG. 1. Smallest PFM.

Since the graphs CG^+ and RG^+ can be defined for PFM and PFFM, it seems natural to seek to determine the class of real rectangular matrices for which these graphs, or at least one of them, can be defined. Section 2 is devoted to the determination of this class of matrices and to some of their properties.

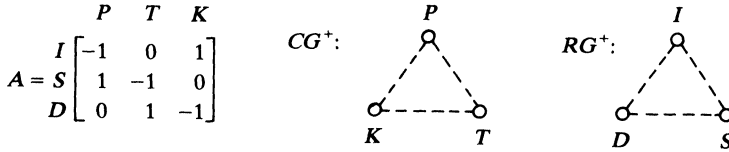


FIG. 2. Smallest PPFM.

Now it turns out that among the matrices for which the signed graphs CG^+ and RG^+ exist there is a subclass with the property that these graphs are balanced (all cycles positive). This subclass includes the PFM but not the PPFM. For such matrices we can develop a satisfactory theory of positivity and we can apply the Perron-Frobenius and its corollaries to the matrices AA^T and $A^T A$. This is the subject matter of § 3.

Finally § 4 is devoted to applications of our results to the classes PFM and PPFM and to certain generalizations of these classes.

All of our results are purely qualitative in character, i.e., they hold regardless of the magnitudes of the matrix elements.

2. Signed matrices. We begin with an embellishment of some fundamental ideas introduced in the paper [6] (see also [9] where similar concepts are used). Let $x = (x_1, \dots, x_N)$ and $y = (y_1, \dots, y_N)$ be vectors in the euclidean space \mathcal{R}^N . We shall call x and y *conformal* if $x_i y_i \geq 0, 1 \leq i \leq N$, and *anticonformal* if $x_i y_i \leq 0, 1 \leq i \leq N$. In the terminology of [6] x and y are conformal if they lie in the closure of the same qualitative cone Q_x in \mathcal{R}^N and anticonformal if one vector lies in a closed qualitative cone and the other in the corresponding negative closed cone.

Let $A = [a_{ij}]$ be an $m \times n$ real matrix. We will call A *row signed* if the row vectors of A regarded as elements of \mathcal{R}^n all lie in the same closed qualitative cone or in its negative, i.e., if they are pairwise either conformal or anticonformal. We define A to be *column signed* if the column vectors of A regarded as elements of \mathcal{R}^m are pairwise either conformal or anticonformal.

LEMMA 1. A is column signed if and only if A is row signed.

Proof. Assume A is column signed. Suppose A is not row signed. Then there exist rows r_s and r_t such that for some i and $j, a_{si} a_{ti} > 0$ and $a_{sj} a_{tj} < 0$. Now consider the products $a_{si} a_{sj}$ and $a_{ti} a_{tj}$. We have $a_{si} a_{sj} a_{ti} a_{tj} < 0$ so these products have different signs. Therefore A is not column signed, a contradiction. It follows that A must be row signed. The proof is similar if we assume A is row signed.

In view of Lemma 1 we shall say henceforth that A is signed without using the adjectives column or row.

The following lemma is a complement to Lemma 1.

LEMMA 2. Let G^+ be a signed graph with n points. Then there exists a matrix A such that $CG^+(A) = G^+$.

Proof. Let e_1, \dots, e_m be the lines of G^+ and p_1, \dots, p_n the points. Construct A as follows:

Column j of A corresponds to point p_j of G^+ .

If $e_i = [c_{i1}, c_{i2}]$, then $a_{ii_1} = 1$ and

$$a_{ii_2} = \begin{cases} 1 & \text{if } e_i \text{ is positive,} \\ -1 & \text{if } e_i \text{ is negative.} \end{cases}$$

Observe that each row of A has only two nonzero elements and each row corresponds to a unique line of $CG^+(A)$. Thus $CG^+(A) = G^+$.

The next result relates the property that A is signed to properties of AA^T and $A^T A$. We require first some preliminary ideas. Recall from [6] that, if $x \in \mathcal{R}^N$, $\text{sgn } x = (\text{sgn } x_1, \dots, \text{sgn } x_N)$ where sgn is the usual signum function. In the same way we can associate with any real $m \times n$ matrix A the matrix $\text{sgn } A = [\text{sgn } a_{ij}]$ (see [6] for more detail). Let us introduce the addition table

+	-1	0	1
-1	-1	-1	x
0	-1	0	1
1	x	1	1

in which we use an x to denote an indeterminate (as to sign) entry. Define $A^T \circ A$ to be the matrix product formed using this addition table.

THEOREM 1. *Let A be an $m \times n$ matrix with at least one nonzero element in each row and column. The following are equivalent:*

- (1) A is signed.
- (2) $\text{sgn } (AA^T) = \text{sgn } A \circ (\text{sgn } A)^T$.
- (3) $\text{sgn } (A^T A) = (\text{sgn } A)^T \circ \text{sgn } A$.

Proof. The proof is left to the reader.

The interpretation of (2) and (3) is that the left-hand side is defined if and only if the right-hand side is. Since $\text{sgn } A \circ (\text{sgn } A)^T$ is symmetric and each row has a nonzero element, the diagonal elements are all positive and we need only calculate the elements above the principal diagonal. If any of these elements equals x ; then A is not signed. Otherwise A is signed.

For very large matrices Theorem 1 may not provide a useful test of whether or not A is signed, especially if AA^T or $A^T A$ is not sparse. The next result provides another criterion.

THEOREM 2. *A is signed if and only if every 4-cycle of BG^+ is positive.*

3. Positivity. For the deeper study of signed matrices we will require some background. First we recall that a signed graph is called *balanced* if every cycle is positive. Secondly the signed graph G^+ is balanced if and only if the points of G^+ can be partitioned into disjoint subsets S_1 and S_2 (one of which may be empty) such that every line joining two points in the same set is positive and every line joining points in different sets is negative (see Harary [4] for further details).

Let $A = [a_{ij}]_1^n$ be a square sign-symmetric matrix. Then we associate with A a graph $G(A)$ as follows: G has n points labelled $1, 2, \dots, n$ and a line joining points i and j ($i \neq j$) if a_{ij} (and a_{ji}) $\neq 0$. The line $[ij]$ will be given a positive label if $a_{ij} > 0$ and a negative label if $a_{ij} < 0$. In this way we arrive at the signed graph $G^+(A)$. The matrix A is called a *Morishima matrix* if $G^+(A)$ is balanced. Moreover, the square matrix A is a Morishima matrix if and only if there exists a permutation matrix \mathcal{P} such that

$$\mathcal{P}^T A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \geq 0$, $A_{22} \geq 0$, $A_{12} \leq 0$, $A_{21} \leq 0$ and where each of the blocks A_{11} and A_{22} is square. In this representation the block A_{22} may be 0×0 in which case A itself is nonnegative which is a special case of the Morishima class.

With this material as background we can proceed. First note that, if BG^+ is balanced, it has nonnegative 4-cycles and so RG^+ and CG^+ are defined.

THEOREM 3. *The following are equivalent:*

- (a) BG^+ is balanced.
- (b) CG^+ is balanced.
- (c) RG^+ is balanced.

Proof. We will only establish the equivalence of (a) and (b), the proof that (a) is equivalent to (c) being similar. Suppose BG^+ is balanced. Then the points of BG^+ can be partitioned into two disjoint subsets B_1 and B_2 such that positive lines join points of the same subset and negative lines join points of different subsets. Define $C_1 = B_1 \cap C$ and $C_2 = B_2 \cap C$.

Suppose first that c_s and c_t in C_1 are joined in CG^+ . Then there exists a row point r_i such that $[c_s, r_i]$ and $[c_t, r_i]$ are lines in BG^+ . If $r_i \in B_1$ both lines are positive and if $r_i \in B_2$ both are negative. In the latter case $a_{is} < 0$ and $a_{it} < 0$ so $[c_s, c_t]$ is positive. A similar argument works if both points are in C_2 .

Now suppose $c_s \in C_1$, $c_t \in C_2$ and $[c_s, c_t] \in CG^+$. Then there are lines $[c_s, r_i]$ and $[c_t, r_i]$ in BG^+ . If $r_i \in B_1$, then $[c_s, r_i]$ is positive, $[c_t, r_i]$ is negative and it follows that $[c_s, c_t]$ is negative. A similar result holds if $r_i \in B_2$.

Thus we have shown that CG^+ is balanced and so (a) implies (b).

Assume next that CG^+ is balanced so that the points in CG^+ can be partitioned into disjoint subsets C_1 and C_2 such that positive lines join points of the same subset and negative lines join points of different subsets. We construct disjoint subsets B_1 and B_2 of BG^+ as follows:

$$\tilde{R} = \{r_i \in R : [r_i, c] \text{ is positive for some } c \in C_1\},$$

$$\hat{R} = \{r_i \in R : [r_i, c] \text{ is negative for some } c \in C_2 \text{ and } r_i \text{ is not adjacent to any } c \in C_1\},$$

$$R_1 = \tilde{R} \cup \hat{R}, \quad R_2 = R - R_1,$$

$$B_1 = R_1 \cup C_1, \quad B_2 = R_2 \cup C_2.$$

It is clear that $B_1 \cap B_2 = \emptyset$ and $B_1 \cup B_2 = R_1 \cup C_1 \cup R_2 \cup C_2 = R \cup C$.

First we show that all lines joining points in B_1 are positive. Suppose $[r_i, c_s]$ is negative for some $r_i, c_s \in B_1$. Then $r_i \notin \tilde{R}$, so $r_i \in \hat{R}$ and there is a point $c_t \in C_1$ such that $[c_t, r_i]$ is positive. But then $a_{is} < 0$ and $a_{it} > 0$ so that $[c_t, c_s]$ is negative in CG^+ , a contradiction. Next we show that all lines between points in B_2 are positive. Suppose $[r_i, c_s]$ is negative for some $r_i, c_s \in B_2$. Since $r_i \notin \hat{R}$, then there is a point $c_t \in C_1$ adjacent to r_i and $[r_i, c_t]$ is negative. But then $a_{is} < 0$ and $a_{it} < 0$, so that $[c_t, c_s]$ is a positive line in CG^+ , a contradiction.

Finally we show that all lines between points in B_1 and B_2 are negative. Suppose there is a line from a point in R_2 to a point in C_1 . Then this line must be negative by definition. Suppose there is a line from a point $r_i \in R_1$ to a point $c_j \in C_2$. If $r_i \in \tilde{R}$, then $[r_i, c_s]$ is a line for some $c_s \in C_1$. Then $[r_i, c_s]$ is positive and $[c_s, c_j]$ is a line in CG^+ so it must be negative. Thus, since $a_{is} > 0$, we have $a_{ij} < 0$ so that $[r_i, c_j]$ is negative. If $r_i \in \hat{R}$, then $[r_i, c_t]$ is negative for some $c_t \in C_2$. But this means c_t and c_j are adjacent in CG^+ , so $[c_t, c_j]$ must be positive and hence $[r_i, c_j]$ must be negative.

We have thus shown that B_1 and B_2 satisfy the conditions for BG^+ to be balanced, so (b) implies (a).

Theorem 3 adds to the list of properties shared by the various graphs associated with a regular rectangular matrix. We now connect the form of the matrix A to balance.

THEOREM 4. BG^+ is balanced if and only if there exist permutation matrices P and Q such that

$$PAQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \geq 0, A_{22} \geq 0, A_{12} \leq 0, A_{21} \leq 0$.

Proof. Suppose BG^+ is balanced. Then the points of B can be partitioned into disjoint subsets B_1 and B_2 (one of which may be empty) such that positive lines join points of the same subset and negative lines join points of different subsets. Furthermore, $B_1 = R_1 \cup C_1$ and $B_2 = R_2 \cup C_2$. If one of the sets, say $B_2 = \emptyset$, then $A \geq 0$ and the result is trivial. In the remaining cases there exist permutation matrices P and Q such that

$$PAQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

From the restrictions on the lines in BG^+ we clearly have that $A_{11} > 0, A_{12} < 0, A_{21} < 0, A_{22} > 0$. Since any of the sets R_i or $C_i, 1 < i < 2$ may be empty, PAQ may contain one block, two blocks or all four blocks. For the converse suppose there exist permutation matrices P and Q such that PAQ has the above form. Then choosing R_i and C_i as before and setting $B_1 = R_1 \cup C_1, B_2 = R_2 \cup C_2$ we see that BG^+ is balanced.

The following simple lemma shows that whenever CG^+ and RG^+ are defined, certain cliques in these graphs must be balanced.

LEMMA 3. *If a k -clique in $CG^+(RG^+)$ arises from the k nonzeros in a row (column) of A , then the k -clique is balanced.*

Proof. Suppose a k -clique in CG^+ is determined by k nonzeros in a row. Let C_1 be the set of columns having a positive entry in the row and C_2 the set of columns having a negative entry. Clearly every line joining points in the same set is positive and every line joining points in different sets is negative. Therefore the k -clique is balanced.

We caution the reader that the lemma does not imply that every clique of $CG^+(RG^+)$ is balanced; it only identifies the existence of a spanning set of balanced cliques. To show this consider the PFFM A with

$$\text{sgn } A = \begin{bmatrix} - & - & 0 & 0 & 0 & + & + \\ 0 & - & - & 0 & 0 & + & 0 \\ + & 0 & + & - & - & 0 & 0 \\ + & + & + & 0 & - & 0 & 0 \\ 0 & 0 & 0 & + & + & - & 0 \\ 0 & 0 & 0 & + & + & 0 & - \end{bmatrix}.$$

The graph RG^+ is shown in Fig. 3. The cliques $\langle 1, 4, 5 \rangle, \langle 1, 2, 4, 5 \rangle$ and $\langle 2, 4, 5 \rangle$, for example, are not balanced. Yet the spanning set $S = \{\langle 1, 3, 4 \rangle, \langle 1, 2, 4 \rangle, \langle 2, 3, 4 \rangle,$

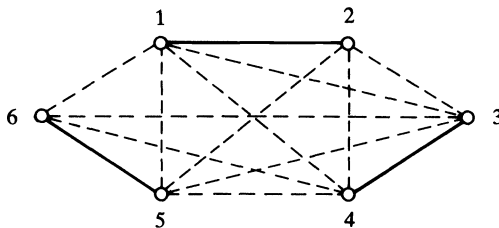


FIG. 3

$\langle 3, 5, 6 \rangle, \langle 3, 4, 5, 6 \rangle, \langle 1, 2, 5 \rangle, \langle 1, 6 \rangle$ consists of balanced cliques. (Here we are using the notation of Harary [11] for the subgraph generated by the set X , namely $\langle X \rangle$.)

In view of our results in Theorems 3 and 4 we are tempted to call the $m \times n$ matrix A a Morishima matrix if CG^+ is balanced. This will, however, introduce an inconsistency when $m = n$. As an example consider the 4×4 matrix A with sign patterns

$$\text{sgn } A = \begin{bmatrix} - & 0 & - & 0 \\ + & + & + & 0 \\ 0 & + & 0 & + \\ + & 0 & + & + \end{bmatrix}.$$

The graphs CG^+, RG^+ and $\tilde{D}^+(A)$, the signed directed graph of A with loops omitted, are illustrated in Fig. 4. Since the graph $\tilde{D}^+(A)$ has negative cycles, this matrix is not a Morishima matrix, yet both CG^+ and RG^+ are balanced. Therefore we introduce the following basic concept.

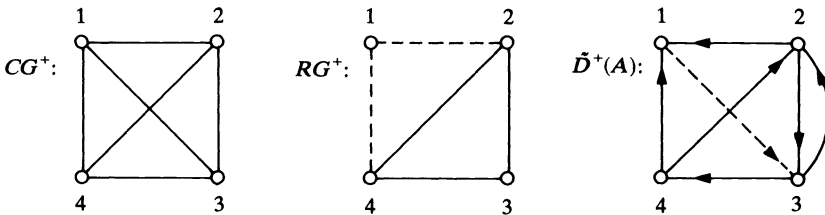


FIG. 4

DEFINITION 1. The $m \times n$ matrix A will be called a *quasi-Morishima matrix* if BG^+ is balanced.

The remainder of our results will be valid for all quasi-Morishima matrices including the case $m = n$.

THEOREM 5. Let A be a quasi-Morishima matrix. Then $A^T A$ and AA^T are (symmetric) Morishima matrices. If A has a nonzero element in each row and column then $A^T A$ and AA^T have positive diagonal elements.

Proof. The last statement of the theorem is trivial so we prove only the first statement. Since A is quasi-Morishima there exist permutation matrices P and Q such that

$$B = PAQ = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

$B_{11} \geq 0, B_{22} \geq 0, B_{12} \leq 0, B_{21} \leq 0$. Then $B^T B = (PAQ)^T PAQ = Q^T A^T P^T PAQ$; hence $B^T B = Q^T A^T A Q$. Calculating $B^T B$, we set

$$B^T B = \begin{bmatrix} B_{11}^T B_{11} + B_{21}^T B_{21} & B_{11}^T B_{12} + B_{21}^T B_{22} \\ B_{12}^T B_{11} + B_{22}^T B_{21} & B_{12}^T B_{12} + B_{22}^T B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

From the signs in the block form of A we get that $C_{11} \geq 0, C_{22} \geq 0, C_{12} \leq 0$ and $C_{21} \leq 0$. Furthermore, if B_{11} is $r \times s$, it is easy to show that C_{11} is $s \times s$ and C_{22} is $(n - s) \times (n - s)$. It follows that $A^T A$ is a Morishima matrix. A similar argument shows that AA^T is a Morishima matrix.

This theorem shows that the class of quasi-Morishima matrices is closely connected to the class of Morishima matrices. We will show how the Perron–Frobenius theorem applies to these matrices, but we require some preliminary results first.

THEOREM 6. *Let A be a regular quasi-Morishima matrix. Then there exist diagonal matrices D_1 and D_2 with diagonal elements ± 1 such that $D_2AD_1 \geq 0$.*

Proof. Without loss of generality we can assume that A itself is not nonnegative; otherwise we can choose D_1 to be the $n \times n$ identity matrix and D_2 the $m \times m$ identity matrix. By Theorem 4 we can find P and Q such that

$$PAQ = \begin{bmatrix} A_{11} & -A_{12} \\ -A_{21} & A_{22} \end{bmatrix},$$

where all of the blocks are nonnegative. Suppose A_{11} is $r \times s$. Let $D_{10} = \text{diag}(1, \dots, 1, -1, \dots, -1)$ where there are s 1's and $n-s$ -1's, and let $D_{20} = \text{diag}(1, \dots, 1, -1, \dots, -1)$ where there are r 1's and $m-r$ -1's. Then $D_{20}PAQD_{10}$ is nonnegative. Define $D_2 = P^{-1}D_{20}P$ and $D_1 = QD_{10}Q^{-1}$, then $D_2AD_1 = P^{-1}D_{20}PAQD_{10}Q^{-1} \geq 0$, proving the theorem.

Note that the matrices D_1 and D_2 are in general not unique if CG^+ is not connected. We will discuss this and related matters in a subsequent paper.

Before applying this theorem let us consider an example. Let A be the quasi-Morishima matrix with sign pattern

$$\text{sgn } A = \begin{bmatrix} + & 0 & - & - & + & 0 \\ 0 & + & - & 0 & 0 & - \\ - & - & 0 & + & - & + \\ 0 & 0 & + & + & 0 & + \end{bmatrix}.$$

We can compute D_1 directly from the balanced graph CG^+ by partitioning its points into subsets C_1 and C_2 so that the lines have the usual properties. We obtain $C_1 = \{1, 2, 5\}$, $C_2 = \{3, 4, 6\}$. If we choose signs so that the scalar product of D_1 with row 1 is positive, we obtain $D_1 = \text{diag}(1, 1, -1, -1, 1, -1)$ and

$$\text{sgn } AD = \begin{bmatrix} + & 0 & + & + & + & 0 \\ 0 & + & + & 0 & 0 & + \\ - & - & 0 & - & - & - \\ 0 & 0 & - & - & 0 & - \end{bmatrix}$$

so that the sign pattern of the columns is $[+ + - -]$. Thus we choose $D_2 = \text{diag}[1, 1, -1, -1]$ and $D_2AD_1 \geq 0$.

Next observe that $(D_2AD_1)^T = D_1^T A^T D_2^T$ hence

$$(D_2AD_1)^T D_2AD_1 = D_1^T A^T D_2^T D_2AD_1.$$

But $D_2^T = D_2$ so $D_2^T D_2$ is the identity matrix and

$$(D_2AD_1)^T D_2AD_1 = D_1^T A^T AD_1 \geq 0.$$

It follows that the matrix $A^T A$ is similar to a nonnegative matrix, the similarity being effected by the diagonal matrix D_1 . Similarly we have

$$D_2AD_1(D_2AD_1)^T = D_2AA^T D_2 \geq 0.$$

Now it is well known that the matrices AA^T and $A^T A$ are symmetric and nonnegative definite (see [12], for example). Therefore the same will be true of the

matrices $D_1^T A^T A D_1$ and $D_2 A A^T D_2^T$. We are now ready to apply the Perron–Frobenius theory (see [12] for an elementary, but thorough, discussion).

THEOREM 7. *Let the $m \times n$ matrix A be a quasi-Morishima matrix with at least one nonzero element in each row and column and suppose that CG is connected. Then the matrix $A^T A$ has a simple eigenvalue r_n equal to its spectral radius and, if λ is any other eigenvalue of $A^T A$, we have $0 \leq \lambda < r_n$. Moreover, the eigenvector y belonging to r_n has all its components different from zero and its sign pattern is such that $D_1 y$ is a strictly positive vector $D_1 y > 0$, i.e., y has the same sign pattern as D_1 .*

Proof. We apply the Perron–Frobenius theorem to the nonnegative matrix $A = D_1^T A^T A D_1$. Since CG is connected, A is irreducible. Also $a_{ii} > 0$, $1 \leq i \leq n$, because each column of A has a nonzero element. Thus A is primitive. The theorem follows.

For completeness we state the corresponding theorem for AA^t .

THEOREM 7'. *Let the hypotheses of Theorem 6 be satisfied. Then the matrix AA^T has a simple eigenvalue r_m equal to its spectral radius and, if λ is any other eigenvalue of AA^T , we have $0 \leq \lambda < r_m$. Moreover, the eigenvector y belonging to r_m has all its components different from zero and the same sign pattern as D_2 .*

We note that it is known that $r_m = r_n$ and, more generally, that the spectrum of AA^T is the same as that of $A^T A$ except that one of these matrices may have more zero eigenvalues than the other depending upon which of m and n is larger.

4. Applications. Let us begin with an examination of elements in the class PFM. Consider first the graph $RG^+(A)$ for $A \in \text{PFM}$. We can partition the points of RG^+ into two disjoint sets S and M with the property that lines joining points of S or points of M are positive and lines joining a point of S and a point of M are negative. It follows that RG^+ is balanced. We thus have as a consequence of Theorem 3 the following result.

THEOREM 8. *If $A \in \text{PFM}$, then A is a quasi-Morishima matrix.*

We now know that for $A \in \text{PFM}$ the graph CG^+ is balanced, but we are interested in the structure of this graph in any case. In fact the points of CG^+ can be partitioned into *three* disjoint sets P , T and K such that each line joining points in the same set is positive, each line joining points in different sets is negative, and no line joins a point in P with a point in K .

If a matrix in the class PFM is regular the graphs RG^+ and CG^+ satisfy the following conditions:

(α) Each S point in RG^+ is adjacent to at least one M point and each M point to at least one S point.

(β) Each P point in CG^+ is adjacent to at least one T point and each K point to at least one T point.

(γ) Each T point in CG^+ is adjacent to at least one P point and at least one K point.

The effect of imposing regularity on a PFM is to insure that such matrices do describe a sort of flow. In fact the conditions guarantee that, when A is permuted into the form (1), to each nonzero element in a row of A_{11} there corresponds a nonzero element in the same row in A_{12} , to each nonzero element in a column of A_{12} there corresponds a nonzero element in the same column of A_{22} , and to each nonzero element in a row of A_{22} there corresponds a nonzero element in the same row of A_{23} . Thus there exists a connection (or flow) between elements of A_{11} and elements of A_{23} .

The structure of a PFM as displayed in (1) suggests a generalization of this class which still retains the properties of balance in CG^+ and RG^+ . A matrix A will be

called a *generalized physical flows matrix* (GPFM) if there exist permutation matrices P and Q such that

$$(4.1) \quad PAQ = \begin{bmatrix} A_{11} & A_{12} & 0 & \cdots & 0 & 0 \\ 0 & A_{22} & A_{23} & & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & & A_{pp} & A_{p,p+1} \end{bmatrix},$$

where $A_{ii} \geq 0, A_{i,i+1} \leq 0, 1 \leq i \leq p$. It is clear that the graphs RG^+ and CG^+ exist for $A \in \text{GPFM}$. Let us show that RG^+ is balanced for $p \geq 3$.

Note that RG^+ can be partitioned into disjoint sets $S_i, i \leq i \leq p$, such that lines joining two points of S_i are positive for $1 \leq i \leq p$, lines joining a pair of points one in S_i and one in S_{i+1} are negative, $1 \leq i \leq p - 1$, and no line joins a point in S_i to a point in S_j if $|i - j| > 1$. Therefore we may partition the points of RG^+ into the disjoint sets U (which is the union of the sets S_i for i odd) and V (which is the union of the sets S_i for i even), and the sets U and V satisfy the conditions for balance. Thus RG^+ , and hence CG^+ is balanced, and every element of GPFM is a quasi-Morishima matrix.

The form of RG^+ and CG^+ for $A \in \text{GPFM}$ suggests the following ideas which lead to another generalization of the class PFM still within the class of quasi-Morishima matrices. Define the signed graph G^+ to be a *signed ladder graph of order* $p \geq 2$ if the points of G^+ can be partitioned in p disjoint sets $L_i, 1 \leq i \leq p$, such that lines joining two points in L_i are positive, lines joining two points one in L_i and one in L_{i+1} are negative, $1 \leq i \leq p - 1$, and no line joins a point in L_i to a point in L_j if $|i - j| > 1$. Then G^+ is balanced. Define the matrix A to be a *quasi physical flows matrix* (QPFM) if $RG^+(A)$ is a signed ladder graph of order p and if CG^+ is a signed ladder graph of order $p + 1$. It is clear that $\text{GPFM} \subset \text{QPFM}$, but the converse is not true. It is also easy to see how to impose regularity conditions on the classes GPFM and QPFM. But even with regularity conditions imposed it is still not true that $\text{GPFM} = \text{QPFM}$. Here is a simple example. Let

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

Figure 5 illustrates $RG^+(A)$ and $CG^+(A)$. It is clear that RG^+ is a signed ladder graph of order $p = 2$ and CG^+ a signed ladder graph of order 3 so that $A \in \text{QPFM}$. It is also clear that $A \notin \text{GPFM}$, and that regularity conditions hold for A .



FIG. 5

We do not have a characterization of matrices in the class QPFM in block form. This is an interesting open question and we feel that its solution would provide some useful insights for energy modelers and others concerned with large scale systems.

We will conclude with a brief discussion of the class PFFM. The example illustrated in Fig. 2 of the smallest regular PFFM shows that not every element in this class is a quasi-Morishima matrix. However the PFFM

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

satisfies the conditions for a QPFM so it is a quasi-Morishima matrix.

It turns out that even if A is a regular PFFM it can be a quasi-Morishima matrix. We do not have a simple criterion that will enable us to characterize which elements of this class are quasi-Morishima and which are not. This must be regarded as an open problem.

As a final application let us turn to an entirely different area. In the theory of Lanchester models of military combat between heterogeneous forces (see [10] for some interesting examples) systems of differential equations having the form

$$\dot{x} = -Ay, \quad \dot{y} = -Bx$$

must be solved. In the simplest cases A is an $m \times n$ nonnegative matrix and B an $n \times m$ nonnegative matrix. The matrix of the system is therefore the quasi-Morishima matrix

$$a = \begin{bmatrix} 0 & -A \\ -B & 0 \end{bmatrix}.$$

The spectral properties of the system are most conveniently obtained by making use of Theorem 7.

A problem more general than the above is obtained when logistic considerations are incorporated into the equations of combat and, in particular, when one or both armies is being reinforced. In this case (see [13]) we obtain systems having the form

$$\dot{x} = A_1x + A_2y, \quad \dot{y} = B_1x + B_2y,$$

where A_1, A_2, B_1 and B_2 are rectangular matrices in general. When this happens one can identify certain key submatrices which are quasi-Morishima matrices. Here again the Perron-Frobenius theorem can be applied to obtain useful results.

REFERENCES

- [1] H. J. GREENBERG, J. R. LUNDGREN AND J. S. MAYBEE, *Graph-theoretic formulations of CAA*, in Proceedings of the Symposium on Computer Assisted Analysis and Model Simplification, H. J. Greenberg and J. S. Maybee, Eds. Academic Press, New York, 1981.
- [2] ———, *Graph theoretic methods for the qualitative analysis of rectangular matrices*, this Journal, 2 (1981), pp. 227–239.
- [3] H. J. GREENBERG, *Measuring complementarity and qualitative determinacy in matricial forms*, in Proceedings mentioned in [1].
- [4] F. HARARY, *On the notion of balance of a signed graph*, Michigan Math. J. 2 (1953–54), pp. 143–146.
- [5] ———, *Structural duality*, Behavioral Science, 2 (1957), pp. 255–265.
- [6] J. S. MAYBEE AND J. M. QUIRK, *Qualitative problems in matrix theory*, SIAM Rev., 11 (1969), pp. 30–51.
- [7] F. S. ROBERTS, *Discrete Mathematical Models with Applications to Social, Biological, and Environmental Problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1976.
- [8] F. HARARY, R. Z. NORMAN AND D. CARTWRIGHT, *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley, New York, 1965.

- [9] J. S. MAYBEE, *Some aspects of the theory of PN-matrices*, SIAM J. Appl. Math., 31 (1976), pp. 397–410.
- [10] ROGER F. WILLIS, *Study of the feasibility of developing useful theories of combat*, unpublished, U.S. Army, TRASANA, White Sands Missile Center, New Mexico.
- [11] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1971.
- [12] E. SENETA, *Non-Negative Matrices*, John Wiley, New York, 1973.
- [13] J. S. MAYBEE, *Models of military combat with logistics, a general approach*, Interim Tech. Rep. to U.S. Army Research Office, 1981.

LEAKY ELECTRICITY: 1-CHAIN FORMULAS FOR THE CURRENT AND VOLTAGE*

KENNETH A. BERMAN†

Abstract. A leaky electrical network N is a mathematical generalization of an electrical network. In this paper we present a formula expressing the current 1-chain of N as a linear combination of leaky uv -flows associated with the spanning 2-arborescences and a formula expressing the voltage 1-chain as a linear combination of coboundaries associated with the spanning 2-arborescences. This generalizes a result of Nerode and Shank [Amer. Math. Monthly, 68 (1961), pp. 244–247] on electrical networks.

1. Introduction. In “leaky” electricity, Kirchhoff’s first law is replaced by the statement that the sum of the currents in the edges directed out of each vertex different from the source or sink is zero (but not necessarily in the edges directed into that vertex). Kirchhoff’s second law that the sum of the voltages around any circuit is zero remains applicable. A leaky electrical network is a mathematical generalization of an electrical network. This can be seen as follows. Suppose N is an electrical network. Each wire in N may be represented by two edges, e and e' , such that e points in the direction of the current flow and e' points counter to the direction of the current flow. The digraph G^* obtained in this way is a *symmetric* digraph. Assign the amount of current in the wire to edge e and negative the amount of current in the wire to edge e' . By Kirchhoff’s first law the sum of the currents in the edges directed into a vertex of N equals the sum of the currents in the edges directed out of that vertex (except at the source or sink). But this implies that in G^* the sum of the values assigned to the edges directed out of a vertex equals zero (except at the source or sink). Thus electrical networks correspond to symmetric leaky electrical networks.

The theory of leaky electricity was developed by Tutte in [5], [6] and jointly by Brooks, Smith, Stone and Tutte in [1]. Tutte applied this theory to the problem of the dissection of an equilateral triangle into equilateral triangles. In his papers, Tutte gives an explicit expression for the current and voltages in the edges of a leaky electrical network in terms of the spanning 2-arborescences.

In this paper we give an elegant linear algebraic proof of this result. Our main theorem is a formula expressing the current 1-chain as a linear combination of leaky uv -flows associated with the spanning 2-arborescences and a formula expressing the voltage 1-chain as a linear combination of coboundaries associated with the spanning 2-arborescences. This theorem generalizes a result of Nerode and Shank [4] on electrical networks.

2. 1-chains. Let G be a strongly connected digraph, i.e., a digraph such that there is a directed path from each vertex to every other vertex. Let $E(G) = \{e_1, e_2, \dots, e_m\}$ denote the edge set of G . A 1-chain C is a mapping from the edges $E(G)$ into the real numbers R . We will refer to $C(e_j)$ as the *weight* of C on edge e_j for each edge $e_j \in E(G)$. The set \mathcal{C} of all 1-chains forms a vector space over R : For $C_1, C_2 \in \mathcal{C}$, $\lambda \in R$ and $e_j \in E(G)$ vector addition is defined by

$$(2.1) \quad (C_1 + C_2)(e_j) = C_1(e_j) + C_2(e_j)$$

* Received by the editors October 7, 1981.

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

and scalar multiplication is defined by

$$(2.2) \quad (\lambda C_1)(e_j) = \lambda C_1(e_j).$$

For u, v two vertices of G , a uv -flow F_{uv} is a 1-chain such that the sum of the weights of F_{uv} on the edges directed out of a vertex equals the sum of the weights on the edges directed into that vertex for every vertex but u and v . A *leaky uv -flow* L_{uv} is a 1-chain such that the sum of the weights of L_{uv} on the edges directed out of a vertex is zero for the every vertex but u and v . A *coboundary* B is a 1-chain such that for $e_j \in E$

$$(2.3) \quad B(e_j) = P(h) - P(t),$$

where h and t denote the head and tail of e_j , respectively, and P is a mapping from the vertices of G into the real numbers R . Mapping P is a 0-chain. It is immediate that the set \mathcal{F}_{uv} of uv -flows, the set \mathcal{L}_{uv} of leaky uv -flows and the set \mathcal{B} of coboundaries each form a subspace of \mathcal{C} .

Let C be a 1-chain from \mathcal{C} such that $C(e_j) > 0$ for all $e_j \in E$. A *leaky electrical network* N with conductivity 1-chain $C \in \mathcal{C}$, having source u and sink v , is a digraph G together with a current 1-chain I and a voltage 1-chain V such that

- (i) I is a leaky uv -flow,
- (ii) V is a coboundary,
- (iii) $I = CV$.

Condition (i) corresponds to the generalized Kirchoff's first law. Conditions (ii) and (iii) correspond to Kirchoff's second law and Ohm's law respectively. The *total current*, I_u (I_v), at vertex u (vertex v) is the sum of the currents in the edges directed out of u (v). Note that the total current at the source u need not equal the negative of the total current at the sink v , as is the case for symmetric networks. In [1] a simple argument is given to show that I and V are uniquely determined by conditions (i), (ii), (iii) up to within a constant multiplier. Thus if the total current at u is given then the current 1-chain I and the voltage 1-chain V are uniquely determined.

An (*in*) *arborescence* A_v rooted at vertex v of G is a tree of G such that there is a directed path in the tree from every vertex in the tree to v . (Arborescence A_v may be the single vertex v .) If A_v spans the vertices then A_v is a *spanning arborescence*. Let \mathcal{A}_v denote the set of all spanning arborescences rooted at v . Set $c_j = C(e_j)$. The implexity α_v of G at vertex v is given by

$$(2.4) \quad \alpha_v = \sum_{A_v \in \mathcal{A}_v} \prod_{e_j \in E(A_v)} c_j,$$

where $E(A_v)$ denotes the edge set of an arborescence A_v . Since G is strongly connected there is at least one spanning arborescence rooted at v . Since $c_j > 0$ for each edge $e_j \in E(G)$ it follows that the implexity α_v is strictly greater than zero.

A *spanning 2-arborescence* $A = (A_u, A_v)$ rooted at vertices u and v is a pair of arborescences, A_u rooted at u and A_v rooted at v , which are vertex disjoint and together span the vertices of G . Let \mathcal{A}_{uv} denote the set of all spanning 2-arborescences rooted at vertices u and v . The weight π_A of a spanning 2-arborescence $A \in \mathcal{A}_{uv}$ is given by

$$(2.5) \quad \pi_A = \prod_{e_j \in E(A)} c_j$$

where $E(A)$ denotes the edge set of A . Let V and W be the sets of vertices spanned by A_u and A_v , respectively, and let $S(A)$ be the set of edges of G having one end vertex in V and the other in W . Let $S^+(A)$ be the subset of edges in $S(A)$ directed

out of vertices in V and $S^-(A)$ be the subset of edges in $S(A)$ directed out of vertices in W . For $e_j \in S(A)$, set

$$(2.6) \quad \text{sign}_A e_j = \begin{cases} 1, & e_j \in S^+(A), \\ -1, & e_j \in S^-(A). \end{cases}$$

Let $S_{uv}(A)$ be the subset of edges in $S(A)$ directed out of either vertex u or vertex v . For $e_j \in E(A)$ let $T_j(A)$ be the subset of edges in $S(A)$ having the same tail as e_j .

With each spanning 2-arborescence A of \mathcal{A}_{uv} associate the coboundary B_A defined by

$$(2.7) \quad B_A(e_j) = \begin{cases} \text{sign}_A e_j, & e_j \in S(A), \\ 0, & \text{otherwise,} \end{cases}$$

and the leaky uv -flow L_A defined by

$$(2.8) \quad L_A(e_j) = \begin{cases} (\text{sign}_A e_j)c_j, & e_j \in S_{uv}(A), \\ \frac{1}{2}(\text{sign } e_j)c_j, & e_j \in S(A) - S_{uv}(A), \\ -\frac{1}{2} \sum_{e_k \in T_j(A)} (\text{sign } e_k)c_k, & e_j \in E(A), \\ 0, & \text{otherwise.} \end{cases}$$

3. Formulas for current and voltage 1-chains. In this section we prove the following main theorem.

THEOREM 3.1. *Let N be a leaky electrical network with 1-chain C such that the total current at the source u is 1. Let α_v be the implexity at the sink v and let π_A , B_A and L_A be defined by (2.5), (2.7), (2.8), respectively. Then the current 1-chain I and the voltage 1-chain V are given by*

$$(3.1) \quad I = \frac{1}{\alpha_v} \sum_{A \in \mathcal{A}_{uv}} \pi_A L_A,$$

$$(3.2) \quad V = \frac{1}{\alpha_v} \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A.$$

Since the set of leaky uv -flows and the set of coboundaries form subspaces of \mathcal{C} it follows that $(1/\alpha_v) \sum_{A \in \mathcal{A}_{uv}} \pi_A L_A$ is a leaky uv -flow and $(1/\alpha_v) \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A$ is a coboundary. Let δ_u^+ be the set of edges directed out of vertex u . Then

$$\begin{aligned} \sum_{e_j \in \delta_u^+} \sum_{A \in \mathcal{A}_{uv}} \pi_A L_A(e_j) &= \sum_{e_j \in \delta_u^+} \sum_{A \in \mathcal{A}_{uv}} \pi_A c_j = \sum_{A \in \mathcal{A}_{uv}} \sum_{e_j \in \delta_u^+} \pi_A c_j \\ &= \sum_{A_v \in \mathcal{A}_v} \prod_{e_k \in E(A_v)} c_k = \alpha_v. \end{aligned}$$

Hence to prove the theorem it suffices to show that

$$(3.3) \quad \sum_{A \in \mathcal{A}_{uv}} \pi_A L_A = C \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A.$$

Consider an edge $e_j \in E(G)$. Let \mathcal{A}_j be the subset of spanning 2-arborescences in \mathcal{A}_{uv} which contain edge e_j and let \mathcal{A}'_j be the subset of spanning 2-arborescences in \mathcal{A}_{uv} which do not contain edge e_j . For $A \in \mathcal{A}_j$ and $e_k \in T_j(A)$ let A_{jk} be the spanning 2-arborescence obtained from A by deleting edge e_j and adding edge e_k .

If e_j is directed out of either vertex u or v then it is immediate from the definition of L_A and B_A that

$$\sum_{A \in \mathcal{A}_{uv}} \pi_A L_A(e_j) = c_j \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A(e_j).$$

Now suppose e_j is neither directed out of vertex u nor vertex v . Then it is immediate from the definition of L_A and B_A that

$$\sum_{A \in \mathcal{A}_j} \pi_A L_A(e_j) = \frac{1}{2} c_j \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A(e_j).$$

We also have that

$$\begin{aligned} \sum_{A \in \mathcal{A}_j} \pi_A L_A(e_j) &= \sum_{A \in \mathcal{A}_j} \pi_A \left(-\frac{1}{2} \sum_{e_k \in T_j(A)} (\text{sign}_A e_k) c_k \right) \\ &= \sum_{A \in \mathcal{A}_j} \sum_{e_k \in T_j(A)} \pi_{A_{jk}} L_{A_{jk}}(e_j) = \sum_{A \in \mathcal{A}_j} \pi_A L_A(e_j). \end{aligned}$$

Hence

$$\sum_{A \in \mathcal{A}_{uv}} \pi_A L_A(e_j) = \sum_{A \in \mathcal{A}_j} \pi_A L_A(e_j) + \sum_{A \in \mathcal{A}_j} \pi_A L_A(e_j) = c_j \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A(e_j).$$

The theorem follows.

The following corollary due to Tutte [1], [5], [6] follows immediately from Theorem 3.1.

COROLLARY (Tutte). *Let e_j be an edge of G with tail t and head h . Let $\alpha_{ut,vh}$ be the sum of the weights over the spanning 2-arborescences such that t lies in the arborescence rooted at u and h lies in the arborescence rooted at v . Similarly define $\alpha_{uh,vr}$. If the total current at the source u is 1 then*

$$(3.4) \quad V(e_j) = \frac{\alpha_{ut,vh} - \alpha_{uh,vr}}{\alpha_v}.$$

4. Electrical networks. An electrical network N with conductivity 1-chain C having source u and sink v is a digraph G together with a current 1-chain I and a voltage 1-chain V such that

- (i) I is a uv -flow,
- (ii) V is a coboundary,
- (iii) $I = CV$.

Conditions (i) and (ii) correspond to Kirchhoff's two laws of electricity and condition (iii) corresponds to Ohm's Law. The *total current* is the current entering the network at the source u (which equals the current leaving the network at the sink v).

With the electrical network N we may associate the leaky electrical network N^* as follows. Replace each edge e_j of G with two edges, e_j^0 and e_j^1 , both joining the same two vertices as e_j but with e_j^0 directed the same as e_j and e_j^1 directed counter to e_j . Call the resulting digraph G^* . Set $I^*(e_j^0) = I(e_j)$ and $I^*(e_j^1) = -I(e_j)$. Also set $V^*(e_j^i) = V(e_j)$, ($i = 0, 1$), and $C^*(e_j^i) = C(e_j)$, ($i = 0, 1$). Then the digraph G^* together with current 1-chain I^* and voltage 1-chain V^* determine a leaky electrical network N^* with conductivity 1-chain C^* .

Let \mathcal{T} denote the set of spanning trees of G . The *complexity* τ of N is given by

$$(4.1) \quad \tau = \sum_{T \in \mathcal{T}} \prod_{e_j \in E(T)} c_j$$

where $E(T)$ is the edge set of a spanning tree T . A *spanning 2-tree*, $T = (T_u, T_v)$,

separating vertices u and v is a tree T_u containing u and a tree T_v containing v which are vertex disjoint and together span the vertices of G . Let \mathcal{T}_{uv} denote the set of all spanning 2-trees separating vertices u and v . The weight π_T of a spanning 2-tree $T \in \mathcal{T}_{uv}$ is given by

$$(4.2) \quad \pi_T = \prod_{e_j \in E(T)} c_j$$

where $E(T)$ is the edge set of T . Let V and W be the sets of vertices spanned by T_u and T_v , respectively. Let B_T be the coboundary defined by

$$(4.3) \quad B_T(e_j) = \begin{cases} 1, & e_j \text{ has tail in } V \text{ and head in } W, \\ -1, & e_j \text{ has tail in } W \text{ and head in } V, \\ 0 & \text{otherwise.} \end{cases}$$

The following theorem of Nerode and Shank [4] is a corollary of Theorem 3.1.

THEOREM 4.1 (Nerode and Shank). *Let N be an electrical network with conductivity 1-chain C such that the total current is 1. Let τ be the complexity of N and let π_T and B_T be defined by (4.2) and (4.3) respectively. Then the voltage 1-chain V is given by*

$$(4.4) \quad V = \frac{1}{\tau} \sum_{T \in \mathcal{T}_{uv}} \pi_T B_T.$$

By applying Theorem 3.1 to the associated leaky electrical network N^* we have

$$V(e_j) = V^*(e_j^0) = \frac{1}{\alpha_v} \sum_{A \in \mathcal{A}_{uv}} \pi_A B_A^*(e_j^0) = \frac{1}{\tau} \sum_{T \in \mathcal{T}_{uv}} \pi_T B_T(e_j).$$

As a corollary of Theorem 4.1 we have the following famous result of Kirchhoff [2], [3].

COROLLARY (Kirchhoff). *Let e_j be an edge of G with tail t and head h . Let $\tau_{ut,vh}$ be the sum of the weights over the spanning 2-trees such that u and t lie in one tree and v and h lie in the other tree. Similarly, define $\tau_{uh,vt}$. If the total current is 1 then*

$$(4.5) \quad V(e_j) = \frac{\tau_{ut,vh} - \tau_{uh,vt}}{\tau}.$$

REFERENCES

[1] R. L. BROOKS, C. A. B. SMITH, A. H. STONE AND W. T. TUTTE, *Leaky electricity and triangulated triangles*, Philips Res. Repts, 30 (issue in honour of C. J. Bouwkamp), 1975, pp. 205-219.
 [2] W. K. CHEN, *Applied Graph Theory*, North-Holland, New York, 1976.
 [3] G. KIRCHHOFF, *Über die Auflösung der Gleichungen, auf Welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geföhrt wird*, Poggendorf's Ann. Phys. Chem., 72, (1847), pp. 497-508.
 [4] A. NERODE AND H. SHANK, *An algebraic proof of Kirchhoff's Network Theorem*, Amer. Math. Monthly, 68 (1961), pp. 244-247.
 [5] W. T. TUTTE, *The dissection of equilateral triangles into equilateral triangles*, Proc. Cambridge Phil. Soc., 44 (1948), pp. 463-482.
 [6] ———, *The rotor effect with generalized electrical flows*, Ars Combinatoria, 1 (1976), pp. 3-31.

INEQUALITIES FOR SUBSETS OF A SET AND KLYM POSETS*

DAVID E. DAYKIN† AND PETER FRANKL‡

Abstract. We strengthen and generalise the LYM inequality for KLYM posets. The set of all subsets of a finite set is a KLYM poset, and for it we discuss conjectured inequalities.

Key words. antichain, chain length, comparable elements, convex, finite poset, KLYM poset, list of maximal chains, LYM inequality, ranks

1. On KLYM posets. Let P be a finite poset with ranks $0, 1, \dots, n$. For $0 \leq i \leq n$ let $r(i)$ be the number of elements of P of rank i . If $x \in P$ we write $r(x)$ for $r(\text{rank}(x))$. A maximal chain of P has length $n + 1$. We call $A \subset P$ an *antichain* if it has no chain of length 2. Kleitman proved [3] that the following two conditions are equivalent for P :

- (1) If $A \subset P$ is an antichain then $\sum (x \in A) 1/r(x) \leq 1$.
- (2) There is a nonempty list $C(1), \dots, C(c)$ of not necessarily distinct maximal chains of P with the property that for every $x \in P$ the number of chains in the list which contain x is $c/r(x)$.

We assume that P satisfies (1), (2) and call it a KLYM poset in view of the writings of Kleitman, Lubell, Yamamoto and Meshalkin (cf. [3]). Usually (1) is called the LYM inequality. An example for P is the lattice L of all subsets of $N = \{1, 2, \dots, n\}$ ordered by inclusion.

Let $q(0), \dots, q(n)$ be the numbers $r(0), \dots, r(n)$ ordered so that $q(0) \geq q(1) \geq \dots \geq q(n)$. Then let $Q(b) = q(0) + \dots + q(b - 1)$, so in particular $Q(n + 1) = |P| = \sum r(i)$. We denote $q(0) = Q(1)$ by $|S|$ because Sperner proved that if A is an antichain in L then $|S| \geq |A|$. For $B \subset P$ let the set of elements comparable with B be

$$\text{comp } B = \{x : x \in P, \exists y \in B, \text{ either } x \leq y \text{ or } y \leq x\}.$$

Then we give the following generalization of (1).

THEOREM 1. *If $B \subset P$ has no chain of length $b + 1$, then*

$$\max \{b|B|/Q(b), \sum (x \in B) 1/r(x)\} \leq b|\text{comp } B|/|P|.$$

Proof. *Part (i).* Let A be an antichain in P . Put $E = P \setminus (\text{comp } A)$ and for $0 \leq i \leq n$ let $E(i)$ be the set of elements of E of rank i . Then there is a j such that $|E|r(j) \leq |P||E(j)|$. Using this fact and applying (1) to the antichain $A \cup E(j)$ we get

$$\begin{aligned} |E|/|P| + \sum (x \in A) 1/r(x) &\leq |E(j)|/r(j) + \sum (x \in A) 1/r(x) \\ &= \sum (x \in A \cup E(j)) 1/r(x) \leq 1. \end{aligned}$$

In other words,

$$(3) \quad \sum (x \in A) 1/r(x) \leq |\text{comp } A|/|P|.$$

Part (ii). We can partition B as $B = A_1 \cup \dots \cup A_b$ where A_j are antichains. Then we apply (3) to each A_j and sum over j . Since $\text{comp } A_j \subset \text{comp } B$, this gives the second inequality of the theorem.

* Received by the editors May 20, 1981.

† Department of Mathematics, University of Reading, Reading RG6 2AX, Berkshire, Great Britain.

‡ Centre National de la Recherche Scientifique, Paris, France.

Part (iii). We have

$$\begin{aligned}
 c|P||B| &= |P| \sum (1 \leq i \leq c) \sum (x \in B \cap C(i))r(x) \\
 &\leq |P| \sum (1 \leq i \leq c, B \cap C(i) \neq \emptyset)Q(b) \\
 &= Q(b) \sum (1 \leq i \leq c, B \cap C(i) \neq \emptyset)|P| \\
 &= Q(b) \sum (1 \leq i \leq c, B \cap C(i) \neq \emptyset) \sum (x \in (\text{comp } B) \cap C(i))r(x) \\
 &\leq Q(b) \sum (1 \leq i \leq c) \sum (x \in (\text{comp } B) \cap C(i))r(x) \\
 &= Q(b)c|\text{comp } B|,
 \end{aligned}$$

and the theorem is proved. \square

Notice that $|B|/|S|$ is a lower bound for both terms in $\max\{\cdot, \cdot\}$. All the inequalities are best possible in the sense that examples give equality.

2. On the lattice L of subsets of N . We say that V is *convex* if $x, z \in V$ and $x \leq y \leq z$ imply $y \in V$. In particular $L \setminus \text{comp } F$ is convex for every $F \subset L$. We let D denote a *down-set* so $x \leq y$ and $y \in D$ imply $x \in D$. We let U denote an *up-set* so $x \leq y$ and $x \in U$ imply $y \in U$.

Conjecture 1. If V is convex, $|S|/|L| \leq |A|/|V|$ for some antichain $A \subset V$.

Question. If V is convex, is

$$|V||V| |V \cap D \cap U| |S| \leq |L| |V \cap D| |V \cap U| |A|$$

for some antichain $A \subset V$? This question is related to:

Conjecture 2. If A is an antichain and $V = L \setminus \text{comp } A$ then

$$|V||V| |V \cap D \cap U| |S| \leq |L| |V \cap D| |V \cap U| (|S| - |A|).$$

When $A = \emptyset$ this inequality reduces to Kleitman's classical [1], [2] inequality $|L| |D \cap U| \leq |D| |U|$.

Next let

$$m = \max\{|L|/4, |S|\} = \begin{cases} |S| & \text{if } 1 \leq n \leq 8, \\ |L|/4 & \text{if } 9 \leq n. \end{cases}$$

Conjecture 3. If $X, Y \subset L$ are such that $x \in X, y \in Y, x \neq y$ imply $x \not\prec y, y \not\prec x$ then $\min\{|X|, |Y|\} \leq m$.

This clearly implies the folklore:

Conjecture 4. If $X \subset L$ is such that $x, y \in X$ implies either $x = N \setminus y$ or both $x \cup y = N$ and $x \cap y = \emptyset$, then $|X| \leq m$.

THEOREM 2. *Conjecture 2 implies Conjecture 3.*

Proof. Given X, Y satisfying the conditions of Conjecture 3, let A be the antichain $X \cap Y$ and let V be $L \setminus \text{comp } A$. Also put $G = X \setminus A \subset V$ and $H = Y \setminus A \subset V$. For any $Z \subset L$, define

below $Z = \{x : x \in L, \exists z \in Z, x \leq z\} = \text{down-set,}$

above $Z = \{x : x \in L, \exists z \in Z, x \geq z\} = \text{up-set.}$

Then $(\text{below } G) \cap (\text{above } H) = \emptyset$, so $|V \cap \text{below } G| + |V \cap \text{above } H| \leq |V|$, and similarly $|V \cap \text{above } G| + |V \cap \text{below } H| \leq |V|$, so we assume $|V \cap \text{below } G| + |V \cap \text{above } G| \leq |V|$, and then $|V \cap \text{below } G| |V \cap \text{above } G| \leq |V|^2/4$. If $V = \emptyset$ then

$X = Y = A$ and $|A| \leq |S|$. If $V \neq \emptyset$ then

$$\begin{aligned} |G| &= |V \cap G| \leq |V \cap (\text{below } G) \cap (\text{above } G)| \\ &\leq |L| |V \cap \text{below } G| |V \cap \text{above } G| (|S| - |A|) / |S| |V| |V| \\ &\leq |L| (|S| - |A|) / 4|S| = t, \quad \text{say.} \end{aligned}$$

If $1 \leq n \leq 8$ then $|L| < 4|S|$ and $t < |S| - |A|$. If $9 \leq n$ then $4|S| < |L|$, so $t \leq (|L|/4) - |A|$. In either case, $|X| = |A| + |G| \leq m$, as required. \square

REFERENCES

- [1] D. E. DAYKIN, *An hierarchy of inequalities*, Stud. Appl. Math., 63 (1980), pp. 263–274.
- [2] D. J. KLEITMAN, *Families of non-disjoint subsets*, J. Combin. Theory, 1 (1966), pp. 153–155.
- [3] ———, *On an extremal property of antichains, etc.*, in Combinatorics, Proc. Conf. Breukelen (1974), M. Hall and J. H. van Lint, eds., Tract 55, Math. Centrum, Amsterdam, 1974, pp.77–90.

THE COMPLEXITY OF SOLVING POLYNOMIAL EQUATIONS BY PRIME ROOT EXTRACTIONS*

GEORG GATI†

Abstract. We show that for each algebraic number field Q of finite degree and for each natural number $n = p_1 \cdots p_k$, p_i prime, there exists an irreducible polynomial f_n of degree n in $Q[x]$ such that f_n is solvable by radicals and $1 + p_1 + p_1 p_2 + \cdots + p_1 p_2 \cdots p_{k-1}$ extractions of p_i th roots are required for obtaining all roots of f_n . This generalizes the linear lower bound on the number of circles one has to draw in order to solve a geometric construction problem by ruler and compass and exponentially improves the obvious lower bound which is the number of prime factors of $\varphi(n)$, the value of Euler's function on n .

Key words. complexity, lower bounds, polynomial equations

Let Q be an algebraic number field of finite degree and $f(x)$ an irreducible polynomial in $Q[x]$. If the equation $f(x) = 0$ is solvable by radicals then the splitting field of f can be constructed by repeated cyclic extensions of Q . Computationally, all roots of f can be calculated if all roots of binomial equations $y^n - a = 0$ can be found. If n is a composite number then the equation $y^n - a = 0$ can be replaced by several equations of degree less than n . However, binomial equations of prime degree cannot be replaced by binomial equations of smaller degree. We were thus led to the following question: How many binomial equations of prime degree have to be solved in order to obtain all roots of an irreducible polynomial $f(x)$ in $Q[x]$ of degree n ?

Consider the case when Q is the field Q of rational numbers and $f(x)$ is the cyclotomic polynomial $\phi_n(x)$ which is of degree $\varphi(n)$ where φ denotes the Euler function. ϕ_n is irreducible over Q and has Galois group $C_{\varphi(n)}$. Hence the equation $\phi_n(x) = 0$ is solvable and the number of binomial equations of prime degree which have to be solved in order to obtain all roots of ϕ_n is the number of prime factors of $\varphi(n)$. For every nonprime natural number n we shall prove the existence of irreducible polynomials of degree n which considerably improve this obvious lower bound, namely to $1 + p_1 + p_1 p_2 + \cdots + p_1 p_2 \cdots p_{k-1}$ if $n = p_1 p_2 \cdots p_k$ is a factorization of n into primes.

The case $n = 2^k$, $k \in \mathbb{N}$, yields the number of circles one has to draw in order to solve a geometric construction problem by ruler and compass and was treated in [1].

THEOREM 1 (Šafarevič [2]). *Let Q be an algebraic number field of finite degree and G a finite solvable group. Then there exists a Galois extension of Q with Galois group G .*

DEFINITION. A subgroup H of a group G is *corefree* if H contains no nontrivial proper normal subgroup of G .

The following result is a consequence of Steinitz's theorem and the fundamental theorem of Galois theory; for a proof see [1].

THEOREM 2. *Let Q be an algebraic number field of finite degree and Q' a finite Galois extension of Q with Galois group G . If H is a corefree subgroup of G then there exists an irreducible polynomial $f(x)$ in $Q[x]$ of degree $(G:H)$ with splitting field Q' .*

Let n be a natural number with prime factorization $n = p_1 \cdots p_k$. We denote by C_n the iterated wreath product $C_{p_1}[C_{p_2}[\cdots [C_{p_k}]\cdots]]$, where the C_{p_i} are cyclic groups in p_i elements. The group C_n is solvable and isomorphic to a subgroup of the automorphism group of the rooted tree in which the root has out degree p_1 , each

* Received by the editors July 28, 1981, and in revised form November 15, 1981. This research was supported by the Swiss National Science Foundation under grant 82.813-0.80.

† Eigenössische Technische Hochschule Zürich, Zürich, Switzerland, and Institut für Statistik und Informatik, Universität Wien, A-1010 Wien, Austria.

successor of the root has out degree p_2 , etc. Let H_n be the stabilizer subgroup in C_n of a leaf of this tree. Then H_n is a corefree subgroup of C_n of index n .

We now can prove:

THEOREM 3. *Let Q be an algebraic number field of finite degree and $n \cong 2$ a natural number with prime factorization $n = p_1 p_2 \cdots p_k$. Then there exists an irreducible polynomial $f(x)$ in $Q[x]$ of degree n such that $f(x)$ is solvable by radicals and one has to solve $1 + p_1 + p_1 p_2 + \cdots + p_1 p_2 \cdots p_{k-1}$ binomial equations of prime degree in order to obtain all roots of $f(x)$.*

Proof. By Theorem 1 there exists a Galois extension Q' of Q with Galois group C_n . By Theorem 2 there exists an irreducible polynomial $f(x)$ in $Q[x]$ of degree n with splitting field Q' . It follows that the Galois group of f over Q is C_n and that f is solvable by radicals. The fundamental theorem of Galois theory implies that the number of binomial equations of prime degree that have to be solved in order to obtain all roots of f is the number of factors in a prime factorization of $|C_n|$, namely $1 + p_1 + p_1 p_2 + \cdots + p_1 p_2 \cdots p_{k-1}$.

Acknowledgments. We have benefited from advice by M. O. Rabin and J. T. Tate.

REFERENCES

- [1] G. GATI, *The complexity of solving polynomial equations by quadrature*, J. Assoc. Comput. Mach., to appear.
- [2] I. R. ŠAFAREVIČ, *On extensions of fields of algebraic numbers solvable in radicals*, Dokl. Akad. Nauk SSSR 95 (1954), pp. 225–227. (In Russian.)
- [3] ———, *Construction of fields of algebraic numbers with given solvable Galois group*, Izv. Akad. Nauk. SSSR. Ser. Mat., 18 (1954), pp. 525–578; AMS Transl. Ser. 2, 4 (1956), pp. 185–237.

SCHEDULING OPPOSING FORESTS*

M. R. GAREY† D. S. JOHNSON†, R. E. TARJAN†‡ AND M. YANNAKAKIS†

Abstract. A basic problem of deterministic scheduling theory is that of scheduling n unit-length tasks on m identical processors subject to precedence constraints so as to meet a given overall deadline. T. C. Hu's classic "level algorithm" can be used to solve this problem in linear time if the precedence constraints have the form of an in-forest or an out-forest. We show that a polynomial time algorithm for a wider class of precedence constraints is unlikely, by proving the problem to be NP-complete for precedence constraints that are the disjoint union of an in-forest and an out-forest (the "opposing forests" of our title). However, for any fixed value of m we show that this problem can be solved in polynomial time for such precedence constraints. For the special case of $m = 3$ we provide a linear time algorithm.

1. Introduction. One of the fundamental problems of deterministic scheduling theory is that of scheduling unit-length tasks on a collection of identical processors subject to precedence constraints so as to meet a given overall deadline. Given a number m of processors, a deadline D , and a task system $S = (T, <)$, where $T = \{T_1, T_2, \dots, T_n\}$ is a set of tasks and $<$ is a partial order on T , an m -processor schedule for S that meets deadline D is an assignment $\sigma : T \rightarrow \{0, 1, \dots, D - 1\}$ which satisfies the following processor and precedence constraints:

- (1) For all t , $0 \leq t \leq D - 1$, $|\{T_i \in T : \sigma(T_i) = t\}| \leq m$.
- (2) For all $T_i, T_j \in T$, $T_i < T_j$ implies $\sigma(T_j) \geq \sigma(T_i) + 1$.

A task T_i with $\sigma(T_i) = t$ is said to *start* at time t and *finish* at time $t + 1$. The set $\{T_i \in T : \sigma(T_i) = t - 1\}$ is called the set of tasks executed in the t th time slot (tasks in the first time slot start at time $t = 0$).

In what follows, we shall refer to the question, "Given m , D , and S , is there an m -processor schedule for S that meets deadline D ?" as the multiprocessor scheduling problem (MS). This problem provides an elementary model for a number of scheduling situations, including the pre-emptive scheduling of multiprocessor systems [2]. It is known to be NP-complete [18] and hence is unlikely to be solvable by any polynomial time algorithm (see [11] for an introduction to the theory of NP-completeness). However, a number of special cases can be solved with polynomial time algorithms.

In a classic 1961 paper [12], T. C. Hu presented an algorithm, called the "level algorithm," which can be used to solve MS in time $O(n)$ when S is either an in-forest, i.e., each task has at most one immediate successor, or an out-forest, i.e., each task has at most one immediate predecessor. (T_i is an *immediate predecessor* of T_j , and T_j is an *immediate successor* of T_i , if $T_i < T_j$ and no other task T_k satisfies $T_i < T_k < T_j$.)

Quite recently, Papadimitriou and Yannakakis [14] have shown that if $<$ is an *interval order* (each task T_i corresponds to an interval $[a_i, b_i]$ on the real line, and $T_i < T_j$ if and only if $b_i < a_j$), then MS can be solved in time $O(n^2)$. Even more recently, Warmuth [19] has shown that if $<$ is a *level order* (any two incomparable tasks with a common predecessor or successor have identical sets of predecessors and successors), then for any fixed value of m MS can be solved in time $O(n^{m-1})$.

* Received by the editors September 18, 1981, and in revised form April 29, 1982.

† Bell Laboratories, Murray Hill, New Jersey 07974.

‡ The work of this author was done in part while he was with the Computer Science Department, Stanford University, and was partially supported by the National Science Foundation under grant MCS75-22870-A02 and by the Office of Naval Research under contract N00014-76-C-0688.

For arbitrary partial orders, the most general case that is known to be solvable in polynomial time is that for $m = 2$. The first polynomial time algorithm for this case was presented in 1969 by Fujii, Kasami and Ninomiya [5], and improved algorithms have been obtained by Coffman and Graham [3], Sethi [16] and Gabow [6].

No other significant subcases have been identified as being solvable in polynomial time, although results of this sort have been obtained for a number of variants on the basic model [1], [2], [7], [8], [9]. In particular, it is shown in [1] that if each task has an individual deadline it must meet, then the problem of scheduling equal length tasks on m processors remains solvable in polynomial time for in-forests but becomes NP-complete for out-forests.

In this paper we return to the basic multiprocessor scheduling model and consider the case in which S can be partitioned into two disjoint and independent task systems, $S_I = (T_I, <_I)$ and $S_O = (T_O, <_O)$, where S_I is an in-forest task system and S_O is an out-forest task system. We call such a composite system S an *opposing forest*.

Opposing forest task systems form essentially the simplest natural generalization of the cases solvable by Hu's level algorithm. They are properly included in the class of series-parallel partial orders, a class to which algorithms for in-forest partial orders have often been generalized in other scheduling domains [13], [14], [17]. However, for MS, we shall show that the opposing forest case is NP-complete. (The same holds for certain less natural generalizations of the polynomially solvable cases, such as when S is the union of interval ordered task systems or the union of an interval ordered task system and an in- or out-forest.)

When the number of processors m is fixed, the outlook is somewhat brighter. We present an algorithm for the opposing forest case that runs in time bounded by the polynomial $O(n^{m^2+2m-5})$. Clearly such a bound cannot be used to justify a claim of having found an "efficient" algorithm for large m , but it does verify that the problem can be solved in polynomial time for fixed m and offers hope that subsequent work may lead to significant reductions in the exponent. Moreover, this algorithm shows that task systems more complicated than opposing forests will be required if one is to resolve the longstanding open problem of MS for fixed m by showing that there is some value of m for which the problem is NP-complete.

The paper is organized as follows: In § 2 we introduce the *profile scheduling problem* and show its relationship to opposing forest MS, proving that both problems are NP-complete for general m . In § 3 we describe our general algorithm, which is based on a subroutine for the "monotone" profile scheduling problem. In § 4 we show how special techniques allow us to solve three-processor opposing forest MS in linear time, a substantial improvement over the general algorithm. We conclude in § 5 with a discussion of the other NP-completeness results mentioned above and some directions for further research.

Before moving on to those results, it should be remarked that the earlier algorithms we have cited [3], [12], [15] all construct *minimum makespan* schedules, i.e., schedules in which the latest finishing task finishes as early as possible. Our new algorithms in contrast only determine whether or not a schedule exists that meets the given deadline, constructing one if it does, but without guaranteeing any additional properties for that schedule. Thus, in the event a minimum makespan schedule is desired for these cases, the algorithms must be applied approximately $\log n$ times in a binary search for the minimum achievable deadline, increasing the running time by a factor of $\log n$. For the three processor case, however, we show how to find minimum makespan schedules without introducing this additional $\log n$ factor.

2. Profile scheduling. In this section we introduce the notion of a “profile” and show its relationship to the scheduling problems we are considering. A *profile* is simply a sequence $\bar{m} = (m_0, m_1, \dots, m_k)$ of nonnegative integers. A schedule σ for a task system S meets the profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ if it meets the deadline D , and if for all i , $0 \leq i \leq D-1$, it satisfies

$$|\{T_j \in T: \sigma(T_j) = i\}| \leq m_i.$$

The schedule *has* the given profile if equality holds in the above for all values of i .

An m -processor schedule σ for an opposing forest task system $S = S_I \cup S_O$ meets the *internal profile* $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ if it meets the deadline D and for all i , $0 \leq i \leq D-1$,

$$|\{T_j \in T_I: \sigma(T_j) = i\}| \leq m_i \quad \text{and} \quad |\{T_j \in T_O: \sigma(T_j) = i\}| \leq m - m_i.$$

Note that if σ meets the internal profile \bar{m} , it can be viewed as a composite schedule made up of a schedule σ_I for S_I that meets \bar{m} and a schedule σ_O for S_O that meets the complementary profile $\bar{m}^C = (m - m_0, m - m_1, \dots, m - m_{D-1})$.

The opposing forest MS problem is closely related to the following problem for S an in-forest or an out-forest task system:

PROFILE SCHEDULING (PS)

Instance. Task system $S = (T, <)$, number m of processors, deadline D , and a profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$, with $0 \leq m_i \leq m$ for $0 \leq i \leq D-1$.

Question. Is there a schedule σ for S that meets profile \bar{m} ?

Our first observation is that in-forest PS is linearly equivalent to out-forest PS. Given an instance of one, we can convert it to an instance of the other by replacing $<$ by the reversed partial order $<^R$ ($T_i <^R T_j$ if and only if $T_j < T_i$) and replacing \bar{m} by the reversed profile $\bar{m}^R = (m_{D-1}, \dots, m_2, m_1, m_0)$.

We shall be particularly interested in the PS problem for “monotone” profiles. A *monotone profile* for an in-forest is one satisfying $m_0 \geq m_1 \geq \dots \geq m_{D-1}$. A *monotone profile* for an out-forest is one satisfying $m_0 \leq m_1 \leq \dots \leq m_{D-1}$. A *monotone internal profile* for an opposing forest task system $S = S_I \cup S_O$ is one satisfying $m_0 \geq m_1 \geq \dots \geq m_{D-1}$. It imposes the monotone profile \bar{m} on the in-forest S_I and the monotone profile \bar{m}^C on the out-forest S_O . We shall call the case of PS in which S is an in-forest (out-forest) task system and \bar{m} is monotone for S the MONOTONE PROFILE SCHEDULING (MPS) problem for in-forests (out-forests). Note that by the above remarks in-forest MPS is equivalent to out-forest MPS; an algorithm for one can be converted directly to an algorithm for the other with essentially no change in the running time. Our interest in these problems derives from the following results:

THEOREM 2.1. *Let $S = (T_I \cup T_O, <_I \cup <_O)$ be an opposing forest task system, let m be a number of processors, and let D be a deadline. Then, if there is an m -processor schedule for S meeting deadline D , there is also such a schedule that meets a monotone internal profile*

$$\bar{m} = (m_0, m_1, \dots, m_{D-1}) \quad \text{with } m_i \leq m \quad \text{for } 0 \leq i \leq D-1.$$

Proof. For any profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ with $m_i \leq m$ for $0 \leq i \leq D-1$, define the *monotonicity measure* $f(\bar{m})$ by

$$f(\bar{m}) = \sum_{i=0}^{D-1} (D-i)m_i.$$

For any schedule σ for S that meets deadline D , define $f(\sigma)$ by

$$f(\sigma) = \max \{f(\bar{m}) : \sigma \text{ meets internal profile } \bar{m}\}.$$

Since $f(\sigma)$ cannot exceed $D \cdot |T_I \cup T_O|$, there must be some schedule σ^* that maximizes $f(\sigma)$ and some internal profile $\bar{m}^* = (m_0^*, m_1^*, \dots, m_{D-1}^*)$ met by σ^* for which $f(\sigma^*) = f(\bar{m}^*)$. We claim that σ^* and \bar{m}^* are the desired schedule and internal profile.

Suppose not. Then for some i , $0 \leq i < D - 1$, we must have $m_i^* < m_{i+1}^*$. Consider the least such value of i . By the definition of \bar{m}^* , we know that S_O must actually have the profile $(\bar{m}^*)^C$, so the number of tasks from T_O that start at time i must be exactly $m - m_i^*$ and must exceed the number $m - m_{i+1}^*$ of tasks from T_O that start at time $i + 1$. Since S_O is an out-forest task system, it follows immediately that some task $T_j \in T_O$ with $\sigma^*(T_j) = i$ must have no successors scheduled to start at time $i + 1$. If σ^* assigns fewer than m tasks to start at time $i + 1$, we could thus reassign T_j to start at time $i + 1$ and obtain a new schedule σ satisfying $f(\sigma) = f(\sigma^*) + 1$. On the other hand, if σ^* assigns exactly m tasks to start at time $i + 1$, then there are exactly m_{i+1}^* tasks from T_I that start at time $i + 1$ and at most $m_i^* < m_{i+1}^*$ tasks from T_I that start at time i . Thus, since S_I is an in-tree task system, there must be some task $T_k \in T_I$ with $\sigma^*(T_k) = i + 1$ that has no predecessors starting at time i . In this case we can interchange the tasks T_k and T_j to form a new schedule σ with $f(\sigma) = f(\sigma^*) + 1$. Therefore, in both cases we obtain a contradiction to the maximality of $f(\sigma^*)$, and the theorem is proved. \square

The following lemma is useful for refining Theorem 2.1:

LEMMA 2.1. *Suppose that $S = (T, <)$ is an in-forest task system, σ is a schedule for S meeting the monotone profile $\bar{m} = (m_0, m_1, \dots, m_q)$ where $m_q > 0$, and k is an index satisfying $0 \leq k < q$ such that $m_k > m_{k+1}$. Then there exists a schedule σ' for S that meets the monotone profile $\bar{m}' = (m_0, m_1, \dots, m_{q+1})$ where*

$$m'_i = \begin{cases} m_i & \text{if } 0 \leq i < k \text{ or } k < i \leq q, \\ m_i - 1 & \text{if } i = k, \\ 1 & \text{if } i = q + 1. \end{cases}$$

Proof. See Fig. 1. We first construct a sequence of tasks as follows. Let T_0 be any task with $\sigma(T_0) = k$. Inductively, suppose that we have just chosen a task T_i such

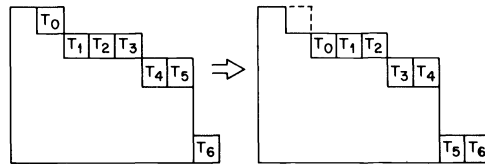


FIG. 1. Schedule transformation for Lemma 2.1.

that $\sigma(T_i) = k + i$. If $k + i < q$, consider the set of tasks scheduled to start at time $k + i + 1$. If this set is empty, set $T_{i+1} = \emptyset$. If the set contains a successor of T_i (there can be at most one, since S is an in-forest), let that successor be T_{i+1} . Otherwise, choose any task from the set to be T_{i+1} .

This procedure results in a sequence T_0, T_1, \dots, T_{q-k} of tasks such that each T_i satisfies $\sigma(T_i) = k + i$ and has no successors, except possibly T_{i+1} , scheduled to start at time $k + i + 1$. Thus our desired schedule σ' can be obtained from σ by adding 1

to the starting time for each task T_i in our sequence and leaving the starting times for all other tasks exactly as they were. \square

THEOREM 2.2. *Let $S = (T_I \cup T_O, <_I \cup <_O)$ be an opposing forest task system, let m be a number of processors, and let D be a deadline. If there is an m -processor schedule for S that meets deadline D , then there is such a schedule that also meets a monotone internal profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ for which either $m_0 < m$ or $m_{D-1} > 0$.*

Proof. Suppose S has an m -processor schedule meeting deadline D . By Theorem 2.1, there is such a schedule σ that meets a monotone internal profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$. If $m_0 < m$ or $m_{D-1} > 0$, we are done. If not, apply the transformation of Lemma 2.1 to the schedule obtained by restricting σ to T_I with k equal to the largest index i for which $m_i = m$, and simultaneously apply the out-forest analogue of Lemma 2.1 to the schedule obtained by restricting σ to T_O with k equal to the largest index j for which $m_j = 0$. The combination of the two resulting schedules is a schedule for S that meets a new monotone internal profile \bar{m}' identical to \bar{m} except that $m'_i = m - 1$ and $m'_j = 1$. Thus repeated application of this operation will yield a schedule of the desired form. \square

Theorem 2.2 leads directly to the following algorithmic result:

COROLLARY 2.2.1. *For any fixed number m of processors, if there exists an algorithm A that solves the m -processor in-forest MPS problem in time $P(n, m, D)$, then there exists an algorithm that solves the m -processor opposing forest MS problem in time $O(D^{m-1}P(n, m, D))$.*

Proof. Theorem 2.2 tells us that we can restrict our attention to schedules for the opposing forest that have an internal profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ that is monotone and that satisfies either $m_0 < m$ or $m_{D-1} > 0$. For each such profile, we can check whether there exists a schedule with that internal profile by applying A to the in-forest with the specified profile and to the reversed out-forest with the reversed complementary profile. The number of internal profiles that must be considered can be counted by first observing that each corresponds to a unique sequence n_m, n_{m-1}, \dots, n_0 where

$$n_i = |\{j: 0 \leq j \leq D-1 \text{ and } m_j = i\}|$$

and where either $n_m = 0$ or $n_0 = 0$. Furthermore, since we can restrict attention to schedules in which the out-forest has the complementary profile, the value of n_0 or n_m , whichever is nonzero, is uniquely determined by the other n_i and the number of tasks in the out-forest. Thus there are at most $2 \cdot D^{m-1}$ such sequences (profiles) that need to be considered, and we need only apply algorithm A $4 \cdot D^{m-1}$ times, from which the result follows. \square

Corollary 2.2.1 indicates that if in-forest MPS is solvable in polynomial time for a fixed number m of processors, then opposing forest MS is solvable in polynomial time for the same fixed number m of processors (we can assume that $D \leq n$). The converse need not hold, although it is not difficult to show that if opposing forest MS can be solved in polynomial time for $m+1$ processors, then in-forest MPS can be solved in polynomial time for m processors. Note also that Corollary 2.2.1 does not imply that if in-forest MPS is solvable in polynomial time for arbitrary m , then so is opposing forest MS, because of the D^{m-1} factor in the time complexity. This last issue is moot, however, (assuming $P \neq NP$ [11]) in light of the following theorem:

THEOREM 2.3. *If the number m of processors is arbitrary, the following problems are all NP-complete:*

- (1) PROFILE SCHEDULING for in-forest task systems;
- (2) MONOTONE PROFILE SCHEDULING for in-forest task systems;
- (3) MULTIPROCESSOR SCHEDULING for opposing forests.

Proof. Since all these problems are easily seen to be in NP, our proof need only consist of a series of polynomial transformations. The first transforms the known NP-complete problem 3-PARTITION into problem (1). We then transform (1) into (2) and (2) into (3), completing the proof. In what follows, we use the symbol “ α ” to denote “transforms to.” The 3-PARTITION problem is defined as follows:

3-PARTITION

Instance. Set A of $3q$ elements, a positive integer weight $w(a)$ for each $a \in A$, and a positive integer bound B , where $B/4 < w(a) < B/2$ for each $a \in A$ and $\sum_{a \in A} w(a) = qB$.

Question. Can A be partitioned into q disjoint sets A_1, A_2, \dots, A_q such that, for $1 \leq i \leq q$, $\sum_{a \in A_i} w(a) = B$?

Note that the constraints on the weights force the sets A_i to contain exactly three elements each. This problem is NP-complete in the strong sense [7], [10], [11], and hence in using it as the source problem for an NP-completeness proof we are allowed to use “pseudo-polynomial” transformations [10], [11], i.e., our transformations and instance sizes need only be polynomial in terms of q and B (rather than q and $\log B$). We proceed as follows:

(A) 3-PARTITION α *in-forest* PS. Suppose we are given an instance of 3-PARTITION specified by A, q, B , and w . Without loss of generality we may assume that B is even and exceeds $3q$ (otherwise we could simply multiply B and all values $w(a)$ by $2q$ without changing the answer for that instance). We construct an in-forest instance of PS as indicated in Fig. 2.

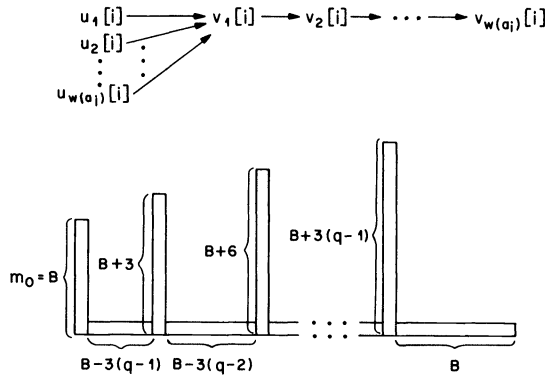


FIG. 2. The constructed instance of in-forest PS in the proof of Theorem 2.3.

The set T is the union of $3q$ sets $T[i]$, one for each element $a_i \in A$. $T[i]$ consists of $2w(a_i)$ tasks, $u_j[i]$ and $v_j[i]$, $1 \leq j \leq w(a_i)$. The partial order on these tasks is specified by:

$$u_j[i] < v_k[i] \quad \text{for } 1 \leq j, k \leq w(a_i),$$

$$v_j[i] < v_k[i] \quad \text{for } 1 \leq j < k \leq w(a_i).$$

Note that each $T[i]$ is an in-tree and hence T is an in-forest.

The deadline D is given by

$$D = q + \sum_{i=0}^{q-1} (B - 3i) = qB - \frac{3q^2 - 5q}{2}.$$

To specify the profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$, we first define certain special times $t_j, 1 \leq j \leq q + 1$, by

$$t_j = (j - 1)B - 3(j - 1)q + \frac{3j^2 - j - 2}{2}.$$

Note that $t_1 = 0$ and $t_{q+1} = D$. For $0 \leq i \leq D - 1$, we set

$$m_i = \begin{cases} B + 3(j - 1) & \text{if } i = t_j \text{ for } 1 \leq j \leq q, \\ 1 & \text{otherwise.} \end{cases}$$

Observe that

$$\sum_{i=0}^{D-1} m_i = D - q + \sum_{j=0}^{q-1} (B + 3j) = 2qB = |T|,$$

and hence any schedule for $S = (T, <)$ that meets profile \bar{m} must in fact *have* profile \bar{m} .

This completes our description of the constructed instance of in-forest PS, which is easily seen to be constructible in pseudo-polynomial time. It remains for us to show that there exists a schedule for S that meets profile \bar{m} if and only if the desired partition exists for A .

First, suppose the desired partition for A exists, and without loss of generality assume that $A_i = \{a_{3i-2}, a_{3i-1}, a_{3i}\}, 1 \leq i \leq q$, is such a partition. We show how to derive a corresponding schedule for S that meets profile \bar{m} .

The tasks of $T[1], T[2]$, and $T[3]$ are scheduled as follows: For $i \in \{1, 2, 3\}$, all tasks $u_j[i], 1 \leq j \leq w(a_i)$, are scheduled at time $t_1 = 0$. Since $m_0 = B$ and $w(a_1) + w(a_2) + w(a_3) = B$, these tasks completely fill up this slot in the profile. The tasks $v_j[i], 1 \leq j \leq w(a_i) - (q - 1)$, are scheduled at times *between* t_1 and t_2 . Since there are $B - 3(q - 1)$ such tasks and $B - 3(q - 1)$ time slots with room for one task each, these use up the remaining slots that occur before t_2 . Finally, for $0 \leq k < q - 1$ and $i \in \{1, 2, 3\}$, we schedule each task $v_{w(a_i)-k}[i]$ at time t_{q-k} . Figure 3 illustrates this ‘‘partial’’ schedule.

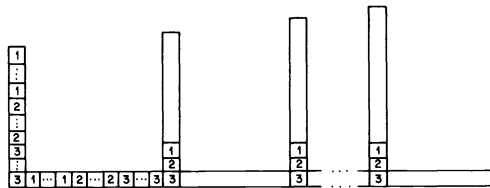


FIG. 3. Scheduling of tasks from $T[1], T[2]$, and $T[3]$.

In general, the tasks in $T[3i - 2], T[3i - 1], T[3i]$ are scheduled analogously. All the $u_k[j]$ tasks, for $j \in \{3i - 2, 3i - 1, 3i\}$ and $1 \leq k \leq w(a_j)$, are scheduled at time t_i , which has precisely enough room for all these tasks plus one representative from each $T[l], 1 \leq l \leq 3(i - 1)$. The tasks $v_k[j], j \in \{3i - 2, 3i - 1, 3i\}$ and $1 \leq k \leq w(a_j) - (q - i)$, are scheduled between t_i and t_{i+1} and exactly fill up the $B - 3(q - i)$ slots there. Finally, for $j \in \{3i - 2, 3i - 1, 3i\}$ and $0 \leq k < q - i$, we schedule each task $v_{w(a_j)-k}[j]$ at time t_{q-k} .

We leave to the reader the straightforward verification that the above description yields a schedule for S that meets profile \bar{m} .

Now suppose that there exists a schedule σ for S that meets, and consequently has, profile \bar{m} . We shall show that this implies the existence of the desired partition for A .

Consider the set of tasks scheduled at time t_q . We first claim that this set can include tasks of the form $u_k[i]$ for at most three distinct values of i . If there were more, then there would be at least four sets of tasks $\{v_k[i]: 1 \leq k \leq w(a_i)\}$ whose members were all scheduled at or following time $t_q + 1$. However, since $w(a_i) > B/4$ for all $a_i \in A$, these sets contain at least $B + 1$ tasks in total, and the time slots after $t_q + 1$ in the profile have room for at most B tasks, a contradiction to the assumption that the schedule has profile \bar{m} .

Next we claim that tasks $u_k[i]$ for *at least* three distinct values of i are scheduled at time t_q . Suppose there were only two. Then the total number of such tasks at t_q could be at most $B - 2$, since $w(a_i) < B/2$ for each $a_i \in A$ and B is even. This leaves room for at least $3(q - 1) + 2$ additional tasks in that time slot, but there can be at most $3(q - 1) + 1$ tasks of type $v_k[i]$ at that time, at most one for each value of i and none for the two values of i for which $u_k[i]$ tasks are scheduled at t_q . Thus again we have a contradiction to the assumption that the schedule has profile \bar{m} . The case in which one or fewer classes of $u_k[i]$ tasks are scheduled at t_q is handled similarly.

Thus there are precisely three indices i such that tasks $u_k[i]$ are scheduled at time t_q . This means that at most $3(q - 1)$ tasks of type $v_k[i]$ can be scheduled at this time, so there must be at least B $u_k[i]$ tasks scheduled at t_q . Let A_q be the set of a_i with the corresponding three indices. We thus have that $\sum_{a_i \in A_q} w(a_i) \geq B$. However, since these tasks have a total of $\sum_{a_i \in A_q} w(a_i)$ successors and since there is room for only B tasks after time t_q , we also must have $\sum_{a_i \in A_q} w(a_i) \leq B$. Thus $\sum_{a_i \in A_q} w(a_i) = B$. This in turn implies that the task $v_{w(a_i)}[i]$ for each $a_i \notin A_q$ also must be scheduled at time t_q .

We proceed by induction. Suppose that in general we define

$$A_j = \{a_i: \text{for some } k, 1 \leq k \leq w(a_i), \sigma(u_k[i]) = t_j\}$$

and $A'_j = A_j \cup A_{j+1} \cup \dots \cup A_q$. When $j = q$, we know that the sets A_l , $j \leq l \leq q$, form a partition of A'_j into three-element sets such that the weights of the elements in each sum to B . We also know that the slots in the profile starting with time t_q are completely filled with the tasks

$$\{u_k[i], v_k[i]: a_i \in A'_j, 1 \leq k \leq w(a_i)\}$$

and

$$\{v_k[i]: a_i \notin A'_j \text{ and } w(a_i) - (q - j) \leq k \leq w(a_i)\}.$$

Suppose this is true for a given value $j = J + 1 \leq q$. An analogue of the above argument for $j = q$ can be used to show that $A_J \cap A'_{J+1} = \emptyset$, $|A_J| = 3$, $\sum_{a \in A_J} w(a) = B$, and the space in the profile starting with t_J and extending through time $t_{J+1} - 1$ is completely filled with the tasks from the sets $T[i]$ for $a_i \in A_J$ that are not scheduled later than $t_{J+1} - 1$, plus $v_{w(a_i) - q + J}[i]$ for each $a_i \notin A'_J = A_J \cup A'_{J+1}$. This implies that the induction hypothesis holds for $j = J$, and we can conclude that it holds for $j = 1$. Hence, the constructed sets A_1, A_2, \dots, A_q form the desired partition of A .

It follows from the above that in-forest PS is NP-complete. Our remaining transformations are considerably simpler.

(B) *In-forest* PS α *in-forest* MPS. Let $S = (T, <)$ and $\bar{m} = (m_0, m_1, \dots, m_{D-1})$

specify an instance of in-forest PS. Define

$$M_0 = 1 + \sum_{i=0}^{D-1} m_i,$$

$$M_j = M_0 - \sum_{i=0}^{j-1} m_i = 1 + \sum_{i=j}^{D-1} m_i, \quad 1 \leq j \leq D-1.$$

Note that we then have

$$M_0 \geq M_0 - m_0 = M_1 \geq M_1 - m_1 = M_2 \geq \cdots \geq M_{D-1} \geq M_{D-1} - m_{D-1} = 1.$$

We now define an in-tree $S' = (T', <')$ as follows:

$$T' = \{x_i[k] : 0 \leq k \leq D-1, 1 \leq i \leq M_k - m_k\},$$

$$x_i[k] <' x_j[l] \Leftrightarrow \{0 \leq k < l \leq D-1 \text{ and either } i = j \text{ or } j = M_l - m_l \text{ and } i > M_l - m_l\}.$$

This tree is illustrated in Fig. 4. Notice that any schedule σ for S' must have makespan at least D , and, if it has makespan exactly D , it must have $\sigma(x_i[k]) = k$ for all $x_i[k] \in T'$ and thus *have profile* $(M_0 - m_0, M_1 - m_1, \dots, M_{D-1} - m_{D-1})$.

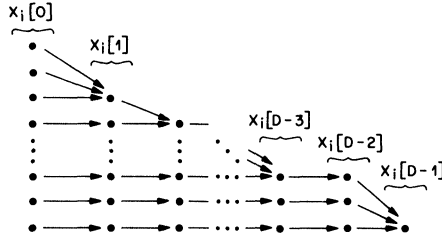


FIG. 4. The in-tree task system S' .

The desired instance of in-forest MPS is simply the task system $(T \cup T', < \cup <')$ and the profile $\vec{M} = (M_0, M_1, \dots, M_{D-1})$. It is easy to see that this combined task system can be scheduled to meet the monotone profile \vec{M} if and only if S can be scheduled to meet the original profile \vec{m} : Since any schedule for the combined task system that meets \vec{M} must satisfy $\sigma(x_i[k]) = k$ for all $x_i[k] \in T'$, the restriction of that schedule to tasks in T must be a schedule for S that meets \vec{m} . Similarly, any schedule σ for S that meets \vec{m} can be extended to a schedule for the combined system that meets \vec{M} by setting $\sigma(x_i[k]) = k$ for all $x_i[k] \in T'$.

It is trivial to construct the above instance in polynomial time, so we have, as required, shown that in-forest PS α in-forest MPS and hence the latter is NP-complete. Our final transformation is quite similar.

(C) *In-forest MPS α opposing forest MS.* Let $S = (T, <)$ and $\vec{m} = (m_0, m_1, \dots, m_{D-1})$ specify an instance of in-forest MPS. We construct an out-tree $S' = (T', <')$ as follows:

$$T' = \{y_i[k] : 0 \leq k \leq D-1, 1 \leq i \leq m_0 - m_k + 1\},$$

$$y_i[k] <' y_j[l]$$

$$\Leftrightarrow 0 \leq k < l \leq D-1 \text{ and either } i = j \text{ or } i = m_0 - m_k + 1 \text{ and } j > m_0 - m_k + 1.$$

Notice that this tree has the form of the tree in Fig. 4, but with all the arrows reversed. If T' is scheduled to have makespan D , it must be the case that $\sigma(y_i[k]) = k$ for each

$y_i[k] \in T'$, and hence the schedule must have the (monotone) profile $(m_0 - m_0 + 1, m_0 - m_1 + 1, \dots, m_0 - m_{D-1} + 1)$. From this it is trivial to conclude that the opposing forest task system $(T \cup T', < \cup <')$ can be scheduled on $m_0 + 1$ processors with makespan D if and only if S can be scheduled to meet profile \bar{m} . Thus we have our desired polynomial transformation from in-forest MPS to opposing forest MS, and it follows that opposing forest MS is NP-complete. \square

Having proved that opposing forest MS is NP-complete when the number of processors is arbitrary, we proceed in the next section to the case where m is fixed. We devise an algorithm for in-forest MPS that runs in polynomial time for any fixed value of m , which, by Corollary 2.2.1, yields a polynomial time algorithm for m -processor opposing forest MS for each fixed value of m .

3. A monotone profile algorithm for in-forests. Our algorithm is of the “divide and conquer” variety. Due to the manner in which we divide into subproblems, all the subproblems that arise will be of one of the two following forms:

Case 1. $|T| = \sum_{i=0}^{D-1} m_i$. In this case we say that a schedule σ *satisfies* the profile \bar{m} if σ maps T into the set $\{0, 1, \dots, D-1\}$ of starting times, it observes the precedence constraints, and it schedules exactly m_i tasks to start at each time i , $0 \leq i \leq D-1$. Note that such a schedule will in fact *have* the profile \bar{m} .

Case 2. $|T| \geq \sum_{i=0}^{D-1} m_i + m_{D-1}$. In this case we say that a schedule σ *satisfies* the profile \bar{m} if σ maps T into the set $\{0, 1, \dots, D\}$ of starting times, it observes the precedence constraints, and it schedules exactly m_i tasks to start at each time i , $0 \leq i \leq D-1$, with the remainder scheduled to start at time D . We can view such a schedule σ as one mapping a subset of T into the starting times $\{0, 1, \dots, D-1\}$, along with a collection of “unscheduled” tasks (those with $\sigma(T_i) = D$), with the restriction that no unscheduled tasks have any successors in T .

In what follows, we shall assume that each task system under consideration falls under either Case 1 or Case 2, and that \bar{m} is a monotone in-forest profile. At the end of the section, we will discuss the conversion of an arbitrary instance of in-forest MPS to the required form.

We will be using the level algorithm of [12] as a basic subroutine in our algorithm. In terms of the MPS problem, this algorithm can be described as follows:

Given an in-forest $S = (T, <)$, the *level* $l(T_i)$ of a task $T_i \in T$ is defined to be 0 if T has no successors and otherwise is given by

$$1 + \max \{l(T_j) : T_j \in T \text{ and } T_i < T_j\}.$$

The level algorithm first reorders the tasks so that $l(T_1) \geq l(T_2) \geq \dots \geq l(T_n)$. It then schedules the tasks by time slot, first choosing the tasks to be executed at time 0, then those to be scheduled at time 1, and so on, always scheduling as many tasks as possible and giving preference to tasks with lower indices. Specifically, in choosing the tasks to start at time t , the algorithm chooses from the set

$$A_t = \{T_i \in T : T_i \text{ has not been scheduled to start before } t, \\ \text{and all predecessors of } T_i \text{ have been scheduled to start before } t\}$$

the m_t tasks with lowest indices (and, hence, highest levels), or, if $|A_t| < m_t$, it assigns *all* tasks in A_t to start at t . This scheduling process continues, with t increased by 1 at each step, until all tasks in T have been assigned starting times. (To handle subproblems falling under Case 2, we set $m_D = \infty$ by convention.)

We now give four lemmas pertaining to the level algorithm, as it applies to problems of types (a) and (b). These will be used later to justify the way in which our

overall algorithm works. The first of these, which we state without proof, is merely an observation about in-forests that will be useful in our subsequent proofs. A *leaf* in an in-forest task system is a task that has no predecessors.

LEMMA 3.1. *Suppose $S = (T, <)$ is an in-forest with exactly k leaves. If any subset of these leaves is deleted from S , then the resulting “pruned” in-forest will have no more than k leaves. If all k of these leaves are deleted, then the maximum level in T is reduced by 1.*

Our second lemma characterizes a class of profiles for which the level algorithm is guaranteed to find satisfying schedules whenever they exist. For a given profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$, define M_i to be the number of tasks that can be scheduled in the first $i + 1$ time slots, i.e., for $0 \leq i \leq D - 1$, $M_i = m_0 + m_1 + \dots + m_i$, with $M_i = 0$ if $i < 0$ and $M_D = \infty$ by convention. For an in-forest task system, define l_i to be the number of tasks with level i , define L_i to be the number of tasks with level i or greater, and let $L_{-1} = \infty$ by convention.

LEMMA 3.2. *Suppose $S = (T, <)$ is an in-forest task system and $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ is a monotone profile satisfying $m_0 - m_{D-1} \leq 1$. Then the following are equivalent:*

- (i) *There is a schedule for S satisfying \bar{m} .*
- (ii) *For all $i \geq 0$, $L_i \leq M_{D-i-1}$ in Case 1, or $L_i \leq M_{D-i}$ in Case 2.*
- (iii) *For any ordering of the tasks by nonincreasing level, the level algorithm produces a schedule for S satisfying \bar{m} .*

Proof. We first note that (iii) trivially implies (i). It is also immediate that (i) implies (ii), since in any schedule that observes the precedence constraints it is necessarily the case that all tasks with level i or greater start at or before time $D - i - 1$ in Case 1 or time $D - i$ in Case 2. We shall show that (ii) implies (iii) by induction on D , thus completing the proof.

Suppose $D = 1$ and that (ii) holds for S and \bar{m} . In Case 1 this means that all tasks have level 0, and they all belong to the set A_0 of tasks available at time 0. Since there are precisely m_0 of them, by the definition of Case 1, the level algorithm will schedule them all at time 0, thus satisfying the profile \bar{m} . In Case 2 there must be at least $2m_0$ tasks, with at most m_0 having level 1 by (ii) and the rest having level 0. Since S is an in-forest, it follows that the set A_0 of leaves must consist of at least m_0 tasks, including all the level-1 tasks. Hence the level algorithm will schedule m_0 tasks, including all the level-1 tasks, at time 0, and the schedule will satisfy the profile \bar{m} .

Now suppose that (ii) implies (iii) for all values of $D < d$ and let S and \bar{m} satisfy (ii) for $D = d$. We first claim that S has at least m_0 leaves. Suppose that S had no more than $m_0 - 1$ leaves. By Lemma 3.1, we know that removing all the leaves from an in-forest with at most $m_0 - 1$ leaves yields a pruned in-forest with no more than $m_0 - 1$ leaves and with the maximum task level reduced by exactly 1. Thus, if we remove all the leaves from S , and then remove all the leaves from the resulting in-forest, and continue to do this until the set of tasks becomes empty, we will remove at most $m_0 - 1$ tasks at each step and the number of steps will be one greater than the maximum task level in S . By (ii), the maximum task level in Case 1 is at most $D - 1$ and in Case 2 is at most D . Thus the number of tasks in T must satisfy

$$|T| \leq (m_0 - 1) \cdot D \leq \sum_{i=0}^{D-1} m_i - 1$$

in Case 1 and

$$|T| \leq (m_0 - 1) \cdot (D + 1) \leq \sum_{i=0}^{D-1} m_i + m_{D-1} - 1$$

in Case 2. Since each of these inequalities contradicts the definition of the corresponding case, it follows that S must have at least m_0 leaves.

It follows that the set A_0 of tasks initially available to be scheduled by the level algorithm satisfies $|A| \geq m_0$, and the level algorithm will choose exactly m_0 of these tasks to start at time 0. Let T' be the set of remaining tasks, $S' = (T', <')$ be the induced task system, and $\bar{m}' = (m'_0, m'_1, \dots, m'_{D-2})$ be the remaining profile, where $m'_i = m_{i+1}$ for $0 \leq i \leq D-2$. Define M'_i and L'_i in terms of the new task system, profile, and deadline $D' = D - 1$. It is easy to see that S' and \bar{m}' satisfy the hypotheses of the lemma and that whichever of Cases 1 and 2 held for S and \bar{m} continues to hold for S' and \bar{m}' . We shall argue that (ii) must continue to hold for S' and \bar{m}' , and hence, by the inductive hypothesis, the level algorithm will construct a schedule for S' satisfying \bar{m}' . Since this is exactly what the level algorithm will do in scheduling S after the first m_0 tasks have been scheduled, it will follow that the level algorithm constructs a schedule for S satisfying \bar{m} , and the lemma will follow by induction.

Let k be the least level (in S) of any task in $T - T'$. Property (ii) will certainly hold for S' and \bar{m}' for all $i \leq k$, since for such i we have $L'_i = L_i - m_0$, $M'_{D-1-i} = M_{D-1-i} - m_0$, and $M'_{D-i} = M_{D-i} - m_0$, in Cases 1 and 2 respectively. Consider any $i > k$. By the definition of k and the operation of the level algorithm, we know that there were at most $m_0 - 1$ tasks of level greater than k available at time 0. Since S is an in-forest, this implies that for each $j \geq i$ there can be at most $m_0 - 1$ tasks of level exactly j in S , i.e., $l_j \leq m_0 - 1$ for all $j \geq i$. In Case 1 the maximum task level in S is at most $D - 1$ and hence the maximum task level in S' is at most $D - 2$. Thus we have

$$L'_i = \sum_{j=i}^{D-2} l'_j \leq \sum_{j=i}^{D-2} l_j \leq (m_0 - 1)(D - i - 1) \leq M'_{D-i-2},$$

as desired. Similarly in Case 2 the maximum level for a task in S' is at most $D - 1$, and hence we have

$$L'_i = \sum_{j=i}^{D-1} l'_j \leq \sum_{j=i}^{D-1} l_j \leq (m_0 - 1)(D - i) \leq M'_{D-i-1},$$

as desired. Thus (ii) holds for S' and \bar{m}' , and the fact that (ii) implies (iii) for all values of D follows by induction. \square

Figure 5 gives an example of a task system that can be scheduled to satisfy the monotone profile $(3, 3, 1, 1, 1)$, but for which the level algorithm fails to find a satisfying schedule, thus showing that Lemma 3.2 cannot be extended to more complicated profiles.

The remaining two lemmas are needed for specific technical reasons that will become clear later.

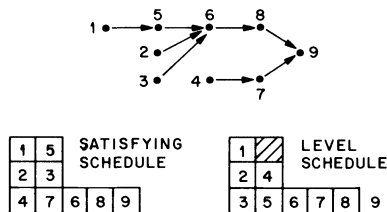


FIG. 5. Task system for which the level algorithm fails to find a satisfying schedule.

LEMMA 3.3. Suppose $S = (T, <)$ is an in-forest task system, $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ is a monotone in-forest profile satisfying $m_0 = m_{D-1}$, and there is a schedule for S satisfying \bar{m} . If $S' = (T', <')$ differs from S only in that a leaf has been deleted from S and replaced in S' by a leaf of no greater level, then there is a schedule for S' satisfying \bar{m} .

Proof. This follows immediately from Lemma 3.2, since such a change in S will yield $L'_i \leq L_i$ for all $i \geq 0$. \square

LEMMA 3.4. Suppose $S = (T, <)$ is an in-forest task system with $|T| = mD + 1$, $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ is a monotone in-forest profile satisfying $m_0 = m_{D-1} = m$, and T_1 and T_2 are two level-0 tasks in T . Let $S_1 = (T - \{T_1\}, <_1)$ and $S_2 = (T - \{T_2\}, <_2)$ be the induced task systems obtained by deleting T_1 and deleting T_2 respectively. If there is a level schedule σ for S_1 that satisfies \bar{m} and that starts all predecessors of T_1 before time $D - 1$, then S_2 can be scheduled to meet \bar{m} .

Proof. Let us consider the hypothetical schedule σ for S_1 . If $\sigma(T_2) = D - 1$, we can obtain the desired schedule for S_2 simply by replacing T_2 with T_1 . If $\sigma(T_2) = i < D - 1$, then there were at most $m - 1$ tasks of level greater than 0 available at time i and hence at most that many available at any later time. Since σ schedules exactly m tasks in each time slot, this means that at each time j , $i \leq j \leq D - 1$, there must be at least one level 0 task. Because S is an in-forest, this implies that at each time j , $i < j \leq D - 1$, there must be some task T'_j that has no predecessors scheduled at time $j - 1$. Thus if we delete T_2 and reschedule each T'_j to start at time $j - 1$ instead of j , we can then schedule T_1 at time $D - 1$ and obtain a schedule for S_2 satisfying \bar{m} . \square

Our first application of the above lemmas will be in proving that any schedule for an in-forest S satisfying a monotone profile \bar{m} can be assumed to have certain structural properties, that is, can be assumed to take on a certain “normal form.”

Any satisfying schedule σ can be partitioned into blocks B_1, B_2, \dots, B_r , with one block B_i for each distinct value h_i taken on by the m_j 's, $0 \leq j \leq D - 1$, where block B_i consists of the w_i consecutive time slots for which $m_j = h_i$ and the blocks are indexed so that $h_1 > h_2 > \dots > h_r$. See Fig. 6. Note that block B_i must contain exactly $h_i w_i$ tasks, since σ satisfies \bar{m} . Any tasks from T that are not included in these blocks (those with $\sigma(T_i) = D$ in Case 2) will be ignored for the moment.

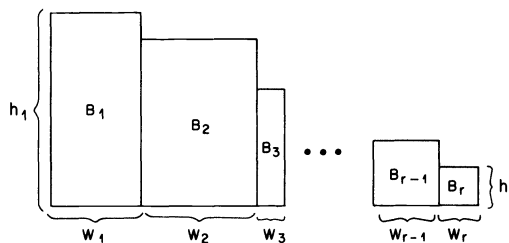


FIG. 6. Block structure of a schedule σ for S satisfying the profile \bar{m} .

For each task T_k in block B_i and each j , $i \leq j \leq r$, define $l_j(T_k)$ to be the level that T_k has in the in-forest obtained by restricting S to only those tasks in blocks B_1, B_2, \dots, B_j . Note, in particular, that $l_r(T_k) = l(T_k)$ for all tasks T_k . Define $\bar{l}_j(T_k)$ to be the vector

$$(l_j(T_k), l_{j+1}(T_k), \dots, l_r(T_k)).$$

For any fixed value of j , we order these vectors lexicographically, i.e., $\bar{l}_j(T_k) < \bar{l}_j(T'_k)$ if and only if either $l_j(T_k) < l_j(T'_k)$ or there is some integer q such that $l_j(T_k) = l_j(T'_k)$, $l_{j+1}(T_k) = l_{j+1}(T'_k)$, \dots , $l_{q-1}(T_k) = l_{q-1}(T'_k)$, and $l_q(T_k) < l_q(T'_k)$. We shall show that we may restrict our attention to schedules that have the following two normal form properties:

Property A. Within each block B_i , the tasks are scheduled according to the level algorithm, with the task ordering determined by $\bar{l}_i(\cdot)$. (Note that this merely says that tasks are ordered by $l_i(\cdot)$, with the values of $l_j(\cdot)$ for $j > i$ used to break ties.)

Property B. If T_k is a task in block B_j whose predecessors are all scheduled before the last time slot in block B_i , where $1 \leq i < j \leq r$, then $\bar{l}_j(T_k) \leq \bar{l}_j(T'_k)$ for all tasks T'_k in block B_i .

We shall say that a schedule is *in normal form* if it has these two properties.

THEOREM 3.1. *If there exists a schedule for the in-forest $S = (T, <)$ that satisfies the monotone profile \bar{m} , then there exists such a schedule which is also in normal form.*

Proof. For Property A, we observe that by Lemma 3.2, since the tasks in each block B_i can be scheduled to satisfy the profile (h_i, h_i, \dots, h_i) for block B_i , any level schedule for the in-forest induced by those tasks will also satisfy this profile. In particular, the level schedule obtained when those tasks are ordered by $\bar{l}_i(\cdot)$ will satisfy this profile.

The argument for Property B is somewhat more complicated. For a task T_k in a block B_i , $1 \leq i \leq r$, define the *vector level* of T_k to be $\bar{l}_i(T_k)$. Define the *characteristic vector* of a schedule σ satisfying profile \bar{m} to be a vector composed of the vector levels of all the tasks, with the vector levels for tasks in block B_j occurring before those for tasks in block B_i whenever $j > i$ and with the vector levels for tasks belonging to the same block occurring in nonincreasing lexicographic order. We claim that a schedule σ satisfying \bar{m} that achieves the lexicographically *minimum* characteristic vector must have Property B (and hence can be converted to such a schedule that also has Property A, since the above conversion to Property A does not affect the characteristic vector).

Let σ be a schedule satisfying \bar{m} that achieves the minimum possible characteristic vector, and suppose that σ does not have Property B. Then there is a block B_j and a task T_k in B_j for which Property B is violated. Choose the largest $i < j$ such that Property B fails for B_i , B_j , and T_k , and let T'_k be the latest scheduled task in B_i for which $\bar{l}_j(T'_k) < \bar{l}_j(T_k)$. By these choices, no successor of T'_k can be scheduled during B_i , B_{i+1} , \dots , B_{j-1} , or that successor would be a later violator than T'_k , contradicting either the choice of i or the choice of T'_k . Thus, if all tasks outside of blocks B_i and B_j remain as scheduled, T'_k would have no successor preventing it from being scheduled in block B_j . By hypothesis, T_k can be scheduled as early as the last time slot in B_j . Thus we propose to interchange T_k and T'_k . By Lemma 3.3, the tasks in B_j can still be scheduled within the same h_j by w_j "rectangle" when T_k is replaced by T'_k , since both are leaves of the in-forest induced by themselves and the remaining tasks in B_j , and $\bar{l}_j(T'_k) < \bar{l}_j(T_k)$ implies $l_j(T'_k) \leq l_j(T_k)$. Similarly, the tasks in B_i can be rescheduled within the same h_i by w_i "rectangle" when T'_k is replaced by T_k , since the hypotheses of Lemma 3.4 are satisfied. This yields a new schedule σ' for S satisfying \bar{m} . Moreover, this does not affect the values of $\bar{l}_j(T_k)$ or $\bar{l}_j(T'_k)$, nor the vector levels of any other tasks occurring in blocks B_j , B_{j+1} , \dots , B_r . Thus the first change in the characteristic vector in going from σ to σ' occurs among the vector levels for tasks in block B_j , where $\bar{l}_j(T_k)$ is replaced by the lower value $\bar{l}_j(T'_k)$, thus yielding a lexicographically smaller characteristic vector. This, however, contradicts the initial choice of σ , so σ must have Property B. \square

The next theorem and its corollaries are consequences of Theorem 3.1 and provide the key to our divide and conquer strategy.

THEOREM 3.2. *Suppose $S = (T, <)$ is an in-forest task system, \bar{m} is a monotone profile, and σ is a normal form schedule for S satisfying \bar{m} and having blocks, B_1, B_2, \dots, B_r . Let l^* be the least value of $l_r(T_j)$ for a task T_j scheduled outside of block B_r , let B_i be the last block prior to B_r that contains a task T_j with $l_r(T_j) = l^*$, and let t_i denote the last starting time in block B_i . Then there exists a task T_i in B_i with $\sigma(T_i) = t_i$ such that $\bar{l}_i(T_i) = (0, 0, \dots, 0, l^*)$, and every task T_j which has $\sigma(T_j) > t_i$ and $l_r(T_j) > l^*$ must have a predecessor scheduled to start at time t_i .*

Proof. Let T_i be the latest task in B_i with $l_r(T_i) = l^*$. By the choice of l^* , T_i has no successors in blocks $B_i, B_{i+1}, \dots, B_{r-1}$ and hence $\bar{l}_i(T_i) = (0, 0, \dots, 0, l^*)$. Thus all tasks T_j in B_i with $l_r(T_j) > l^*$ must have $\bar{l}_i(T_j) > \bar{l}_i(T_i)$. If $\sigma(T_i) < t_i$, then the time $\sigma(T_i) + 1$ is in block B_i and must be the starting time for at least one task T_j that is not a successor of any task scheduled at time $\sigma(T_i)$, since S is an in-forest and T_i has no successors in B_i . By our choice of T_i , it must also be the case that $l_r(T_j) > l^*$. However, this implies that $\bar{l}_i(T_j) > \bar{l}_i(T_i)$, which is a violation of normal form property A, since T_j was available at time $\sigma(T_i)$ but T_i was scheduled instead. Thus T_i must be as required.

In a similar way it can be argued that the earliest-scheduled task T_j with $\sigma(T_j) > t_i$, $l_r(T_j) > l^*$, and no predecessor scheduled at time t_i must violate normal form Property B with respect to B_i and T_i , and hence no such tasks can exist. \square

COROLLARY 3.2.1. *Suppose there exists a schedule σ for an in-forest task system $S = (T, <)$ satisfying a monotone profile \bar{m} with $r \geq 2$ blocks. Then there exist integers l^* and i , $0 \leq l^* \leq |T|$ and $1 \leq i \leq r$, and a set $T' \subseteq T$ with $|T'| = h_i$, such that if we define*

$$U = T' \cup \{T_j \in T : l(T_j) > l^* \text{ and } T_j \text{ has no predecessor in } T'\},$$

$$W = \{T_j \in T : l(T_j) = l^* \text{ and } T_j \text{ has no predecessor in } T'\},$$

$$V = \{T_j \in T : l(T_j) < l^* \text{ or } T_j \text{ has a predecessor in } T'\},$$

$$\bar{m}_1 = \text{the initial portion of } \bar{m} \text{ corresponding to the first } i \text{ blocks,}$$

$$\bar{m}_2 = \text{the remaining portion of } \bar{m} \text{ corresponding to the last } r - i \text{ blocks,}$$

$$n_0 = \sum_{j=1}^i h_j \cdot w_j,$$

then

(1) for some $W' \subseteq W$ with $|W'| = |U| + |W| - n_0$, the in-forest S_1 induced by $U \cup (W - W')$ can be scheduled to meet \bar{m}_1 , and

(2) for any $W' \subseteq W$ with $|W'| = |U| + |W| - n_0$, the in-forest S_2 induced by $V \cup W'$ can be scheduled to satisfy \bar{m}_2 .

Proof. Let σ be a normal form schedule for S satisfying \bar{m} , choose l^* and i as in Theorem 3.2, and let T' be the set of tasks scheduled by σ in the last time slot of block B_i . Claim (1) then follows by Theorem 3.2.

For claim (2), observe that Theorem 3.2 tells us that for some $W' \subseteq W$ with $|W'| = |U| + |W| - n_0$, $V \cup W'$ is scheduled by σ so that all tasks from W' go in block B_r or later. Let X be the set of tasks scheduled by σ in block B_r or later. Note that the tasks in W' are all leaves of the in-forest induced by X and all have level l^* . Furthermore, this in-forest is scheduled by σ to satisfy the profile (h_r, h_r, \dots, h_r) . Thus if W'' is any other subset of W with $|W''| = |W'|$, the in-forest induced by $(X - W') \cup W''$ can also be scheduled to satisfy the same profile, by Lemma 3.2. Since all tasks in W

have no successors except in block B_r or later, this new schedule for block B_r (and the time slot following B_r in Case 2) can be appended to the schedule of blocks B_{i+1}, \dots, B_{r-1} under σ to yield the schedule required for Claim (2). \square

COROLLARY 3.2.2. *Suppose $S = (T, <)$ is an in-forest task system, \bar{m} is a monotone profile, and there exist l^*, i and T' as in Corollary 3.2.1, with $U, V, W, \bar{m}_1, \bar{m}_2$, and n_0 defined in the same way, such that there is a $W' \subseteq W$ with $|W'| = |U| + |W| - n_0$ satisfying*

- (1) *the in-forest induced by $U \cup (W - W')$ can be scheduled to meet \bar{m}_1 , and*
- (2) *the in-forest induced by $V \cup W'$ can be scheduled to satisfy \bar{m}_2 .*

Then S can be scheduled to satisfy \bar{m} .

Proof. The schedules for the two subproblems can be juxtaposed to obtain the desired schedule for S . \square

Theorem 3.2 and its corollaries provide a divide and conquer strategy for solving the problem of scheduling an in-forest $S = (T, <)$ to satisfy a monotone profile \bar{m} with blocks B_1, B_2, \dots, B_r . We shall denote the resulting Monotone Profile Algorithm by A , for short, and let $A[S, \bar{m}]$ stand for the schedule constructed by the algorithm for S and \bar{m} , with $A[S, \bar{m}] = \phi$ if no schedule exists. (We remind the reader that we continue to assume that either Case 1 or Case 2 holds. At the end of this section, we shall observe how the algorithm can be extended to the standard MPS problem by the introduction of dummy tasks having level 0.)

MONOTONE PROFILE ALGORITHM

Step 1. If $m_0 - m_{D-1} \leq 1$, apply the level algorithm to S and \bar{m} . If the resulting schedule satisfies \bar{m} , return it and halt. If the schedule fails to satisfy \bar{m} , return ϕ and halt (no schedule exists, by Lemma 3.2).

Step 2. If $m_0 - m_{D-1} > 1$, then for all $l, 0 \leq l \leq \max \{l(T_j) : T_j \in T\}$, for all $i, 1 \leq i < r$, and for all sets $T' \subseteq T$ such that $|T'| = h_i$ and $\min \{l(T_j) : T_j \in T'\} = l$, define $U, V, W, \bar{m}_1, \bar{m}_2$, and n_0 as in Corollary 3.2.1, and do one of the following, depending on the value of n_0 :

2A. If $n_0 > |U| + |W|$ or $n_0 < |U|$, go on to the next choice for (l, i, T') .
(No schedule exists for the current choice.)

2B. If $0 < |U| + |W| - n_0 < h_i$, then for all subsets $W' \subseteq W$ with $|W'| = |U| + |W| - n_0$, do the following:

Let S_1 and S_2 be the task systems induced by $U \cup (W - W')$ and $V \cup W'$ respectively. Compute $\sigma_1 = A[S_1, \bar{m}_1]$ and $\sigma_2 = A[S_2, \bar{m}_2]$. If both σ_1 and σ_2 are satisfying schedules, return the schedule obtained by combining them and halt. If either is ϕ , go on to the next choice for W .

If all possible choices for W' have been exhausted, go on to the next choice for (l, i, T') .

2C. If $|U| + |W| - n_0 \geq h_i$, let S_1 be the task system induced by $U \cup W \cup \{T_0\}$, where T_0 is a new task with level 0 that is a successor of all tasks in U and that is placed last in the level ordering for S_1 . (Note that S_1 and \bar{m}_1 form a Case-2 instance of our problem.)

Compute $\sigma_1 = A(S_1, \bar{m}_1)$. If $\sigma_1 = \phi$, go on to the next value for (l, i, T') . If σ_1 is a satisfying schedule, let W' be the set of tasks started *after* the last block of \bar{m}_1 . (Note that we must have $W' \subseteq W \cup \{T_0\}$, since all tasks in U precede T_0 and hence have level exceeding 0. We also claim that T_0 must belong to W' . This is because T_0 is a successor of all tasks with level exceeding 0 in S_1 (all tasks in W

have level 0 in S_1). Hence, when T_0 was scheduled, all tasks with level exceeding 0 were already completed, so all tasks in $W \cup \{T_0\}$ were available. Since T_0 was last in the level ordering, it must have been the last to be scheduled, and thus it must be in W' .)

Let S_2 be the task system induced by $V \cup (W' - \{T_0\})$ and compute $\sigma_2 = A[S_2, \bar{m}_2]$. If $\sigma_2 = \phi$, go on to the next value for (l, i, T') . (Note that, if there exists a schedule σ for S satisfying \bar{m} , and if $l, i,$ and T' are the values for that schedule as in Theorem 3.2, then, no matter what choice of W' is obtained from σ_1 , we cannot have $\sigma_2 = \phi$, by Corollary 3.2.1.) If σ_2 is a satisfying schedule, return the schedule obtained by combining σ_1 , restricted to those tasks in $U \cup (W - W')$, with σ_2 and halt.

Step 3. If all choices for (l, i, T') have been exhausted without returning a schedule, return ϕ and halt.

We leave to the reader the formal verification, using Theorem 3.2 and its corollaries, that the above algorithm will find a schedule for S satisfying \bar{m} if one exists and will return ϕ otherwise. We now estimate the running time for the algorithm.

The overall time for the algorithm is dominated in the worst case by the time spent in recursive calls that enter Step 1 and apply the level algorithm. Each of these is reached by making a sequence of choices for (l, i, T') and possibly W' . For each triple (l, i, T') in such a sequence the value of i is fixed by the requirement that $h_i = |T'|$, and the same value of h_i cannot occur twice in a sequence. Each particular value of h_i might occur with any of the $n = |T|$ possible values for l and can be associated with at most $\binom{n}{h_i} \binom{n}{h_i-1}$ possibilities for the sets T' and W' , for at most n^{2h_i} possibilities. The total number of values for h_i in a particular sequence is at most $m_0 - 2$, since each choice of a triple (l, i, T') must reduce the number of blocks by at least one, and the algorithm stops when there are only two blocks left (unless the heights of the two blocks differ by more than one, in which case $r \leq m_0 - 1$). Thus the total number of sequences is at most

$$\prod_{j=3}^{m_0} n^{2j} = n^{m_0^2 + m_0 - 6}.$$

Each sequence can lead to at most two applications of the level algorithm, requiring time $O(n)$, so we obtain an overall (no doubt pessimistic) time bound of $O(n^{m_0^2 + m_0 - 5})$.

To use this algorithm to solve the in-forest MPS problem, we first compute $N = \sum_{i=0}^{D-1} m_i$. If $|T| > N$, no schedule can exist. If $|T| \leq N$, we add $N - |T|$ dummy tasks, each with level 0 and having no predecessors, to obtain a Case-1 instance of our problem. Applying the above algorithm to this instance, we will obtain a schedule σ satisfying (indeed, having) profile \bar{m} if and only if there is a schedule σ' for the original instance that meets \bar{m} , where σ' is derived from σ simply by deleting the dummy tasks. Letting $m = m_0$, the running time for this is $O(N^{m^2 + m - 5})$.

Applying Corollary 2.2.1, we can extend this algorithm for solving the in-forest MPS problem to one that solves the opposing forest MS problem in time

$$O(D^{m-1} N^{m^2 + m - 5}) = O(m^{m^2 + m - 5} D^{m^2 + 2m - 6})$$

since $N \leq mD$. Noting that we may assume $n \geq D$ (or the problem would be trivial), we obtain our final running time bound for fixed m of $O(n^{m^2 + 2m - 6})$.

Although this bound is clearly polynomial for any fixed value of m , it is by no means the sort of bound that could be used to justify any claims of having an "efficient"

algorithm. Indeed, even for $m = 3$, the bound is $O(n^9)$. However, at least for this special case, it is possible to do significantly better. In the next section we shall see how special techniques can be used to solve the $m = 3$ case of opposing forest MS in linear time. In the concluding section we will comment on the possibilities for achieving significant improvements for general m .

4. The three-processor problem. Our algorithm for the three-processor case of opposing forest MS is based on Theorem 2.2 and the following “algorithmic” analogue of Lemma 2.1 for the case of $m = 2$:

LEMMA 4.1. *Suppose that $S = (T, <)$ is an in-forest task system and σ is a schedule for S meeting the monotone profile $\bar{m} = (m_0, m_1, \dots, m_q)$, where $m_0 = 2$, $m_q > 0$, and k is the largest index such that $m_k = 2$. Then for any j , $0 \leq j \leq k$, there is a schedule σ' for S meeting the monotone profile $\bar{m}' = (m'_0, m'_1, \dots, m'_{q+k-j+1})$ where $m'_i = 2$ for $0 \leq i < j$ and $m'_i = 1$ for $j \leq i \leq q + k - j + 1$. Furthermore, given σ and j , the schedule σ' can be constructed in linear time.*

Proof. Trivial. \square

We are now prepared to analyze the opposing forest MS problem for $m = 3$. Suppose $S = (T_I \cup T_O, <_I \cup <_O)$ is an opposing forest task system and D is a desired deadline. By Theorem 2.2 we may restrict our search to schedules with monotone internal profiles $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ for which either $m_0 \leq 2$ or $m_{D-1} \geq 1$. Let us consider the latter possibility (the former will be treated symmetrically by our algorithm, and we shall omit description of the details).

Define t_1 to be the least value of j such that the in-forest $S_I = (T_I, <_I)$ can be scheduled to meet a monotone profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ with $m_0 \leq 3$ and $m_j < 3$. Define t_2 to be the minimum makespan for any two-processor schedule for the out-forest $S_O = (T_O, <_O)$. Note that t_2 , and a schedule realizing it, can be constructed in linear time using the level algorithm. Furthermore, by Lemma 3.2, we can use the level algorithm to decide for any integer t whether $t_1 \leq t$ and to construct a corresponding schedule, all in linear time, simply by applying the level algorithm to S_I with the profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ where $m_i = 3$ for $0 \leq i < t$ and $m_i = 2$ for $t \leq i \leq D - 1$.

THEOREM 4.1. *Suppose $S = (T_I \cup T_O, <_I \cup <_O)$ is an opposing forest task system and D is a deadline such that $|T_I| + |T_O| \leq 3D$. Then there exists a three-processor schedule σ for S with makespan D and a monotone internal profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ satisfying $m_{D-1} \geq 1$ if and only if $t_1 + t_2 \leq D$.*

Proof. Suppose S can be so scheduled, and let $\bar{m} = (m_0, m_1, \dots, m_{D-1})$, with $m_{D-1} \geq 1$, be the internal profile for some such schedule. Let $t_0 = \min\{t: m_t < 3\}$. By the definition of t_1 , we must have $t_1 \leq t_0$. Furthermore, since all of S_2 is scheduled from t_0 on and uses at most two processors, we must also have $t_2 \leq D - t_0$. Thus $t_1 + t_2 \leq D$.

For the converse, suppose $t_1 + t_2 \leq D$. Let σ_1 be a schedule for $S_I = (T_I, <_I)$ that has a monotone profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$ satisfying $m_i = 3$ for $0 \leq i < t_1$ and $m_i \leq 2$ for $t_1 \leq i < D$. By Lemma 4.1 we can transform σ_1 into another schedule σ'_1 for S_I that has a monotone profile $\bar{m}' = (m'_0, m'_1, \dots, m'_{D-1})$ satisfying $m'_i = 3$ for $0 \leq i < t_1$ and either $1 \leq m'_i \leq 2$ for $t_1 \leq i < D$ or $0 \leq m'_i \leq 1$ for $t_1 \leq i < D$. (See Fig. 7(a).) Let σ_2 be a two-processor schedule for $S_O = (T_O, <_O)$ having makespan t_2 . By the analogue of Lemma 4.1 for out-forests we can transform σ_2 into another schedule σ'_2 for S_O that has a monotone profile and that either has makespan exactly $D - t_1$ or never executes more than one task at a time. (See Fig. 7(b).) It is now straightforward to observe that σ'_1 and σ'_2 can be combined to form the desired schedule for S , simply by adding t_1 to the starting time for each task in T_O . (If the two schedules did not fit

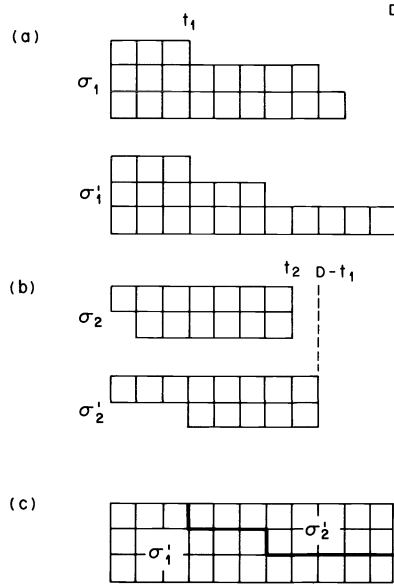


FIG. 7. Constructing a schedule for S when $t_1 + t_2 \leq D$.

together, it would necessarily be the case that the total number of tasks exceeded $3D$, contrary to our assumption.) \square

This theorem leads to the following algorithm for the $m = 3$ case of opposing forest scheduling:

THREE-PROCESSOR OPPOSING FOREST ALGORITHM

Step 1. If $|T_I| + |T_O| > 3D$, halt with no schedule possible.

Step 2. (Find a schedule with $m_{D-1} \geq 1$, if one exists.)

2A. Apply the level algorithm to compute t_2 and the corresponding schedule σ_2 .

2B. Apply the level algorithm to determine whether $t_1 \leq D - t_2$. If not (no schedule with $M_{D-1} \geq 1$ exists), go to Step 3. Otherwise, let σ_1 be the corresponding schedule obtained for S_I .

2C. Construct the transformed schedules σ'_1 and σ'_2 as in the proof of Theorem 4.1.

2D. Output the schedule σ defined by

$$\sigma(T_i) = \begin{cases} \sigma'_1(T_i) & \text{if } T_i \in T_I, \\ \sigma'_2(T_i) + t_1 & \text{if } T_i \in T_O \end{cases}$$

and halt.

Step 3. (Find a schedule with $m_0 \leq 2$, if one exists.) [Analogous to Step 2 above.]

Step 4. If neither Step 2 nor Step 3 produces a schedule, halt with no schedule possible.

THEOREM 4.2. *The above algorithm constructs a three-processor schedule with makespan D or less for S , if one exists, and can be implemented to run in linear time.*

Proof. That the algorithm works follows from Theorems 2.2 and 4.1. That it can be implemented to run in linear time follows from Lemmas 3.2 and 4.1 and the fact that the level algorithm can be implemented to run in linear time. \square

Recent work of Dolev [4] has led to a slightly more complicated, but still linear time, algorithm for the three processor problem, which has the additional property of always finding a *minimum makespan* schedule. Note that this saves a factor of $\log_2 n$ over the time complexity of using our algorithm in a binary search mode on D to find the minimum makespan. However, there is an alternative way to use Theorem 4.1 to find minimum makespan schedules that runs in linear time. We describe it as follows:

Once again we consider the two possibilities, $m_0 \leq 2$ and $m_{D-1} \geq 1$, separately and discuss only the latter, since the former can be handled symmetrically. Thus we need to show how to find a schedule for S with minimum makespan among all such schedules that have a monotone profile with $m_{D-1} \geq 1$. Consider t_1 and t_2 , defined as before. The value of t_2 is independent of the deadline D (which is now unspecified), and we can again compute it in linear time using the level algorithm. The value of t_1 , however, depends on D , so we shall write it as $t_1(D)$. By Theorem 4.1, the minimum makespan D^* for S is exactly the least deadline D such that $t_1(D) \leq D - t_2$. From the definition of t_1 , it follows that D^* is the least value of D such that S_I can be scheduled to satisfy the monotone profile $\bar{m} = (m_0, m_1, \dots, m_{D-1})$, where $m_i = 3$ for $0 \leq i \leq D - t_2 - 1$ and $m_i = 2$ for $D - t_2 \leq i \leq D - 1$.

Now, this profile has $m_0 - m_{D-1} \leq 1$, so we can apply Lemma 3.2 (the proof, as well as the statement, of Lemma 3.2 is easily seen to remain valid when there are fewer tasks than required for Case 1 of § 3 to hold). From part (ii) of Lemma 3.2, a schedule for S_I of the desired form exists for a particular D if and only if the following inequalities are satisfied, where h denotes the maximum level of a task in S_I :

$$L_i \leq 3(D - t_2) + 2t_2 - 2i \quad \text{for } 0 \leq i \leq t_2,$$

$$L_i \leq 3(D - t_2) - 3i \quad \text{for } t_2 < i \leq h.$$

The values of h , L_h , and the increments $L_i - L_{i+1}$, $0 \leq i < h$, can all be computed easily in linear time, and it is then straightforward to find D^* in linear time as the least value of D satisfying the above inequalities. Moreover, by part (iii) of Lemma 3.2, we can construct a schedule for S_I that satisfies the required profile, with $D = D^*$, by applying the level algorithm, again a linear time operation. Finally, as in our previous algorithm, this schedule can be combined with the one for S_O that achieves t_2 to obtain an overall schedule for S that has makespan D^* .

Our three-processor makespan minimization algorithm can therefore be summarized as follows:

- Step 1.* Compute t_2 and a corresponding schedule for S_O using the level algorithm.
- Step 2.* Compute D^* as the least D satisfying the above inequalities and construct a corresponding schedule for S_I using the level algorithm.
- Step 3.* Combine the schedules from Steps 1 and 2 to form a schedule for S with makespan D^* .
- Step 4.* Perform the analogous operations to Steps 1 through 3 for the symmetric case of $m_0 \leq 2$.
- Step 5.* Choose the better of the two schedules constructed in Steps 3 and 4.

From the preceding discussion, we have:

THEOREM 4.3. *The above algorithm constructs a three-processor schedule with minimum makespan and can be implemented to run in linear time.*

5. Concluding comments. In this paper we have concentrated on the question of whether the multiprocessor scheduling problem can be solved in polynomial time

for opposing forest task systems, a natural and comparatively slight generalization of the in-forest and out-forest cases, both of which can be solved in linear time using the level algorithm. One could ask similar questions about a variety of other generalizations of the known polynomially solvable subcases. Recall that MS is solvable in polynomial time for interval ordered task systems [15] and level ordered task systems [19]. A special case of both these classes is what might be called a *layered order*: the tasks are divided into classes C_1, C_2, \dots, C_k and, for any two tasks $T_i \in C_i$ and $T_j \in C_j$, T_i precedes T_j if and only if $i < j$. It is a straightforward exercise to modify the constructions used in proving Theorem 2.3 to prove that MS remains NP-complete for the following classes of partial orders:

1. Disjoint union of an in-forest (or out-forest) and a layered order.
2. Union of layered orders.
3. Intersection of two layered orders (in fact, two total orders) on the same set of tasks.

The last of these follows from the fact that any opposing forest can be represented as the intersection of two total orders, a fact whose proof we leave to the reader.

Our results leave open the question of whether there exists any fixed value of m for which m -processor MS is NP-complete for arbitrary task systems, although the algorithm in § 3 shows that task systems more complicated than opposing forests will be needed for proving any such result (unless $P = NP$). A next logical step might be to consider the case of series-parallel task systems.

The running time for our general algorithm also leaves considerable room for improvement. The techniques used for obtaining our efficient three-processor algorithm do not obviously extend to larger values of m . However, there does seem to be some hope [20] that techniques from [4] and [19] may be useful in this regard.

REFERENCES

- [1] P. M. BRUCKER, M. R. GAREY AND D. S. JOHNSON, *Scheduling equal-length tasks under tree-like precedence constraints to minimize maximum lateness*, Math. Oper. Res., 2 (1977), pp. 275–284.
- [2] E. G. COFFMAN, JR., ed. *Computer and Job/Shop Scheduling Theory*, John Wiley, New York, 1976.
- [3] E. G. COFFMAN, JR., AND R. L. GRAHAM, *Optimal scheduling for two-processor systems*, Acta Inform., 1 (1972), pp. 200–213.
- [4] D. DOLEV, *Scheduling wide graphs*, Computer Science Dept. Rep. CS80-832, Stanford Univ., Stanford, CA, December 1980.
- [5] M. FUJII, T. KASAMI AND K. NINOMIYA, *Optimal sequencing of two equivalent processors*, SIAM J. Appl. Math., 17 (1969), pp. 784–789. *Erratum*, SIAM J. Appl. Math., 20 (1971), p. 141.
- [6] H. GABOW, *An almost linear algorithm for two-processor scheduling*, J. Assoc. Comput. Mach., to appear.
- [7] M. R. GAREY AND D. S. JOHNSON, *Complexity results for multiprocessor scheduling under resource constraints*, SIAM J. Comput., 4 (1975), pp. 397–411.
- [8] ———, *Scheduling tasks with non-uniform deadlines on two processors*, J. Assoc. Comput. Mach., 23 (1976), pp. 461–467.
- [9] ———, *Two-processor scheduling with start-times and deadlines*, SIAM J. Comput., 6 (1977), pp. 416–426.
- [10] ———, *Strong NP-completeness results: Motivation, examples, and implications*, J. Assoc. Comput. Mach., 25 (1978), pp. 499–508.
- [11] ———, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.
- [12] T. C. HU, *Parallel sequencing and assembly line problems*, Oper. Res., 9 (1961), pp. 841–848.
- [13] E. L. LAWLER, *Sequencing problems with series-parallel precedence constraints*, in Proc. Conference on Combinatorial Optimization (1979), Urbino, Italy, to appear.
- [14] C. L. MONMA AND J. B. SIDNEY, *Sequencing with series-parallel precedence constraints*, Math. Oper. Res., 4 (1979), pp. 215–224.

- [15] C. H. PAPANITRIOU AND M. YANNAKAKIS, *Scheduling interval-ordered tasks*, SIAM J. Comput., 8 (1979), pp. 405–409.
- [16] R. SETHI, *Scheduling graphs on two processors*, SIAM J. Comput., 5 (1976), pp. 73–82.
- [17] J. B. SIDNEY, *The two-machine maximum flow time problem with series-parallel precedence constraints*, Oper. Res., 27 (1979), pp. 782–791.
- [18] J. D. ULLMAN, *NP-complete scheduling problems*, J. Comput. System Sci., 10 (1975), pp. 384–393.
- [19] M. WARMUTH, *M Processor unit-execution-time scheduling reduces to M-1 weakly connected components*, M.S. Thesis, Department of Computer Science, Univ. of Colorado, Boulder, 1980.
- [20] ———, Private communication, 1981.

CHOPPED ORTHOGONAL POLYNOMIAL EXPANSIONS—SOME DISCRETE CASES*

MARCI PERLSTADT†

Abstract. We study expansions of functions $f(x)$ in terms of certain discrete families of orthogonal polynomials, $\{p_i(x)\}$ where $x = 0, 1, \dots, N$, N finite or infinite. We assume f is known for $x \leq M$ ($M < N$) and that the expansion in terms of the p_i 's is chopped after L terms ($L < N$). This results in the need to study the eigenstructure of a certain "integral-type" operator. This eigenstructure is determined by producing a commuting second order difference operator.

1. Introduction. Our investigations are motivated by problems of the following type: Let f be a function with Fourier transform \hat{f} . Let A be an operator that restricts ("timelimits") f and let B be the operator that restricts \hat{f} ("bandlimits") f , i.e.,

$$Af = f \cdot \chi_{\mathcal{A}}, \quad B\hat{f} = \hat{f} \cdot \chi_{\mathcal{B}}$$

where $\chi_{\mathcal{A}}, \chi_{\mathcal{B}}$ are the respective characteristic functions of the compact sets \mathcal{A}, \mathcal{B} . Let F, F^{-1} denote the operations of Fourier transform and inverse Fourier transform:

$$F(f) = \hat{f}, \quad F^{-1}(\hat{f}) = f.$$

Suppose that \hat{f} is known only on the set \mathcal{B} and that f is known to have support in the set \mathcal{A} . Formally we have

$$\begin{aligned} BFf &= g, \quad \text{known,} \\ Af &= f. \end{aligned}$$

Combining these two equations we are faced with the problem of solving

$$Ef = BFAf = g$$

or, multiplying by E^* ,

$$E^*Ef = AF^{-1}BFAf = E^*g.$$

This leads to the study of E^*E .

The operator E^*E has been investigated in a number of cases. In general, E^*E is a finite convolution integral operator and thus determination of its eigenstructure presents a formidable problem. The problem has been tackled successfully in certain special cases by finding a second order differential operator \tilde{D} such that \tilde{D} and E^*E commute. If \tilde{D} and E^*E can be shown to have simple spectrum, then they share their eigenfunctions.

Slepian, Landau, and Pollak ([1], [2], [3]) consider the problem for the standard Fourier transform on the real line with $\mathcal{A} = [-T/2, T/2]$, $\mathcal{B} = [-\Omega, \Omega]$. They produce a second order differential operator that commutes with E^*E . That this phenomenon is not to be expected in general is indicated by Morrison [6]. Here it is noted that if \mathcal{A} and \mathcal{B} are chosen in any form other than symmetric intervals about the origin, then no commuting \tilde{D} can be found.

In [4] Slepian extends these results to the standard Fourier transform on \mathbb{R}^n with \mathcal{A}, \mathcal{B} chosen as symmetric balls about the origin. Once again a commuting \tilde{D} is found. In [7] Grünbaum considers the problem in \mathbb{R}^2 for a different choice of \mathcal{A} and \mathcal{B} . In

* Received by the editors December, 1981, and in final revised form May, 1982.

† Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

this case such a \tilde{D} cannot be found. Slepian [5] studies the situation for the Fourier series of a function on $[-\frac{1}{2}, \frac{1}{2}]$ taking $\mathcal{A} = [-W, W]$, $0 < W < \frac{1}{2}$ and taking B to correspond to chopping the series after a finite number of terms. Grünbaum [8] extends these results to the discrete Fourier transform.

Grünbaum, Longhi, and Perlstadt [9] have generalized these results to several other situations including Fourier expansions for $f \in L^2(\text{SO}(n))$. For properly chosen A and B once again a commuting \tilde{D} can be found. In particular, these results specialize to include expansions of f in terms of Gegenbauer polynomials where $\mathcal{A} = [b, 1]$, $-1 < b < 1$. Grünbaum has further noted [10] that similar results hold for expansions of f in terms of Jacobi, Hermite, and Laguerre polynomials. In each case a commuting \tilde{D} is found by appropriate modification of the Sturm–Liouville type differential equation for the family. We extend these results to the “discrete” orthogonal polynomial families satisfying second order Sturm–Liouville type difference equations.

2. The operator E^*E . Let $\{p_i(x)\}$ be a family of orthogonal polynomials on $x = 0, 1, 2, \dots, N$ with respect to the discrete weight function $w(x)$. Expanding f in terms of the p_i 's we have

$$f(x) = \sum_{i=0}^N c_i p_i(x), \quad c_i = \sum_{x=0}^N f(x) p_i(x) w(x)$$

where N is a positive integer or infinite. Suppose that f is known on $\mathcal{A} = \{0, 1, \dots, M\}$ and that the expansion of f in terms of the p_i 's is chopped after L terms ($M, L < N$). Then, letting B represent the chopping, we have

$$E^*E f(x) = A F^{-1} B F A f(x) = \sum_{y=0}^M f(y) \sum_{i=0}^L p_i(x) p_i(y) w(y).$$

We will produce a second order difference operator \tilde{D} that commutes with E^*E for those orthogonal polynomial families satisfying a second order difference equation of the form:

$$D = a_0(x) \Delta \nabla p_n(x) + a_1(x) \Delta p_n(x) + a_2(x) p_n(x) = \lambda_n p_n(x),$$

where

$$\Delta f(x) = f(x + 1) - f(x), \quad \nabla f(x) = f(x) - f(x - 1)$$

and where it is assumed $\lambda_i \neq \lambda_j$ for $i \neq j$.

Lesky [11] has shown that these polynomial families are precisely the

(1) Poisson–Charlier polynomials

$$c_n(x) = {}_2F_0\left(-n, -x; -; -\frac{1}{a}\right), \quad a > 0, \quad x = 0, 1, 2, \dots,$$

$$w(x) = \frac{e^{-a} a^x}{x!};$$

(2) Meixner polynomials

$$M_n(x) = {}_2F_1\left(-n, -x; -N; 1 - \frac{1}{c}\right), \quad \beta > 0, \quad 0 < c < 1$$

$$w(x) = \frac{c^x (\beta)_x}{x!}, \quad x = 0, 1, 2, \dots, \quad (\beta)_x = (\beta)(\beta + 1) \cdots (\beta + x - 1);$$

(3) Krawtchouk polynomials

$$K_n(x) = {}_2F_1\left(-n, -x; -N; \frac{1}{p}\right), \quad 0 < p < 1, \quad p + q = 1,$$

$$w(x) = \binom{N}{x} p^x q^{N-x}, \quad x = 0, 1, 2, \dots, N;$$

(4) Hahn polynomials

$$h_n(x) = {}_3F_2(-n, -x, n + \alpha + \beta + 1; -N, \alpha + 1; 1),$$

$$w(x) = \frac{\binom{\alpha + x}{x} \binom{\beta + N - x}{N - x}}{\binom{N + \alpha + \beta + 1}{N}}, \quad \alpha, \beta > -1, \quad x = 0, 1, 2, \dots, N.$$

Furthermore Lesky shows that for these cases the difference equation D can be recast in the Sturm–Liouville form

$$(*) \quad D = \frac{1}{w(x)} \Delta[Q(x)w(x-1)\nabla] - (R(x) + \lambda) = 0.$$

In [10] Grünbaum has shown that in the case of a continuous weight function, the basic elements needed to produce a differential operator \tilde{D} commuting with E^*E are the

- (i) Sturm–Liouville type differential equation for the p_i 's,
- (ii) Christoffel–Darboux formula,
- (iii) differentiation formula for $p_i(x)$ in terms of $p_i(x)$ and $p_{i-1}(x)$.

For the discrete case the same basic pieces are needed except that (i) and (iii) are replaced by difference equations. The equation for (i) is derived in Lesky [11]. Formula (iii) can be readily derived using the contiguous hypergeometric function formulas [14]. Writing (i) in the form above, the recipe for \tilde{D} becomes

$$\tilde{D} = \frac{1}{w(x)} \Delta[b(x-1)w(x-1)Q(x)\nabla] + G(L)c(x).$$

One should note the similarity to the continuous case [10].

In a manner analogous to the methods used in [9] we note that if \tilde{D} and E^*E are to commute it suffices to choose $b(x) = x - M$ and to choose $c(x)G(L)$ so that

$$\tilde{D}_x K_L(x, y) = \tilde{D}_y K_L(x, y),$$

where $K_L(x, y) = \sum_{i=0}^L p_i(x)p_i(y)$. This follows since

$$E^*E\tilde{D}_x f(x) = \sum_{y=0}^M K_L(x, y)\tilde{D}_y f(y)w(y)$$

$$(**) \quad = (A - B + C) \Big|_{y=-1}^M + \sum_{y=0}^M f(y)[\tilde{D}_y K_L(x, y)]w(y),$$

where

$$A = K_L(x, y + 1)b(y)w(y)Q(y + 1)\Delta f(y),$$

$$B = b(y + 1)w(y + 1)Q(y + 2)f(y)\Delta K_L(x, y + 1),$$

$$C = f(y + 1)\Delta[b(y)w(y)Q(y + 1)\nabla K_L(x, y + 1)].$$

The equality (**) follows from repeated application of the summation by parts formula.

We further note that by expanding the Δ term in C and combining terms we get that $A - B + C|_{y=-1}^M = 0$ if $b(x) = x - M$. Thus since

$$E^*E\tilde{D}_x f(x) = \sum_{y=0}^M f(y)[\tilde{D}_y K_L(x, y)]w(y)$$

we have that it suffices to show

$$\tilde{D}_x K_L(x, y) = \tilde{D}_y K_L(x, y).$$

We carry out the necessary details to show the equivalence of $D_x K_L(x, y)$ and $D_y K_L(x, y)$ for the case of the Poisson–Charlier polynomials in the next section and then give the necessary formulas for the Meixner, Krawtchouk and Hahn polynomials.

3. Poisson–Charlier polynomials. We will always use $p_i(x)$ to indicate the normalized family of orthogonal polynomials, i.e.,

$$\sum_{x=0}^N p_i(x)p_j(x)w(x) = \delta_{ij}.$$

For the Poisson–Charlier polynomials we have

(i) Second-order difference equation

$$e^a \frac{x!}{a^x} \Delta \left[e^{-a} \frac{a^{x-1}}{(x-1)!} (-a) \nabla p_i(x) \right] = ip_i(x);$$

(ii) Christoffel–Darboux

$$K_L(x, y) = \sum_{k=0}^L p_k(x)p_k(y) = \frac{\sqrt{a}\sqrt{L+1}}{x-y} [p_{L+1}(x)p_L(y) - p_L(x)p_{L+1}(y)];$$

(iii) Difference formula

$$\Delta p_n(x) = \sqrt{\frac{n}{a}} p_{n-1}(x).$$

CLAIM:

$$\tilde{D} = \frac{x!e^a}{a^x} \Delta \left[\left(-a(x-1-M) \frac{a^{x-1}e^{-a}}{(x-1)!} \right) \nabla \right] - Lx.$$

We must show $\tilde{D}_x K_L(x, y) = \tilde{D}_y K_L(x, y)$. Since (using (i) and (iii))

$$\tilde{D}_x [p_i(x)] = xp_i(x)[i-L] - (1+M)ip_i(x) - \sqrt{i}\sqrt{a}p_{i-1}(x)$$

we have

$$\begin{aligned} [\tilde{D}_x - \tilde{D}_y]K_L(x, y) &= (x-y) \sum_{i=0}^L (i-L)p_i(x)p_i(y) \\ &\quad - \sum_{i=0}^L \sqrt{i}\sqrt{a} [p_{i-1}(x)p_i(y) - p_i(x)p_{i-1}(y)] \\ &= (x-y) \sum_{i=0}^L (i-L)p_i(x)p_i(y) + \sum_{i=0}^L \sum_{k=0}^{i-1} (x-y)p_k(x)p_k(y) \\ &= (x-y) \left[\sum_{i=0}^L (i-L)p_i(x)p_i(y) + \sum_{k=0}^{L-1} \sum_{i=k+1}^L p_k(x)p_k(y) \right] = 0. \end{aligned}$$

4. The Meixner, Krawtchouk and Hahn polynomials. Once again we assume the $p_i(x)$ are normalized. The proper normalizations can be found in [12].

(a) Meixner polynomials.

(i) Second order difference equation [11]

$$\frac{x!}{c^x(\beta)_x} \Delta \left[\frac{c^{x-1}(\beta)_{x-1}}{(x-1)!} (x-1+\beta) \nabla p_i(x) \right] = i \left(1 - \frac{1}{c} \right) p_i(x);$$

(ii) Christoffel–Darboux [13]

$$(x-y) \left(1 - \frac{1}{c} \right) \sum_{i=0}^L p_i(x) p_i(y) = \sqrt{\frac{(L+1)(L+\beta)}{c}} (p_{L+1}(x) p_L(y) - p_L(x) p_{L+1}(y));$$

(iii) Difference formula [14]

$$(\beta+x) \Delta p_n(x) = n p_n(x) - \sqrt{\frac{n}{c}} (\beta+n-1) p_{n-1}(x).$$

Take $\tilde{D} = (1/w(x)) \Delta[(x-1-M)w(x-1)(x-1+\beta)\nabla] - Lx$.

(b) Krawtchouk polynomials.

(i) Second order difference equation [11]

$$\frac{1}{w(x)} \Delta[w(x-1)p(x-1-N)\nabla p_i(x)] = i p_i(x);$$

(ii) Christoffel–Darboux [13]

$$\sqrt{(L+1)(N-1)pq} [p_{L+1}(x)p_L(y) - p_L(x)p_{L+1}(y)] = (y-x) \sum_{i=0}^L p_i(x)p_i(y);$$

(iii) Difference formula [14]

$$\frac{x-N}{n} \Delta p_n(x) = \left[\sqrt{\frac{q}{p}} \left(\frac{N-n+1}{n} \right) \right] p_{n-1}(x) + p_n(x).$$

Take $\tilde{D} = (1/w(x)) \Delta[(x-1-M)w(x-1)p(x-N-1)\nabla] - Lx$.

(c) Hahn polynomials.

(i) Second order difference equation [11]

$$\frac{1}{w(x)} \Delta[-w(x-1)(\alpha+x)(N+1-x)\nabla p_i(x)] = (i)[i+\alpha+\beta+1]p_i(x);$$

(ii) Christoffel–Darboux [13]

$$(y-x) \sum_{i=0}^L p_i(x)p_i(y) = \sqrt{d_{L+1}} \sqrt{b_L} [p_{L+1}(x)p_L(y) - p_L(x)p_{L+1}(y)]$$

where

$$d_n = \frac{(n)(n+\beta)(n+\alpha+\beta+N+1)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)},$$

$$b_n = \frac{(n+\alpha+\beta+1)(n+\alpha+1)(N-n)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)};$$

(iii) Difference formula [14], [15]

$$\begin{aligned} & \frac{1}{n\sqrt{\pi_n}}(2n + \alpha + \beta)(\alpha + 1 + x)(N - x)\Delta p_n(x) \\ &= [(n + \beta)(N - x) - (n + \alpha)(n + \alpha + \beta + x + 1)]\frac{P_n(x)}{\sqrt{\pi_n}} \\ & \quad - (n + \beta)(n + \alpha + \beta + N + 1)\frac{P_{n-1}(x)}{\sqrt{\pi_{n-1}}}, \end{aligned}$$

where

$$\pi_0 = 1, \quad \pi_i = \frac{b_0 b_1 \cdots b_{i-1}}{d_1 d_2 \cdots d_i}, \quad i > 0.$$

Take $\tilde{D} = (1/w(x))\Delta[-(x - M - 1)w(x - 1)(\alpha + x)(N + 1 - x)\nabla] - L(\alpha + \beta + L + 2)x$.

5. $EE^* = BFAF^{-1}B$. It has been noted [9] that one can equally well study the operator $EE^* = BFAF^{-1}B$. Namely if f is an eigenfunction of $AF^{-1}BFA$ with eigenvalue $\lambda \neq 0$ then Bf is an eigenfunction of $BFAF^{-1}B$ with eigenvalue λ . We can represent EE^* by an $(L + 1) \times (L + 1)$ matrix with entries

$$(EE^*)_{ij} = \sum_{x=0}^M p_i(x)p_j(x)w(x), \quad 0 \leq i, j \leq L.$$

A tridiagonal matrix T that commutes with EE^* can be found by applying \tilde{D} to $p_i(x)$. A three-term recurrence formula for $\tilde{D}p_i(x)$ in terms of $p_{i-1}(x), p_i(x), p_{i+1}(x)$ results and from this we can read off matrix T . For a detailed example of this sort see [9].

If we represent T in the form

$$T = \begin{pmatrix} \alpha_0 & \gamma_0 & & & \\ \gamma_0 & \alpha_1 & \gamma_1 & & \\ & \ddots & \ddots & \ddots & \\ & & & \gamma_{L-1} & \alpha_L \end{pmatrix}$$

then we note that T has simple spectrum if $\gamma_i \neq 0, i = 0, 1, \dots, L - 1$. In the case of the four families studied here we have $\gamma_i \neq 0$. In general, however, EE^* will not have simple spectrum (take $L = M = N - 2$ for the Hahn polynomials) but the eigenvectors of T still provide an orthogonal basis of eigenfunctions for EE^* .

6. Remarks. We note that if one generalizes the form of operator D to include divided difference operators

$$\Delta f(\lambda_x) = \frac{f(\lambda_{x+1}) - f(\lambda_x)}{\lambda_{x+1} - \lambda_x}$$

then several additional families of orthogonal polynomials arise including the dual Hahn and Racah polynomials and basic hypergeometric extensions of these families. The construction of the operator \tilde{D} for these cases is discussed in [16].

Acknowledgments. It is a pleasure to thank Professors F. Alberto Grünbaum and Jeffrey Geronimo for many helpful conversations.

REFERENCES

- [1] D. SLEPIAN AND H. O. POLLAK, *Prolate spheroidal wave functions*, *Fourier analysis and uncertainty: I*, Bell System Tech. J., 40 (1961), pp. 43–64.
- [2] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions*, *Fourier analysis and uncertainty: II*, Bell System Tech. J., 40 (1961), pp. 65–84.
- [3] ———, *Prolate spheroidal wave functions*, *Fourier analysis and uncertainty: III*, Bell System Tech. J., 41 (1962), pp. 1295–1336.
- [4] D. SLEPIAN, *Prolate spheroidal wave functions*, *Fourier analysis and uncertainty: IV*, Bell System Tech. Journal, 43 (1964), pp. 3009–3058.
- [5] ———, *Prolate spheroidal wave functions*, *Fourier analysis and uncertainty: V*, Bell System Tech. J., 57 (1978), pp. 1371–1430.
- [6] J. MORRISON, *On the commutation of finite integral operators with difference kernels, and linear self-adjoint differential operators*, Abstract, Notices AMS, 9 (1962), p. 119.
- [7] F. A. GRÜNBAUM, *A study of Fourier space methods for “limited angle” reconstruction*, Numerical Functional Analysis and Optimization, 2 (1980), pp. 31–42.
- [8] ———, *Eigenvectors of a Toeplitz matrix: Discrete version of the prolate spheroidal wave functions*, this Journal, 2 (1981), pp. 136–141.
- [9] F. A. GRÜNBAUM, L. LONGHI AND M. PERLSTADT, *Differential operators commuting with finite convolution integral operators: Some nonabelian examples*, SIAM J. Appl. Math., 42 (1982), pp. 941–955.
- [10] F. A. GRÜNBAUM, *A new property of a reproducing kernel for classical orthogonal polynomials*, to appear.
- [11] P. LESKY, *Orthogonale Polynomsysteme als Lösungen Sturm–Liouvillescher Differenzgleichungen*, Monat. Math., 66 (1962), pp. 203–214.
- [12] R. ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.
- [13] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1967.
- [14] E. D. RAINVILLE, *The contiguous function relations for ${}_pF_q$* , Bull. AMS, 51 (1945), pp. 714–727.
- [15] S. KARLIN AND J. L. MCGREGOR, *The Hahn polynomial formulas and an application*, Scripta Math., 26 (1961), pp. 33–46.
- [16] M. PERLSTADT, *A property of orthogonal polynomial families with polynomial duals*, to appear.

OPTIMAL DETECTION OF TWO COMPLEMENTARY DEFECTIVES*

C. CHRISTEN†

Abstract. This paper is concerned with the problem of detecting two defective coins of respective weight $w+e$ and $w-e$, mixed with $n-2$ coins of standard weight w , in the minimum number of weighings on a single-dish spring scale. The exact solution of the problem for the worst case is obtained. In particular, it is shown that asymptotically $((1+\log_2 3)/\log_2 3) \log_2 n$ weighings are required, an improvement of almost 30% over the information-theoretic lower bound.

Introduction: A candy factory problem. Workers of a candy factory pack boxes containing a fixed number of equally heavy candy pieces. After a day's work, a mischievous employee informs his supervisor that he has put a piece of candy from one of his 2000 packed boxes into another, so that one box is too heavy and another too light. The supervisor has to correct this situation; of course he doesn't want to open and repack that many boxes. He has at his disposal a single-dish spring scale, on which the weight difference due to a single candy piece can be detected. How many weighings does he need, and how is he to proceed? One may assume that about thousand boxes can be weighed simultaneously on the scale.

Clearly, this is one of the cases where group testing helps. Using a divide-and-conquer method, one may easily come up with 21 weighings, instead of 1998.

On the other hand, one weighing shows that the weighted boxes either contain an additional piece of candy or lack one piece or contain the correct number of pieces (in which case the defective boxes may be or not be among the weighed boxes). But there are at least $2000 \cdot 1999$ possible (ordered) pairs of boxes; thus at least $\lceil \log_3 2000 \cdot 1999 \rceil = 14$ weighings are necessary.

In fact, the minimum number of weighings here is 17.

1. General formulation and results. The above problem may be formulated in a more usual way as a coin-weighing problem:

Given are n coins, of which $n-2$ have the standard weight w , one the excess weight $w+e$ and one the overweight weight $w-e$. What is the minimum number of weighings on a single-dish spring scale necessary in the worst case to detect the overweight and the overheavy coin, assuming that the simultaneous weighing of arbitrarily many coins is possible?

In a slightly more general variant, the possibility that all n coins have standard weight will also be allowed.

Unless specifically mentioned, all logarithms are base 2.

A counting argument similar to the above shows that at least

$$\lceil \log_3 n(n-1) \rceil \sim \frac{2}{\log 3} \log n$$

weighings are required, whereas a rude divide-and-conquer approach (repeatedly using the binary method to: find a deviation; find out in which of the subsets the defective pair is located; separately find the heavy and the light coin) shows that at most

$$2 \lceil \log n \rceil - 1$$

* Received by the editors July 28, 1981, and in revised form June 3, 1982. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

† Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Canada H3C 3J7.

weighings are required. However, it should be expected that a more refined method, looking simultaneously for the heavy and the light coin, should save some fraction of the weighings.

On the other hand, the above counting argument does not take into account the structure of the problem. It seems in fact that the size of the three outcomes cannot be balanced, so that one should expect more weighings than predicted by the lower bound. In similar situations, the derivation of an improved lower bound is often a tedious affair, usually implying complicated counterstrategies. As it turns out, a surprisingly simple optimal counterstrategy exists here.

Relying on this, the exact solution to the problem is obtained. Let f be defined for $t \geq 0$ by:

$$f(t) = \begin{cases} 7 & \text{if } t = 4, \\ (3^{\lfloor (1+t)/(1+\log 3) \rfloor} - 1 + 2^{1+t - \lfloor (1+t)/(1+\log 3) \rfloor})/2 & \text{if } 1 + 2^{t - \lfloor (1+t)/(1+\log 3) \rfloor} < 3^{\lfloor (1+t)/(1+\log 3) \rfloor}, \\ \lfloor 3(3^{\lfloor (1+t)/(1+\log 3) \rfloor} + 2^{2 - \lfloor (1+t)/(1+\log 3) \rfloor})/4 \rfloor & \text{otherwise.} \end{cases}$$

The first twenty values of f are given in Table 1.

TABLE 1.

t	=	0	1	2	3	4	5	6	7	8	9
$f(t)$	=	1	2	3	5	7	12	18	29	44	68
t	=	10	11	12	13	14	15	16	17	18	19
$f(t)$	=	104	156	249	374	566	876	1314	2082	3141	4712

MAIN THEOREM. *The detection of one, or at most one, complementary defective pair requires in the worst case t weighings for $f(t-1) < n \leq f(t)$.*

In particular, this shows that the correct value is asymptotically

$$((1 + \log 3)/\log 3) \log n,$$

thus yielding an improvement of about 18.5% from the previously stated upper bound and of about 29.2% over the information-theoretic lower bound.

This is in sharp contrast with the similar problem using a two-dishes beam scale, where the lower bound $(2/\log 3) \log n$ can essentially be achieved.

The detection of a single defective coin using a beam scale is a well-known and well-solved puzzle (see [1] and [10]). The work by Bellman and Gluss [2] is a try on the detection of two defective heavy coins with a beam scale. Cairns [3] solves the problem within 1 weighing.

The detection of a single defective coin on a spring scale is of course trivial for the worst-case analysis. Christen [5] shows that the detection of two defective coins of equal bias using a spring scale requires considerably less weighings than the present case, although the number of possibilities is the same. However, no sharp lower bound is known in that case. The detection of an arbitrary number of identical false coins with a spring scale has been treated by Söderberg and Shapiro [11], Erdős and Rényi [6], Lindström [6], [7], [8] and Cantor and Mills [4].

2. Some illustrative cases. The first few cases are easily determined:

If fewer than two coins are given, no weighing is required, since there cannot be any defective pair.

If two coins are given, there are at least two possibilities for the defective pair, so one weighing is required. Of course, weigh a single coin; depending whether the result is $w + e$, $w - e$ or w , the weighed coin is heavy or light or there is no defective coin.

If three coins are given, there are at least six possibilities for the defective pair, so that at least two weighings are needed. First weigh a single coin. If the result is $w + e$ (respectively $w - e$), the weighed coin is heavy (light) and the complementary defective is one of the two unweighed coins, so one more weighing is sufficient; if the result is w , two possible defective coins are left, so one more weighing suffices too.

If more than three coins are given, there are at least ten possibilities for the defective pair, so that at least three weighings are required.

The terminology used in the paper will now be introduced and illustrated on the next cases.

Suppose x of the given n coins are weighed. A result of $xw + e$ indicates a *positive bias*: one of the x weighed coins must be heavy and one of the $n - x$ unweighed coins light. This situation will be called an $x:(n - x)$ *configuration*. The first component is always the number of heavy candidates. Similarly, a result of $xw - e$ indicates a *negative bias*; this symmetrical situation is of course an $(n - x):x$ configuration. Finally, in the unbiased case (result xw), either the defective pair is among the weighed coins or it is among the unweighed coins or possibly there is no defective pair.

To every detection scheme is associated a rooted ternary tree, the decision tree of the scheme, whose nodes correspond to the situations after the weighings and whose edges leading out of a node correspond to the result of the weighing. The *central branch* of height h of a decision tree is the branch of length h corresponding to the unbiased result for each of the first h weighings. *Central situations* are those corresponding to the nodes of a central branch.

In an $x:(n - x)$ configuration, suppose y from the x and z from the $n - x$ coins are weighed. The positively and negatively biased results respectively lead to an $y:(n - x - z)$ and an $(x - y):z$ configuration. However, in the unbiased case, the defective pair consists of either one of the y weighed heavy candidates and one of the z weighed light candidates or of one of the $x - y$ unweighed heavy candidates and one of the $n - x - z$ unweighed light candidates. This situation will be called an $y:z + (x - y):(n - x - z)$ *configuration*; the two alternatives are the *summands* of the configuration. Similarly, if the possible defective pair is known to be either in a subset of x coins or in a subset of $n - x$ coins and if y from the x and z from the $n - x$ coins are weighed, an $y:(x - y) + z:(n - x - z)$ (respectively $(x - y):y + (n - x - z):z$) configuration is obtained in a case of a positive (negative) bias, whereas four subsets of coins emerge otherwise. More generally, in a central situation, a weighing produces two subsets out of one (unless the weighing does not split the coins of the subset), so that in a central situation of height h the coins are partitioned into at most 2^h subsets. Similarly, in a noncentral situation, a weighing produces two summands out of one in the unbiased case, unless one set of candidates is unsplit. In the biased case, a summand arises out of each subset or summand, unless the subset or the subset of corresponding candidates is unsplit. Thus in noncentral situations of height h , the configurations consist of at most 2^h summands.

The *magnitude* of a situation is the number of possibilities for the defective pair consistent with the situation. Clearly, the magnitude of a summand is the product of its components, and the magnitude of a configuration is the sum of the magnitudes of its summands. In a central situation, the magnitude of a subset of s elements is $s(s - 1)$, and the magnitude of the partition is the sum of the magnitudes of the subsets (plus one if the case of no defective pair is allowed).

Example 1. From the $2:1+1:1$ configuration, one single weighing detects the defective pair.

Indeed, weigh one of the heavy candidates from the first summand and the light candidate from the second summand. If the bias is negative, the light coin is the weighed light candidate and the heavy coin the corresponding heavy candidate. Otherwise, the light coin is the unweighed light candidate and the heavy coin the weighed heavy candidate or the unweighed one from the first summand, depending whether the bias is positive or zero.

Example 2. From the $3:2+1:2+1:1$ configuration, two weighings detect the defective pair.

Weigh one heavy and one light candidate from the first summand, the heavy candidate from the second and the light candidate from the third. Zero bias leads to a $1:1+2:1$ configuration generated from the first summand only, while positive and negative bias lead to a $1:1+1:2$ configuration generated from the first and the second summand, respectively to a $2:1+1:1$ configuration generated from the first and the third summand. Of course, by suitably interchanging light and heavy and first and second, all these configurations can be treated as in Example 1; thus two weighings suffice in each case.

In general, it is worth noting that configurations differing only in the order of the summands or in the order of the components of a summand can be treated in a similar way. This elementary fact will be constantly used in the following without further mention.

One may further note that interchanging the weighed against the unweighed coins simply inverts the bias; thus it is never necessary to weigh more than half of the coins.

Example 3. Three weighings are sufficient to detect a possible defective pair among five coins.

Weigh first two coins. If the bias is zero, weigh one coin from each subset; if the bias is again zero, the defective pair can only be in the resulting subset of two coins; otherwise a $1:2+1:1$ or a $2:1+1:1$ configuration arises. If the result of the first weighing indicates a bias, a $2:3$ or a $3:2$ configuration is generated, for which two more weighings suffice by Example 2.

Of course, at least four weighings are required for six coins, since the magnitude is here at least thirty.

The next examples show that structure may be more important than number of possibilities (but see § 3).

Example 4. Three weighings are required in the worst case to detect the defective pair from the $2:2+2:2$ configuration.

When a $2:2$ summand is split by a weighing, two possibilities remain open in the unbiased case (as one $2:1$ or $1:2$ summand or as two $1:1$ summands). Thus if both $2:2$ summands are split, the magnitude of the produced central situation is four, so one more weighing is insufficient. But if a summand is unsplit, it appears in one of the resulting configurations; so that one more weighing is insufficient there too.

Given nine coins, there are still less than 81 possibilities for the defective pair. But the 56 possibilities for eight coins are already too many:

Example 5. In the worst case, five weighings are required to detect the defective pair out of eight coins.

After two unbiased results, either four subsets of two coins are obtained or one subset of three and two subsets of two or two subsets of at least three coins. In the second and third case, the magnitude is at least ten, so three more weighings are required. But to have obtained four subsets of two, four coins must have been weighed

first and two of each subset next. Thus if the second result had been biased, a 2:2+2:2 configuration would have been generated, for which three weighings are required, as shown in Example 4.

By contrast, four weighings suffice if the first result is biased (reduction to Example 2).

Of course, four weighings suffice for seven coins, the only critical case for eight being then dominated by Example 2 (four weighings suffice for a 2:2+2:1 configuration since they suffice for a 3:2+1:2+1:1 configuration).

3. Derivation of the lower bound. Throughout this section, the case of no defective pair is excluded. Of course, the derived lower bound holds also when this case is allowed.

LEMMA 1. (i) *The minimum of $2 \sum_{i=1}^m \binom{s_i}{2}$ under the constraint $\sum_{k=1}^m s_k = n$ occurs exactly when $m - n + \lfloor n/m \rfloor m$ of the s_k are equal to $\lfloor n/m \rfloor$ and the others to $\lceil n/m \rceil$.*

(ii) *The above minimum is equal to $(\lceil n/m \rceil - 1)(2n - \lceil n/m \rceil m)$.*

(iii) *For fixed $n \geq m$, decreasing m strictly increases the value of this minimum.*

Proof. (i) Note first that for $s'' \leq s'$ and $s'' + s' = s$ the minimum of $2 \binom{s''}{2} + 2 \binom{s'}{2}$ occurs only when $s'' = \lfloor s/2 \rfloor$ and $s' = \lceil s/2 \rceil$. Indeed, for $i > 0$

$$2 \binom{s'' - i}{2} + 2 \binom{s' + i}{2} = 2 \binom{s''}{2} + 2 \binom{s'}{2} + 2i(s' - s'') + 2i^2 > 2 \binom{s''}{2} + 2 \binom{s'}{2}.$$

It follows from this observation that if $2 \sum_{k=1}^m \binom{s_k}{2}$ is minimum under $\sum_{k=1}^m s_k = n$ then $|s_i - s_j| \leq 1$ must hold for all i, j with $1 \leq i, j \leq m$. This is because when $s_j + 2 \leq s_i$, replacing the pair $\{s_i, s_j\}$ by the pair $\{ \lfloor (s_i + s_j)/2 \rfloor, \lceil (s_i + s_j)/2 \rceil \}$ would strictly decrease the objective function while preserving the constraint.

But if $n \equiv r \pmod m$ (with $0 \leq r < m$), the only way to satisfy $\sum_{k=1}^m s_k = n$ under the condition $|s_i - s_j| = 1$ for all i, j is to take r of the s_k equal to $\lceil n/m \rceil$ and the $m - r$ others equal to $\lfloor n/m \rfloor$. The assertion follows immediately from the fact that the remainder of $n \pmod m$ is equal to $n - \lfloor n/m \rfloor m$.

(ii) The minimum of the objective function is therefore $(n - \lfloor n/m \rfloor m) \lceil n/m \rceil (\lceil n/m \rceil - 1) + (m - n + \lfloor n/m \rfloor m) \lfloor n/m \rfloor (\lfloor n/m \rfloor - 1)$, which is easily seen to be equal to $(\lceil n/m \rceil - 1)(2n - \lceil n/m \rceil m)$ by distinguishing the cases of zero and nonzero remainder mod m .

(iii) When m is decreased by one, one of the s_k has to be set from (say) $\lfloor n/m \rfloor$ to zero. By (i), to produce the minimum, $\lfloor n/m \rfloor$ of the smallest possible others have to be increased by one. This increases the objective function by at least $2 \lfloor n/m \rfloor \lfloor n/m \rfloor - \lfloor n/m \rfloor (\lfloor n/m \rfloor - 1) = \lfloor n/m \rfloor (\lfloor n/m \rfloor + 1)$, which is positive for $n \geq m$. \square

It turns out now that Examples 4 and 5 are somewhat atypical, because except for five weighings, the large magnitudes of the central situations supersede other considerations.

THEOREM 1. *The detection of the complementary defective pair among $f(t) + 1$ coins requires always at least $t + 1$ weighings in the worst case.*

Proof. For $t = 4$, the assertion was shown in Examples 4 and 5. Otherwise, let $h = t - \lfloor (1+t)/\log 6 \rfloor$. From $\lfloor (1+t)/\log 6 \rfloor < (1+t)/\log 6 < 1 + \lfloor (1+t)/\log 6 \rfloor$ follows $6^{\lfloor (1+t)/\log 6 \rfloor} < 2^{1+t} < 6^{1 + \lfloor (1+t)/\log 6 \rfloor}$, hence $6^{t-h} < 2^{1+t} < 6^{t-h+1}$. Therefore

$$(*) \quad 2^h \leq 3^{t-h+1} - 1$$

and

$$(**) \quad 3^{t-h} + 1 \leq 2^{h+1}.$$

Lemma 1 will now be applied with $n = f(t) + 1$ to obtain a lower estimate on the magnitude of an appropriate situation. By part (iii) of this lemma, it suffices to consider the largest possible m . Two cases have to be distinguished.

Case 1. If $2^h < 3^{t-h} - 1$, consider the central situation of height h . Here $n = f(t) + 1 = (3^{t-h} + 1 + 2^{h+1})/2$ and $m = 2^h$. From (*) and from the hypothesis,

$$1 < n/m = (3^{t-h} + 1 + 2^{h+1})/2^{h+1} < 2,$$

hence $\lceil n/m \rceil = 2$.

By part (ii) of Lemma 1, the magnitude of the considered situation is thus at least $2(n - m) = 3^{t-h} + 1$. Therefore at least $t - h + 1$ more weighings are required in this situation to detect the defective pair.

Case 2. If $3^{t-h} - 1 \leq 2^h$, consider the central situation of height $h - 1$. Here $n = f(t) + 1 = \lfloor 3(3^{t-h} 2^h)/4 \rfloor + 1$ and $m = 2^{h-1}$. By part (ii) of Lemma 1, the magnitude of this situation is at least

$$(\lceil n/m \rceil - 1)(2n - \lceil n/m \rceil m) = 4n - 6m + (\lceil n/m \rceil - 3)(2n - \lceil n/m \rceil m - 2m).$$

From the hypothesis and from (**),

$$2 < n/m = (\lfloor 3(3^{t-h} 2^h)/4 \rfloor + 1)/2^{h-1} \leq 4,$$

hence $3 \leq \lceil n/m \rceil \leq 4$. The last term of the above expression is zero when $n/m \leq 3$ and positive otherwise. But

$$\begin{aligned} 4n - 6m &= 4 \lfloor 3(3^{t-h} + 2^h)/4 \rfloor + 4 - 3 \cdot 2^h \\ &\geq 4 \lfloor (3(3^{t-h} + 2^h) + 1)/4 \rfloor - 3 \cdot 2^h \\ &\geq 3^{t-h+1} + 1. \end{aligned}$$

Thus at least $t - h + 2$ further weighings are necessary in this situation to detect the defective pair.

4. Derivation of the upper bound. The first three easy results show that there are no structural complications once heavy splitting has been done; thus balanced ternary subdivision is possible.

LEMMA 2. *From a configuration consisting of 3^t 1:1 summands, the defective pair can be detected in at most t weighings.*

Proof. By induction on t .

One single 1:1 summand means that there are only one heavy and one light candidate; thus no weighing is required.

Suppose the assertion holds for t . Given 3^{t+1} 1:1 summands, weigh 3^t heavy candidates and 3^t light candidates from distinct summands. If the bias is positive, one of the 3^t weighed heavy candidates must be heavy; if it is negative, one of the 3^t weighed light candidates must be light; otherwise the defective pair may only be among the 3^t remaining summands. Thus $t + 1$ weighings are sufficient. \square

LEMMA 3. *From a configuration consisting of $(3^t - 1)/2$ 2:1 configurations and of one 1:1 configuration, the defective pair can be detected in at most t weighings.*

Proof. By induction on t .

The assertion is trivial for $t = 0$; the case of $t = 1$ is in fact Example 1.

Suppose the assertion holds for t . Given $(3^{t+1} - 1)/2$ 2:1 summands and one 1:1 summand, weigh one heavy candidate from the first 3^t summands and the light candidate from the remaining $(3^t + 1)/2$ summands (including the single 1:1 summand). If the bias is zero or positive, a configuration consisting of 3^t 1:1 summands is obtained,

for which t weighings are sufficient by Lemma 2. If the bias is negative, there must be a light element among the $(3^t + 1)/2$ light candidates; thus by the induction hypothesis $t + 1$ weighings are sufficient. \square

LEMMA 4. t weighings are sufficient to detect the possible defective pair among $(3^t - 1)/2$ pairs.

Proof. By induction on t .

For $t = 0$, the assertion is trivial. Suppose it holds for t . Given $(3^{t+1} - 1)$ pairs, weigh one element from each of 3^t pairs. If the result is biased, the defective pair must contain one of the weighed elements; the obtained configuration consists of 3^t 1:1 summands, thus t further weighings are sufficient by Lemma 2. If the result is unbiased, the defective pair is among the $(3^t - 1)/2$ remaining pairs, so that t further weighings are sufficient by the induction hypothesis. \square

In the following proofs, use will be made of an elementary fact about configurations: If a $(x + y):z$ summand is dissected into one $x:z$ and one $y:z$ summand, no more weighings are required in the resulting configuration than in the start configuration. Indeed, the dissection means that instead of having $x + y$ heavy candidates from a subset H and z light candidates from a subset L , one has x heavy candidates from a subset H' and z corresponding light candidates from a subset L' or y heavy candidates from a subset H'' and z corresponding light candidates from a subset L'' . If a weighing from the start configuration involves u heavy candidates from H and v light candidates from L , emulate it by involving v light candidates from L' , v light candidates from L'' and a total of u heavy candidates from the union of H' and H'' in the weighing. The resulting configurations are dissections of those resulting of the start configuration.

The next three results generalize Lemma 3 for some larger summands.

LEMMA 5. From a configuration consisting of x 2:2 summands, y 2:1 summands and z 1:1 summands, t weighings are sufficient to detect the defective pair, as long as $4x + 2y + z = 3^t$, except when $t = x = 2$.

Proof. By induction on t .

For $t \leq 1$, the assertion reduces to previous cases, since x must be zero. Since 2:1 and 1:1 summands may be viewed as dissections of 2:2 summands, it is sufficient to prove the assertion for minimal y and z .

For $t = 2$, it was shown in Example 4 that $x = 2$ is impossible in two weighings; thus let $x = 1$, $y = 2$ and $z = 1$. Weigh one heavy candidate from the 2:2 summand and from one 2:1 summand and one light candidate from each of the remaining summands. The resulting configurations are then those of Example 1, so that one more weighing is sufficient.

Suppose the assertion holds for $2t$. For $2t + 1$, minimal y and z are 1. Weigh one heavy candidate from $(3^{2t} - 1)/4$ 2:2 summands and from the single 2:1 summand, one light candidate from $(3^{2t} - 1)/4$ other 2:2 summands and from the single 1:1 summand, both heavy candidates from half the remaining summands and both light candidates from the other half. If the bias is zero, the resulting configuration consists of $(3^{2t} - 1)/4$ 1:2 summands, $(3^{2t} - 1)/4$ 2:1 summands and one 1:1 summand; by Lemma 4, $2t$ more weighings are sufficient. Otherwise, the resulting configuration consists of $(3^{2t} - 1)/4$ 1:2 (resp 2:1) summands, $(3^{2t} - 1)/8$ 2:2 summands and one 1:1 summand; by the induction hypothesis, $2t$ more weighings are sufficient (note that the exceptional case is avoided).

For $2t + 2$, minimal y and z are respectively 0 and 1. Weigh one heavy candidate from a first 2:2 summand, one light candidate from a second one, one heavy and one light candidate from a third one, both heavy candidates from $(3^{2t+1} - 3)/4$ other summands and both light candidates from $(3^{2t+1} - 3)/4$ further ones. In the biased

cases, the resulting configuration consists of $(3^{2t+1}-3)/4$ 2:2 summands, one 1:2 (respectively 2:1) summand and one 1:1 summand. In the unbiased case, it consists of $(3^{2t+1}-7)/4$ 2:2 summands, one 1:2, one 2:1 and three 1:1 summands. By the induction hypothesis, $2t+1$ more weighings are sufficient. \square

LEMMA 6. *From a configuration consisting of x 3:2 summands, y 2:2 summands, u 2:1 summands and v 1:1 summands, t weighings are sufficient to detect the defective pair, as long as $6x+4y+2u+v=3^t$, except when $t=y=2$.*

Proof. By induction on t .

For $t \leq 1$, the assertion follows from the previous lemma. Because 2:1 and 1:1 summands may be viewed as dissections of the larger summands, it is sufficient to prove the assertion for minimal u, v (the same is not true of 2:2 summands, since there may not be enough 2:1 summands).

Suppose the assertion holds for $t \geq 2$. If x is odd, minimal u and v are 1 and $y \equiv 0 \pmod{3}$. If $y=0$, weigh two heavy and one light candidate from one 3:2 summand, both candidates from the 1:1 summand, all heavy candidates from $(3^{t-1}-1)/2$ 3:2 summands and all light candidates from $(3^{t-1}-1)/2$ 3:2 summands, one 2:1 or 1:2 summand and one 1:1 summand. If $y > 0$, replace groups of two similarly treated 3:2 summands by groups of three similarly treated 2:2 summands or one 3:2 summand with weighed light candidates and one without weighed candidates by three 2:2 summands with each one weighed light candidate. Thus the induction hypothesis is always satisfied (in particular, the exceptional case cannot occur).

If x is even, minimal u and v are 0 and 1 and $y \equiv 2 \pmod{3}$. If $x=0$, the assertion holds by the previous lemma, so suppose there is at least one 3:2 summand. If $y=2$, weigh two heavy and one light candidate from one 3:2 summand, one light candidate from one 2:2 summand, all heavy candidates from $(3^{t-1}-1)/2$ other 3:2 summands, all light candidates from $(3^{t-1}-1)/2$ further such summands and the heavy candidate from the single 1:1 summand. In the biased cases, the resulting configuration consists of $(3^{t-1}-1)/2$ 3:2 summands, one 2:1 and one 1:1 summand; in the unbiased case, one 3:2 summand is dissected into one 2:2 and one 2:1 summand. If $y > 2$, replace groups of two 3:2 summands by groups of three 2:2 summands as above. The exceptional case cannot occur, and the induction hypothesis is then satisfied. \square

LEMMA 7. *From a configuration consisting of x 3:3 summands, y 3:2 summands, u 2:1 summands and v 1:1 summands, t weighings are sufficient to detect the defective pair, as long as $x < 3^{t-2}$ and $9x+6y+2u+v=3^t$.*

Proof. By induction on t .

For $t \leq 2$, the assertion follows from Lemma 6. Suppose the assertion holds for t . Minimal u and v are here 1. If x is even, $y \equiv 1 \pmod{3}$. If $y=1$, weigh one light and two heavy candidates from the 3:2 summand, the light candidate from the 2:1 summand, the heavy candidate from the 1:1 summand, two heavy candidates from one 3:3 summand, two light candidates from another, all heavy candidates from $(3^{t-1}-1)/2$ other 3:3 summands and all light candidates from $(3^{t-1}-1)/2$ further such summands. In the biased cases, the resulting configuration consists of $(3^{t-1}-1)/2$ 3:2 summands, one 2:3 (respectively 3:2) summand, one 2:1 and one 1:1 summand; in the unbiased case, the 2:3 summand is replaced by one 3:1 and one 1:3 summand (a dissection with an inversion). If $y > 1$, replace groups of two similarly treated 3:3 summands by groups of three similarly treated 3:2 summands or one 3:3 summand with weighed light candidates and one without weighed candidates by one 3:2 summand with weighed light candidates, one 3:2 summand without weighed candidates and one 3:2 summand with one weighed light candidate.

If x is odd, $y \equiv 0 \pmod 3$ and $y > 0$. If $y = 3$, weigh one light and two heavy candidates from one 3:2 summand, one heavy and two light from another, two heavy candidates from half the 3:3 summands and two light candidates from the remaining 3:3 summands. The resulting configuration directly satisfies the hypothesis of Lemma 6 in the biased cases and is a dissection with some inversions of such a configuration in the unbiased case. If $y > 3$, replace groups of two 3:3 summands by groups of three 3:2 summands as above. \square

In the following, to *halve* a subset of s coins means to take $\lfloor s/2 \rfloor$ of its coins in the next weighing, while to halve an $x:y$ summand means to take $\lfloor x/2 \rfloor$ of its heavy candidates and $\lfloor y/2 \rfloor$ of its light candidates in the next weighing.

The optimal algorithm can now be stated:

DETECTION ALGORITHM. As long as a component or a subset of at least four coins remains, halve all summands or subsets.

In a central situation, if all subsets have at most two elements, apply the method of Lemma 4. If some subset of three coins remains, split all subsets of three elements and $(2^{1+m} + 1 - 3^p)/2$ subsets of two elements, where m is the number of weighings already made and p the largest exponent such that $3^p < 2^{1+m}$. Apply then the method of Lemma 5 in the biased cases and the method of Lemma 4 in the unbiased case.

In a noncentral situation, apply the methods of Lemma 5, 6 or 7, depending on the size of the largest remaining summands.

THEOREM 2. *The given algorithm requires at most t weighings to detect a single possible complementary defective pair among $f(t)$ coins.*

Proof. (i) Note first that as long as halving is used, at any stage the cardinalities of the obtained subsets or components differ by at most 1; this is because halving sets of cardinalities c and $c + 1$ produces only sets of cardinality between $\lfloor c/2 \rfloor$ and $\lceil (c + 1)/2 \rceil$ and that $\lfloor c/2 \rfloor + 1 \geq \lceil (c + 1)/2 \rceil$. Furthermore, with the given rule, it is impossible that both $c:c$ and $(c + 1):(c + 1)$ summands occur in the same configuration. Indeed, in the unbiased case, summands are of type $\lfloor x_i/2 \rfloor : \lfloor y_i/2 \rfloor$ or $\lfloor x_i/2 \rfloor : \lceil y_i/2 \rceil$; having both $c:c$ and $(c + 1):(c + 1)$ summands would imply $x_i + y_i \leq 4c + 1$ for some i and $x_j + y_j \geq 4c + 3$ for some j , contradicting the first remark. In the biased cases the argumentation is similar.

(ii) Note then that as long as halving is used, the magnitude of a noncentral situation never exceeds the magnitude of the central situation of same height. Indeed, for $k2^h \leq n \leq (2k + 1)2^{h-1}$, the magnitude of the central situation of height h is $(k^2 - k)2^h + (n - k2^h)2k$, whereas the magnitude of a noncentral situation of height h is $k^22^{h-1} + (n + k2^h)k$. But here $k \geq 2$, else the halving would have stopped after $h - 1$ weighings, the cardinality of every component being then at most 3; thus the assertion holds. For $(2k + 1)2^{h-1} < n \leq (k + 1)2^h$, the magnitude of the central situation of height h is $k^22^h + (n - (2k + 1)2^{h-1})2k$ and that of the noncentral situation of same height $(k^2 + k)2^{h-1} + (n - (2k + 1)2^{h-1})(k + 1)$. But here the assertion holds again, since $k \geq 1$ (or the halving would have stopped before).

(iii) The magnitude of some situations will now be evaluated. Let again $h = t - \lfloor (1 + t)/\log 6 \rfloor$. Since the smaller values have already been handled in § 2, we may suppose $t \geq 5$; hence $h \geq 2$.

Case 1. If $2^h < 3^{t-h} - 1$, the detection algorithm executes h halving steps. Indeed, since $n = (3^{t-h} - 1 + 2^{h+1})/2$, there are exactly $(3^{t-h} - 1)/2$ subsets of two coins among the 2^h subsets of the central situation of height h . As this is (by the case hypothesis) more than half of the subsets, there was still at least one subset of four coins after $h - 1$ halvings, so that the algorithm had to halve one more. The magnitude of the situation at the end of the halvings is thus 3^{t-h} .

Case 2. If $3^{t-h} - 1 \leq 2^h$, the detection algorithm executes only $h - 1$ halving steps. Indeed, since $n = \lfloor 3(3^{t-h} + 2^h)/4 \rfloor$, there are exactly $\lfloor 3^{t-h+1}/4 \rfloor - 2^{h-2}$ subsets of three coins and $3 \cdot 2^{h-2} - \lfloor 3^{t-h+1}/4 \rfloor$ subsets of two coins in the central situation of height $h - 1$. By (*) and by the case hypothesis, both quantities are nonnegative. The magnitude of this situation is thus $4 \lfloor 3^{t-h+1}/4 \rfloor + 1 \leq 3^{t-h+1}$. (Note that in case all subsets have exactly 3 coins the magnitude is strictly less than 3^{t-h+1} ; hence the restriction in Lemma 7 does not apply.)

At step h , the algorithm splits all subsets of three coins and $(2^h + 1 - 3^{t-h})/2$ subsets of two coins. The arising central situation has $(3^{t-h} - 1)/2$ subsets of two coins and thus magnitude 3^{t-h} , whereas the arising noncentral situations have $\lfloor 3^{t-h+1}/4 \rfloor - 2^{h-2}$ 2:1 summands and $(2^h + 1 - 3^{t-h})/2$ 1:1 summands and thus magnitude $\leq 3^{t-h}$.

(iv) By (iii), Lemma 4 and Lemma 5, the detection algorithm uses only $t - h$ weighings in Case 1 and $t - h + 1$ weighings in Case 2 to detect the defective pair from the central situation at the end of the halvings.

By (ii), (iii), Lemma 5, Lemma 6 and Lemma 7, the detection algorithm uses only $t - h$ weighings in Case 1 and $t - h + 1$ weighings in Case 2 to detect the defective pair from a noncentral situation at the end of the halvings. The theorem is thus proved. \square

Putting both theorems together yields the main result.

COROLLARY. *Asymptotically, $((1 + \log 3)/\log 3) \log n$ weighings are required to detect a possible complementary defective pair among n coins.*

Proof. The exact formula implies that for some constants c_1 and c_2 ,

$$c_1 2^{t \log 3 / (1 + \log 3)} \leq f(t) \leq c_2 2^{t \log 3 / (1 + \log 3)}.$$

Thus the assertion follows from the main theorem.

REFERENCES

- [1] *Problems and Solutions*, Amer. Math. Monthly, E. D. SCHELL, Problem E651, 52 (1945), p. 42, Sol., (1945), p. 397; D. EVES, Problem E712, 53 (1946), p. 156, Sol., 54 (1947), pp. 46-48; N. J. FINE, Problem 4203, 53 (1946), p. 278, Sol., 54 (1947), pp. 489-491.
- [2] R. BELLMAN AND B. GLUSS, *On various versions of the defective coin problem*, Inform. and Control, 4 (1961), pp. 118-131.
- [3] S. S. CAIRNS, *Balance scale sorting*, Amer. Math. Monthly, 70 (1963), pp. 136-148.
- [4] D. G. CANTOR AND W. H. MILLS, *Determination of a subset from certain combinatorial properties*, Canad. J. Math., 18 (1966), pp. 42-48.
- [5] C. CHRISTEN, *A Fibonacci algorithm for the detection of two elements*, Publ. 341, Dépt. d'I.R.O., Université de Montréal, Montréal, 1980.
- [6] P. ERDŐS AND A. RÉNYI, *On two problems of information theory*, Publ. Hung. Acad. Sci., 8 (1963), pp. 241-254.
- [7] B. LINDSTRÖM, *On a combinatory detection problem I*, Publ. Hung. Acad. Sci., 9 (1964), pp. 195-207.
- [8] ———, *On a combinatory detection problem II*, Stud. Sci. Math. Hung., 1 (1966), pp. 353-361.
- [9] ———, *On a combinatorial problem in number theory*, Canad. Math. Bull., 4 (1965), pp. 477-490.
- [10] C. A. B. SMITH, *The counterfeit coin problem*, Math. Gazette, 31 (1947), pp. 31-39; see also 29 (1945), pp. 227-229, 30 (1946), pp. 231-234.
- [11] S. SÖDERBERG AND H. S. SHAPIRO, *A combinatory detection problem*, Amer. Math. Monthly, 70 (1963), pp. 1066-1070.

EXACT SOLUTION OF SYSTEMS OF LINEAR EQUATIONS WITH ITERATIVE METHODS*

SILVIO URSIC† AND CYRO PATARRA‡

Abstract. An algorithm is presented to compute the exact solution of a system of linear equations with integer coefficients from any method capable of providing a sufficiently accurate approximate solution.

Key words. Algebraic algorithms, continued fractions, rational rounding.

1. Introduction. Numerical methods for the solution of systems of linear equations are usually classified in two main categories: direct and iterative. Most textbooks on the subject then continue by stating that direct methods are potentially capable of finding the exact solution, if exact arithmetic is used, in a finite number of steps. By contrast, iterative methods are presented under the framework that they can only provide us with an approximate solution.

This paper shows that the classification of methods for the solution of linear systems of equations as direct, implying exact, and iterative, implying approximate, is not entirely accurate. In fact, we show that any sufficiently close approximation to the solution leads to the exact rational solution of a system with integer coefficients with very little additional work.

As a consequence, the existing iterative methods and their huge supporting literature become available for utilization in the exact solution of systems of linear equations.

2. The main observation. We are interested in finding the exact rational solution to a system of linear equations with integer coefficients. We assume that the system has a unique solution.

The main observation to be made concerns the discrete nature of the problem. The solution vector can be found, for example with Gaussian elimination, in a finite number of arithmetic operations. As a consequence, the numerator and denominator of each rational in the solution cannot be arbitrarily large. So, there are only a finite number of rationals to be considered as candidates for the solution.

It is therefore possible, in principle, to find the solution to such a system simply by trying one-by-one all rationals, candidates to the solution. This brute force trial algorithm will obviously have an exponential computing time. Trying all candidates for the solution one-by-one is not a very good strategy.

The idea is stated more precisely as follows. Let

$$(1) \quad Ax = B$$

be the linear system to be solved. The coefficients of the array A , $a_{i,j}$, $1 \leq i, j \leq N$, and of the vector B , b_i , $1 \leq i \leq N$, are integers smaller, in absolute value, than some integer d .

Lemma 1 provides a tight bound on the size of the components of x in (1).

LEMMA 1 (Hadamard inequality). *Let $\det(A)$ be the determinant of the array A .*

* Received by the editors May 25, 1978. This research was supported in part by the National Science Foundation under grant GJ-28339A1, and by FAPESP, Fundacao de Amparo a Pesquisa do Estado de Sao Paulo, under grant mat. 70/725.

† 6E University Houses, Madison, Wisconsin 53705.

‡ Mathematics Department, University of Sao Paulo, Sao Paulo, Brazil.

Then

$$(2) \quad (\det(A))^2 \leq \prod_{1 \leq i \leq N} \left(\sum_{1 \leq j \leq N} a_{i,j}^2 \right).$$

With our bound, d , to the coefficients of (1) we can write

$$(3) \quad |\det(A)| \leq N^{N/2} * d^N \leq D,$$

for a suitable integer D . For a proof of Lemma 1 see, for example, [8, exercise 4.6.1.15].

So, if

$$(4) \quad x_i = \frac{p_i}{q_i}, \quad 1 \leq i \leq N,$$

is the solution to (1), we will have $|q_i| \leq D, 1 \leq i \leq N$.

The next lemma tells us that candidates for the solution of (1) are not too close to each other.

LEMMA 2 (minimum distance). *Let p/q and r/s be two rationals with $p/q \neq r/s$ and $|q| \leq D, |s| \leq D$. Then*

$$(5) \quad \min \left| \frac{p}{q} - \frac{r}{s} \right| \leq \frac{1}{D^2}.$$

Not only are there a finite number of candidates for the solution, but they are also reasonably far apart from each other.

3. A system of integer linear inequalities and its solution with continued fractions. Let us suppose we were able to find an approximation a/b to the true value p/q of some component of the solution of (1). If the distance between a/b and p/q is less than half the minimum distance between two candidates for the solution, then the nearest candidate to the approximation a/b will be p/q .

More precisely, the system of inequalities

$$(6) \quad \left| \frac{a}{b} - \frac{\alpha}{\beta} \right| \leq \frac{1}{(2 * D^2)}, \quad 1 \leq \beta \leq D,$$

with α and β as integer unknowns, has at most one solution. The uniqueness of a possible solution to (6) is guaranteed by Lemma 2.

Inequalities (6) can be rewritten as follows:

$$(7) \quad \begin{aligned} 2 * D^2 * b * \alpha - (2 * D^2 * a + b) * \beta &\leq 0, \\ -2 * D^2 * b * \alpha + (2 * D^2 * a - b) * \beta &\leq 0, \quad \beta \leq D, \quad -\beta \leq -1. \end{aligned}$$

The problem of determining whether a system of inequalities like (7) has a solution in integers and then, if some solution exists, actually finding one is in general NP-complete. Hirschberg and Wong [4] showed that integer systems of inequalities with only two unknowns are special. They can be solved in polynomial time. In our case it is simpler to find a solution of (7) by going back directly to the continued fraction algorithm.

Continued fractions are an old and venerable topic and have close ties with Euclid's algorithm. The first documented use of their approximating powers seems to have been done by Huygens [7]. He used continued fractions to compute the best number of teeth in pairs of gears to be used in the driving mechanism of a model of the solar system.

The key result that permits us to solve system (7) efficiently is contained in the following theorem.

THEOREM C (continued fractions approximations). *If*

$$\left| \frac{p}{q} - \frac{a}{b} \right| \leq \frac{1}{(2 * q^2)},$$

then p/q is a convergent in the continued fraction series for a/b .

Proof. For an algebraic proof, see [3, Thm. 184]. For a more geometric and somewhat more revealing approach, see [11, Thm. 7.19]. \square

Proofs of Theorem C lead directly to Algorithm C.

ALGORITHM C (computation of the continued fraction approximation). Given the integers $a \geq 0, b > 0, D > 0$, the algorithm computes, when they exist, two integers p', q' , such that $|a/b - p'/q'| \leq 1/(2 * D^2)$ and $q' \leq D$.

C1. [Initialize.] Set $p \leftarrow 0, q \leftarrow 1, p' \leftarrow 1, q' \leftarrow 0, A \leftarrow a, B \leftarrow b$.

C2. [Test for end.] If $B = 0$, then go to step C5.

C3. [Compute new approximation.] Set $W \leftarrow \lfloor A/B \rfloor, p'' \leftarrow p + W * p', q'' \leftarrow q + W * q'$. If $q'' > D$, then go to step C5.

C4. [Shift and go back.] Set $p \leftarrow p', q \leftarrow q', p' \leftarrow p'', q' \leftarrow q'', T \leftarrow A - B * W, A \leftarrow B, B \leftarrow T$; then go back to step C2.

C5. [Test for goodness and terminate.] If $|2 * D^2 * (a * q' - b * p')| \leq b * q'$ then the approximation to a/b is p'/q' ; otherwise the algorithm returns "NO SOLUTION" and terminates.

Algorithm C is an implementation of the extended Euclidean algorithm with the addition of tests in steps C3 and C5.

Step C3 selects one of the continued fraction convergents, namely, the one with the largest denominator smaller than the bound D . The choice follows from the following facts:

Fact 1. Each successive convergent, p/q , approximates a/b better and better.

Fact 2. Inequalities (6) have at most one solution.

Hence the only candidate to a solution of (6) is the convergent with the largest possible denominator.

Step C5 is not necessary if we know in advance that two integers, p, q , exist satisfying (6). For then we have

$$\left| \frac{a}{b} - \frac{p}{q} \right| \leq \frac{1}{(2 * D^2)} \leq \frac{1}{(2 * q^2)}$$

because $q \leq D$, and we can apply Theorem C.

In general, however, steps C1–C4 may fail to produce the required approximation. The test in step C5 becomes necessary to differentiate between a true solution to (6) and simply a continued fraction approximation p/q to a/b having $q \leq D$.

The worst case computing time of Algorithm C is of $O((\log N)^2)$, when all the inputs are bounded by some integer N . For a computing time analysis of Algorithm C, consult [8]. For many details of its practical implementation, consult Collins [1] and its bibliography.

4. The procedure. The results in §§ 2 and 3 suggest a two-step procedure for the computation of the exact solution of a linear system of equations with integer coefficients:

Step 1. Obtain an approximate solution that differs, in each component, from the true solution by less than $1/(2 * D^2)$. D is an integer bound for the denominators of the solution vector.

Step 2. Use Algorithm C to obtain the exact rational solution.

A computing time analysis of step 1 is difficult. The analysis is complicated by the fact that performance of iterative methods depends strongly on the particular problem being solved. Two improvements of a general nature in the computing time of step 1 are possible.

First, the bound D is in many cases too large. A much smaller bound and considerably fewer iterations might do. It might be more convenient to apply Algorithm C as a test of termination than to iterate up to the precision necessary to be certain that step 2 will produce the exact solution.

Second, the use of exact arithmetic to compute successive approximations to the solution will cause an increase in the size of the integers to be manipulated during each iteration. Algorithm C can be used to reduce the size of the integers in the approximation. Some care must be exercised, however, not to destroy the convergence of the underlying iterative method.

5. Conclusion. Continued fractions approximations can be used to obtain the exact rational solution to a problem whenever:

- (A) The denominator of the sought rational a/b is bounded* by some known integer D ;
- (B) It is possible to obtain an approximation p/q to the rational a/b , satisfying $|p/q - a/b| \leq 1/(2 * D^2)$.

The continued fractions algorithm closely resembles the extended Euclidean algorithm applied to the integers p and q .

As a last observation, consider all the real roots of all the polynomials of degree not greater than N and with integer coefficients bounded by D . Sufficiently small intervals will contain at most one of those roots. We would like to have an efficient algorithm that, given such an interval and given the bounds N and D , would choose a polynomial in our set of polynomials having a root in the given interval. Such an algorithm would allow the use of approximate methods for a wide range of exact computations with algebraic numbers.

The continued fractions algorithm solves the problem efficiently for $N = 1$.

Note added in proof. Since this paper was written, at the beginning of 1977, and presented at the SIAM 1978 Spring Meeting, following a seminar on the subject given at the University of Sao Paulo in 1976, a number of references relevant to its contents have been published. We list some of them:

- [1] O. ABERTH, *A method for exact computation with rational numbers*, J. Comp. Appl. Math., 4 (1978), pp. 285-288.
- [2] R. P. BRENT, F. G. GUSTAVSON AND D. Y. Y. YUN, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*. J. Algorithms, 1 (1980), pp. 259-295.
- [3] J. H. CARTER, *Power series and exact solution of systems of linear equations*, M.Sc. thesis, Dept. Computer Science, Univ. of Toronto, 1978.
- [4] M. GROTSCHER, L. LOVASZ AND A. SCHRIJVER, *The ellipsoid method and its consequences in combinatorial optimization*, Report 80151-OR, Univ. of Bonn, W. Germany, 1980.
- [5] F. GUSTAVSON AND D. Y. Y. YUN, *Fast algorithms for rational Hermite approximation and solution of Toeplitz systems*, IEEE Trans. Circuits and Systems, 9 (1979), pp. 750-755.
- [6] D. W. MATULA AND P. KORNERUP, *Approximate rational arithmetic systems: analysis of recovery of simple fractions during expression evaluation*, Lecture Notes in Computer Science 72, Springer-Verlag, New York, 1979, pp. 383-397.

- [7] R. MOENCK AND J. CARTER, *Approximate algorithms to derive exact solutions to systems of linear equations*, Lecture Notes in Computer Science 72, Springer-Verlag, New York, 1979, pp. 65–73.
- [8] S. URSIC, *The ellipsoid algorithm for linear inequalities in exact rational arithmetic*, IEEE Symposium on Foundations of Computer Science 23, 1982.
- [9] D. YUN AND F. GUSTAVSON, *Fast computation of rational Hermite interpolation and solving Toeplitz systems of equations via the extended Euclidean algorithm*, Lecture Notes in Computer Science 72, Springer-Verlag, New York, 1979, pp. 58–64.

REFERENCES

- [1] G. E. COLLINS, *Computer algebra of polynomial and rational functions*, Amer. Math. Monthly, 80 (1973), pp. 725–755.
- [2] G. E. COLLINS AND E. HOROWITZ, *The minimum root separation of a polynomial*, Math. Comp., 28 (1974), pp. 589–597.
- [3] G. E. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Oxford Univ. Press, Cambridge, 1960, pp. 129–151.
- [4] D. S. HIRSCHBERG AND C. K. WONG, *A polynomial-time algorithm for the knapsack problem with two variables*, J. Assoc. Comput. Mach., 23 (1976), pp. 147–154.
- [5] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.
- [6] C. HUYGENS, *Descriptio automati planetarii*, Opuscula Posthuma, Amsterdam, 1728, pp. 174–179.
- [7] R. M. KARP, *On the computational complexity of combinatorial problems*, Networks, 5 (1975), pp. 45–68.
- [8] D. KNUTH, *The Art of Computer Programming. Vol. II: Seminumerical Algorithms*, Addison-Wesley, Reading, MA, 1968.
- [9] M. T. MCCLELLAN, *The exact solution of systems of linear equations with polynomial coefficients*, J. Assoc. Comput. Mach., 20 (1973), pp. 563–588.
- [10] ———, *A comparison of algorithms for the exact solution of systems of linear equations*, ACM Trans. Math. Software, 3 (1977), pp. 147–158.
- [11] H. M. STARK, *An Introduction to Number Theory*, Markham, Chicago, 1971, pp. 181–239.
- [12] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
- [13] D. M. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

LINEAR TRANSFORMATIONS ON NONNEGATIVE MATRICES PRESERVING PROPERTIES OF IRREDUCIBILITY AND FULL INDECOMPOSABILITY*

RICHARD A. BRUALDI† AND LI QIAO‡

Abstract. We characterize linear transformations on the vector space of $n \times n$ matrices of the form $A \rightarrow PAQ$ which preserve the combinatorial properties of irreducibility, reducibility, near reducibility, full indecomposability, decomposability, and near decomposability for nonnegative matrices. The minimality of such pairs P, Q is also considered.

1. Introduction. Let M_n denote the vector space of $n \times n$ matrices over the complex number field. The problem of determining those linear transformations T of M_n which preserve a given invariant has been extensively studied. For a survey of results and an extensive bibliography see Marcus [6], [7]. A classical result due to Frobenius [3] asserts that if T preserves determinant, that is $\det(T(A)) = \det A$ for each $n \times n$ matrix A , then there exist matrices P and Q with $\det(PQ) = 1$ such that either

$$(1.1) \quad T(A) = PAQ \quad \text{for all } A \in M_n,$$

or

$$(1.2) \quad T(A) = PA^tQ \quad \text{for all } A \in M_n.$$

The conclusion that T have one of the forms (1.1) or (1.2) has been shown to hold for other invariants. Thus, for instance, Marcus and Moysl [8] proved that if $T(A)$ is a rank 1 matrix for each rank 1 matrix A , then T satisfies (1.1) or (1.2) where P and Q are invertible matrices. In this note we consider linear transformation T on M_n of the form (1.1) which take nonnegative matrices to nonnegative matrices and preserve properties of irreducibility, reducibility, near reducibility, full indecomposability, decomposability, and near decomposability for *these* matrices. Results similar to ones we obtain hold also for linear transformations T of the form (1.2). Minc [9] characterizes linear transformations T which take nonnegative matrices to nonnegative matrices and preserve the spectrum of each nonnegative matrix.

Suppose $P = [p_{ij}]$ and $Q = [q_{ij}]$ are $n \times n$ matrices different from the zero matrix such that PAQ is a nonnegative matrix for each $n \times n$ nonnegative matrix A . Let E_{ij} be the $n \times n$ matrix all of whose entries are 0 except for the (i, j) -entry which is 1. By considering the matrices E_{ij} , it follows that $p_{rs}q_{uv} \geq 0$ for each entry p_{rs} of P and each entry q_{uv} of Q . Hence there exists a real number θ such that $P = e^{i\theta}P_1$ and $Q = e^{-i\theta}Q_1$ where P_1 and Q_1 are nonnegative matrices. In addition $PAQ = P_1AQ_1$. This is why we restrict our attention to linear transformation T of the form (1.1) where P and Q are nonnegative matrices.

2. Preliminaries. Let A be an $n \times n$ matrix. If each entry of A is nonnegative, we say A is *nonnegative* and write $A \geq 0$. Let $K, L \subseteq \{1, \dots, n\}$. Then $A[K, L]$ denotes the submatrix of A whose rows are indexed by the integers in K and whose columns are indexed by the integers in L . If $L = \{1, \dots, n\}$, we write $A[K, \cdot]$; if $K = \{1, \dots, n\}$, we write $A[\cdot, L]$. For $K \subseteq \{1, \dots, n\}$, \bar{K} denotes the *complement* of K .

* Received by the editors January 25, 1982, and in revised form June 1, 1982.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

‡ Department of Mathematics, China University of Science and Technology, People's Republic of China, and University of Wisconsin, Madison, Wisconsin 53706.

The $n \times n$ matrix $A = [a_{ij}]$ is called *reducible* if there exists a $K \subseteq \{1, \dots, n\}$ with both K and \bar{K} nonempty such that $A[K, \bar{K}]$ is a zero matrix. It follows that A is reducible if and only if there is a permutation matrix R such that

$$RAR^t = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$$

where A_1 and A_2 are square (nonvacuous) matrices. The matrix A is *irreducible* if it is not reducible. Finally, the matrix A is *nearly reducible* if it is irreducible and the replacement of any nonzero entry by a 0 results in a reducible matrix. All entries on the main diagonal of a nearly reducible matrix equal 0. With the matrix A there is associated a directed graph $D(A)$. The vertices of $D(A)$ are $1, 2, \dots, n$. There is an arc from i to j provided $a_{ij} \neq 0$. It is well known [11] that A is irreducible if and only if $D(A)$ is strongly connected. Here, that a directed graph is *strongly connected* means that for each ordered pair of vertices r, s there is a path from r to s . The directed graph of a nearly reducible matrix is strongly connected and the removal of any arc results in a directed graph which is not strongly connected. Irreducible matrices and their directed graphs have an important role in the spectral theory of nonnegative matrices.

The $n \times n$ matrix A is called (*partly*) *decomposable* if there exist $K, L \subseteq \{1, \dots, n\}$ with both K and L nonempty and with $|K| + |L| = n$ such that $A[K, L]$ is a zero matrix. Thus A is decomposable if and only if there are permutation matrices R and S such that

$$RAS = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$$

where A_1 and A_2 are square (nonvacuous) matrices. The matrix A is *fully indecomposable* provided it is not decomposable. Finally the matrix A is *nearly decomposable* if it is fully indecomposable but the replacement of any nonzero entry by a 0 results in a decomposable matrix.

The connection between irreducible and fully indecomposable matrices is well known [2, p. 33]. The matrix A is fully indecomposable if and only if for some permutation matrix R , RA has no zeros on the main diagonal and is irreducible. However, the relation between nearly reducible and nearly decomposable matrices is not so conclusive. For more information see [1].

A nearly reducible matrix must have at least one nonzero entry in each row and in each column; a nearly decomposable matrix must have at least two. Later we shall need one additional structural property for each of these two types of matrices. For nearly decomposable matrices this property was found by Hartfiel [4] and later in the context of bipartite graphs by Lovász and Plummer [5].

LEMMA 2.1. *Let A be an $n \times n$ nearly decomposable matrix with $n > 2$. Then every 2×2 submatrix of A contains a 0.*

A similar conclusion holds for nearly reducible matrices:

LEMMA 2.2. *Let A be an $n \times n$ nearly reducible matrix. Then every 2×2 submatrix of A contains a 0.*

Proof. Suppose there exist $K, L \subseteq \{1, \dots, n\}$ such that $|K| = |L| = 2$ and $A[K, L]$ contains no 0's. If $K \cap L \neq \emptyset$, then A has a nonzero entry on its main diagonal so that A is not nearly reducible. Hence $K \cap L = \emptyset$. Let $K = \{i, j\}$ and $L = \{r, s\}$ so that i, j, r, s are distinct. Now consider the directed graph $D(A)$. Define a *chord* of a circuit (i.e. a path joining a vertex to itself) to be an arc joining two vertices of the circuit which is not an arc of the circuit. Since A is nearly reducible, no circuit of $D(A)$ can

have a chord, since the removal of such a chord would leave a strongly connected directed graph. In $D(A)$ there are arcs (i, r) , (i, s) , (j, r) and (j, s) . Since $D(A)$ is strongly connected there is a simple path γ from r to i for which (i, r) is not an arc. Similarly, there is a simple path δ from s to j for which (j, s) is not an arc (see Fig. 1). Suppose (i, r) is not an arc of δ . Then the sequence $\gamma, (i, s), \delta, (j, r)$ determines a circuit for which (i, r) is a chord. Hence (i, r) is an arc of δ , and similarly, (j, s) is an arc of γ . Then γ contains a path γ' from r to j , and δ contains a path δ' from s to i (see Fig. 2). Since γ is a simple path, (j, r) is not an arc of γ' ; since δ is a simple path, (j, r) is not an arc of δ' . A similar conclusion holds for (i, s) . But then the sequence $\gamma', (j, s), \delta', (i, r)$ determines a circuit for which (j, r) is a chord. This is a contradiction, and the proof is complete.

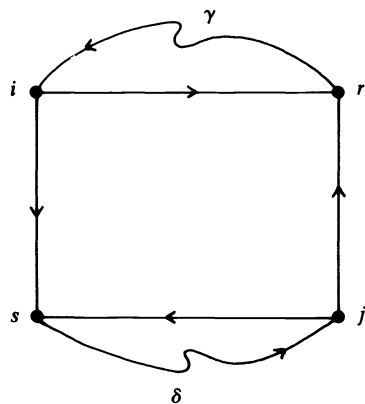


FIG. 1

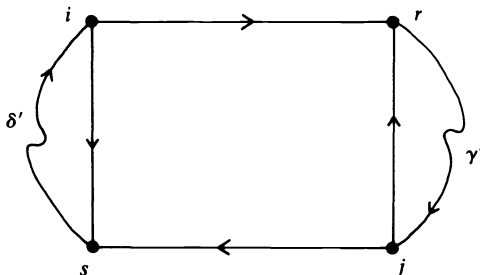


FIG. 2

3. Irreducibility. We consider in this section linear transformations of M_n of the form $A \rightarrow PAQ$ satisfying:

- (3.1) PAQ is nonnegative whenever A is;
- (3.2) PAQ is irreducible (or nearly reducible, or reducible) whenever the non-negative matrix A is.

As already observed, (3.1) implies that P and Q are nonnegative matrices. Since the properties of being irreducible, nearly reducible, and reducible are combinatorial properties of a matrix, there is no loss of generality in assuming that P and Q are

matrices of 0's and 1's. In addition, it suffices to consider in (3.2) only matrices A of 0's and 1's. The $n \times 1$ column of all 1's is denoted by e .

LEMMA 3.1. *Let P and Q be $n \times n$ matrices 0's and 1's and let R be a permutation matrix. Then the following are equivalent:*

(3.3) PAQ is irreducible for each $n \times n$ irreducible matrix A of 0's and 1's;

(3.4) $(PR)A(R'Q)$ is irreducible for each $n \times n$ irreducible matrix A of 0's and 1's.

Proof. This is a simple consequence of the fact that A is irreducible if and only if RAR' is.

THEOREM 3.2. *Let $P = [p_{ij}]$ and $Q = [q_{ij}]$ be $n \times n$ matrices of 0's and 1's. Let the columns of P be $\alpha_1, \dots, \alpha_n$ and the rows of Q be β_1, \dots, β_n . Then the following are equivalent:*

(3.5) For each $n \times n$ irreducible matrix A of 0's and 1's, PAQ is irreducible.

- (3.6) (i) P and Q' have no zero rows;
 (ii) for each $i = 1, \dots, n$, either
 (ia) $\alpha_i = e$ or $\beta_i = e'$, or
 (ib) $\beta_i \alpha_i > 0$.

Proof. First suppose that (3.5) holds. If P has a zero row, so does PAQ ; if Q has a zero column, so does PAQ . Since an irreducible matrix can have no zero rows or columns, (3.6)(i) holds. Suppose (3.6)(ii) does not hold. It follows from Lemma 3.1 that we may assume that both (3.6)(ia) and (ib) fail for $i = 1$. Hence there exists $K \subseteq \{1, \dots, n\}$ with $K \neq \emptyset$ and $\bar{K} \neq \emptyset$ such that $p_{i1} = 0$ for $i \in K$ and $q_{1j} = 0$ for $j \in \bar{K}$. Let $A = [a_{ij}]$ be the $n \times n$ irreducible matrix for which $a_{12} = \dots = a_{1n} = a_{21} = \dots = a_{n1} = 1$ and $a_{ij} = 0$, otherwise. Then for $B = PAQ = [b_{ij}]$ and for $i \in K$ and $j \in \bar{K}$, we calculate that

$$b_{ij} = \sum_{s=1}^n \sum_{r=1}^n p_{ir} a_{rs} q_{sj} = \sum_{s=2}^n p_{i1} a_{1s} q_{sj} + \sum_{r=2}^n p_{ir} a_{r1} q_{1j} = 0 + 0 = 0.$$

Hence $B[K, \bar{K}] = 0$ so that B is reducible, a contradiction. Thus (3.6)(ii) holds.

Now suppose (3.6) holds. Let $K \subseteq \{1, \dots, n\}$ with $K \neq \emptyset$ and $\bar{K} \neq \emptyset$. Let $1 \leq j \leq n$. Then it follows from (3.6)(ii) that either there exists $r \in K$ for which $p_{rj} = 1$ or there exists $s \in \bar{K}$ for which $q_{js} = 1$. Since K and \bar{K} are nonempty, it now follows from (3.6)(i) that there exists $I \subseteq \{1, \dots, n\}$ with $I \neq \emptyset$ and $\bar{I} \neq \emptyset$ such that for $i \in I$ there exists $r \in K$ with $p_{ri} = 1$, and for $k \in \bar{I}$ there exists $s \in \bar{K}$ with $q_{ks} = 1$. Now let $A = [a_{ij}]$ be an $n \times n$ irreducible matrix of 0's and 1's and let $B = PAQ = [b_{ij}]$. Since A is irreducible, we can choose $i \in I$ and $k \in \bar{I}$ such that $a_{ik} = 1$. But then using the above notation, we see that

$$b_{rs} = \sum_{u,v} p_{ru} a_{uv} q_{vs} \geq p_{ri} a_{ik} q_{ks} = 1.$$

It follows that $B[K, \bar{K}] \neq 0$. Since this is true for each $K \subseteq \{1, \dots, n\}$ with K and \bar{K} nonempty, it follows that $B = PAQ$ is irreducible. Since this is true for each irreducible A , (3.5) holds.

COROLLARY 3.3. *Let P and Q be $n \times n$ matrices of 0's and 1's such that P and Q have neither zero rows nor zero columns. Then PAQ is irreducible for each $n \times n$ irreducible matrix A of 0's and 1's if and only if $QP \geq I_n$, the $n \times n$ identity matrix.*

Proof. First suppose PAQ is irreducible for each $n \times n$ irreducible matrix A of 0's and 1's. Then by Theorem 3.2, (3.6)(ii) holds. Since P has no zero column and

Q has no zero row, it follows that (3.6)(iib) holds for each $i = 1, \dots, n$. Hence $QP \geq I_n$.

Conversely, if $QP \geq I_n$, then (3.6)(iib) holds for each $i = 1, \dots, n$. Since P has no zero rows and Q has no zero columns, we conclude (3.6) holds. Hence by Theorem 3.2, PAQ is irreducible for each irreducible matrix A of 0's and 1's.

COROLLARY 3.4. *Let Q be an $n \times n$ permutation matrix and let P be an $n \times n$ matrix of 0's and 1's. Then PAQ is irreducible for each $n \times n$ irreducible matrix A of 0's and 1's if and only if $P \geq Q^t$.*

Proof. If $P \geq Q^t$, then for an $n \times n$ matrix A of 0's and 1's, $PAQ \geq Q^tAQ$; thus PAQ is irreducible when A is. Conversely, suppose PAQ is irreducible for each irreducible matrix A of 0's and 1's. Then (3.6) holds. Since Q has no row of all 1's, P can have no column of all 0's. Hence (3.6)(iib) holds, that is $QP \geq I_n$ or $P \geq Q^t$.

An example of a pair of matrices P, Q satisfying the equivalent conditions of Theorem 3.2 for which $QP \not\geq I_n$ is the following:

$$P = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

We now consider nearly reducible matrices. Let P and Q be $n \times n$ matrices such that PAQ is nearly reducible for each $n \times n$ nearly reducible matrix A of 0's and 1's. Let B be an $n \times n$ irreducible matrix of 0's and 1's. Then there exists a nearly reducible matrix C of 0's and 1's with $B \geq C$. So $PBQ \geq PCQ$. Since PCQ is nearly reducible, PBQ is irreducible. Hence PBQ is irreducible for each irreducible matrix B of 0's and 1's. But the converse does not hold, as is easily seen. Indeed we have the following.

THEOREM 3.5. *Let $P = [p_{ij}]$ and $Q = [q_{ij}]$ be $n \times n$ matrices of 0's and 1's. Then PAQ is nearly reducible for each $n \times n$ nearly reducible matrix A of 0's and 1's if and only if P and Q are permutation matrices and $Q = P^t$.*

Proof. If P is a permutation matrix and $Q = P^t$, then by definition PAQ is nearly reducible for each nearly reducible matrix A . Conversely, suppose PAQ is nearly reducible for each reducible matrix A of 0's and 1's. We prove that P is a permutation matrix and $Q = P^t$ by consideration of a number of cases.

Case 1. Q is a permutation matrix. It follows from Corollary 3.4 that $P \geq Q^t$. Suppose $P \neq Q^t$. Then for some i , row i of P contains at least two 1's. Let r and s be chosen so that $p_{ir} = q_{ri} = 1$ and $p_{is} = 1$, where $r \neq s$. We may choose a nearly reducible matrix A whose r th row is different from its s th row. It follows that $PAQ \geq Q^tAQ$ but $PAQ \neq Q^tAQ$. Since Q^tAQ is nearly reducible, PAQ is not nearly reducible. This contradiction implies that $P = Q^t$.

Case 2. P is a permutation matrix. This case is similar to Case 1.

Case 3. Some column of P contains at least two 1's. Using the analogue of Lemma 3.1 for nearly reducible matrices, we may suppose column 1 of P contains 1's in rows i and j where $i \neq j$. Let A be the $n \times n$ nearly reducible matrix

$$\begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & & & \\ \vdots & & 0 & \\ 1 & & & \end{bmatrix}.$$

Then rows i and j of PA equal $[*1 \dots 1]$. If $Q \geq R$ for some permutation matrix R , then PAQ has a 1 on the main diagonal and cannot be nearly reducible. So $Q \not\geq R$ for each permutation matrix R . Since Q has no zero columns, it now follows that

some row of Q contains at least two 1's. Let row k of Q contain 1's in columns u and v . First suppose $k \neq 1$. Then it follows that the 2×2 submatrix of PAQ formed by rows i and j and columns u and v contains no 0's. Hence by Lemma 2.2, PAQ is not nearly reducible, a contradiction. Thus $k = 1$ and only row 1 of Q can contain more than one 1.

Let there be $r \geq 2$ 1's in row 1 of Q . Suppose $n - r \geq 2$, and let row 1 of Q contain 0's in columns p and q . Since each column of Q contains at least one 1, it follows that the 2×2 submatrix of PAQ formed by rows i and j and columns p and q contains no 0's. Using Lemma 2.2 again, we obtain a contradiction. Hence $n - r = 0$ or 1. First suppose $n - r = 0$. Then row 1 of Q contains only 1's, and AQ has no 0's in rows $2, \dots, n$. If P had a 1 in a column other than column 1, then PAQ would have a row with no 0's and would not be nearly reducible. It follows that

$$P = \begin{bmatrix} 1 & & \\ \vdots & & \\ 1 & & 0 \end{bmatrix}.$$

Since PBQ is irreducible for each irreducible matrix B of 0's and 1's, and row 1 of Q contains only 1's, it now follows from Theorem 3.2 that all entries of Q equal 1. This contradicts the fact that only row 1 of Q can contain more than one 1 (or, use the fact that PAQ now contains no 0's).

Now suppose $n - r = 1$. Without loss of generality we may assume the unique 0 in row 1 of Q occurs in column n . Since no column of Q contains only 0's, there is a 1 in column n of Q in one of rows $2, \dots, n$. Hence the only 0's of AQ occur in columns $1, \dots, n - 1$ of row 1 and rows $2, \dots, n$ of column n . Suppose two columns other than column 1 of P contained a 1. Then PAQ would have two rows each containing at most one 0 and hence a nonzero entry on the main diagonal. We conclude that at most one column of P other than column 1 contains a 1. It now follows from Theorem 3.2, that Q has at least $n - 2$ rows of all 1's. This contradicts the fact that only row 1 of Q can contain more than one 1. This contradiction now means that this case cannot occur.

Case 4. Some row of Q contains at least two 1's. An argument similar to the above shows that this case cannot occur.

Since each row of p and each column of Q contains a 1, there are no cases left to be considered. We conclude that P is a permutation matrix and $Q = P^t$, and the proof is complete.

Finally we consider reducible matrices, but first we prove the following lemma concerning decomposable matrices.

LEMMA 3.6. *Let P and Q be $n \times n$ matrices of 0's and 1's such that P has no zero rows and Q has no zero columns. Then PAQ is decomposable for every $n \times n$ matrix A of 0's and 1's having a zero row or column if and only if P and Q are permutation matrices.*

Proof. Suppose PAQ is decomposable for every A with a zero row or column. Let A_1 be the matrix all of whose entries equal 1 except those in the first row which equal 0. Since Q has no zero columns, it follows that $A_1Q \geq A_1$. Let $P' = P[\cdot, \{2, \dots, n\}]$ and let $A'_1 = A_1[\{2, \dots, n\}, \cdot]$ so that all entries of A'_1 equal 1. Then

$$PA_1Q \geq PA_1 = P'A'_1.$$

Since PA_1Q is decomposable, it now follows that P' has a zero row. Since P has no zero rows, we conclude that some row of P equals $(1, 0, \dots, 0)$. By considering the

matrix A_i all of whose entries equal 1 except those in row i which equal 0 ($i = 1, \dots, n$), we conclude in a similar way, that for each $i = 1, \dots, n$ some row of P contains only 0's except for a 1 in column i . Hence P is a permutation matrix. In the same way, by considering the matrices A_i^j we conclude that Q is also a permutation matrix.

The converse is evident.

THEOREM 3.7. *Let P and Q be $n \times n$ matrices of 0's and 1's. Then PAQ is reducible for each $n \times n$ reducible matrix A of 0's and 1's if and only if one of the following holds:*

$$(3.7) \quad P \text{ has a zero row or } Q \text{ has a zero column.}$$

$$(3.8) \quad P \text{ and } Q \text{ are permutation matrices and } Q = P^t.$$

Proof. First suppose PAQ is reducible for each reducible A . Suppose (3.7) does not hold. Since a matrix with a zero row or column is reducible and since a reducible matrix is decomposable, it follows from Lemma 3.5 that P and Q are permutation matrices. Since $PAQ = (PQ)(Q^tAQ)$, it follows that $(PQ)B$ is reducible for each $n \times n$ reducible matrix B of 0's and 1's. Since PQ is a permutation matrix, it follows easily that $PQ = I_n$. Hence $Q = P^t$. The converse is clear.

If we insist upon both the properties of being irreducible and being reducible be preserved, we obtain the following.

THEOREM 3.8. *Let P and Q be $n \times n$ matrices of 0's and 1's. Then the following are equivalent:*

$$(3.9) \quad \text{For each } n \times n \text{ matrix } A \text{ of 0's and 1's } PAQ \text{ is irreducible if and only if } A \text{ is.}$$

$$(3.10) \quad P \text{ and } Q \text{ are permutation matrices and } Q = P^t.$$

Proof. Suppose (3.9) holds. Then P can have no zero row and Q can have no zero column, and it follows from Theorem 3.7 that (3.10) holds. That (3.10) implies (3.9) is obvious.

To conclude this section we consider $n \times n$ matrices P and Q of 0's and 1's which satisfy the equivalent conditions (3.5) and (3.6) of Theorem 3.2, and which have the additional property that the replacement of a 1 by a 0 (in either P or Q) always results in matrices P' and Q' not satisfying (3.5) and (3.6). We call such a pair of matrices (P, Q) a *minimal pair for irreducibility*.

Let n be a positive integer. By \mathcal{P}_n we denote the set of all $n \times n$ matrices P of 0's and 1's for which there exists an $n \times n$ matrix Q of 0's and 1's such that (P, Q) is a minimal pair for irreducibility. The set \mathcal{Q}_n is defined in an analogous way. It follows that if (P, Q) is minimal pair of $n \times n$ matrices for irreducibility, then $P \in \mathcal{P}_n$ and $Q \in \mathcal{Q}_n$, but the converse need not hold. For a $P \in \mathcal{P}_n$, we denote by $\mathcal{Q}_n(P)$ the set of all $n \times n$ matrices Q for which (P, Q) is a minimal pair for irreducibility. We investigate the sets \mathcal{P}_n and $\mathcal{Q}_n(P)$ for $P \in \mathcal{P}_n$.

Let P be an $n \times n$ matrix of 0's and 1's. We say that P is in *standard form* (with respect to columns) provided

$$(3.11) \quad P = [P_1 J_{n,v} O_{n,w}]$$

where $J_{n,v}$ is an $n \times v$ matrix of all 1's (possibly vacuous), $O_{n,w}$ is an $n \times w$ matrix of all 0's (possibly vacuous), and P_1 is an $n \times (n - v - w)$ matrix (possibly vacuous) having at least one 1 and at least one 0 in each column. It follows from Lemma 3.1 that there is no loss of generality in assuming that P is in standard form.

THEOREM 3.9. *Let P be an $n \times n$ matrix of 0's and 1's in standard form (3.11) where $v > 0$ and $w > 0$. Then $P \in \mathcal{P}_n$ if and only if each column of P_1 contains exactly one 1. If $P \in \mathcal{P}_n$, then*

$$\begin{bmatrix} P_1^t \\ O_{v,n} \\ J_{w,n} \end{bmatrix}$$

is the unique matrix in $Q_n(P)$.

Proof. This theorem is readily checked using the conditions of (3.6). The fact that $v > 0$ guarantees that P has no zero rows; the fact that $w > 0$ guarantees that a matrix $Q \in Q_n(P)$ will have row of 1's and hence no zero column.

THEOREM 3.10. *Let P be an $n \times n$ matrix of 0's and 1's in standard form (3.11) where $v > 0$ and $w = 0$. Assume the rows of P have been permuted so that*

$$P = \begin{bmatrix} P_2 & \\ 0 & J_{n,v} \end{bmatrix}$$

where P_2 has no zero rows. Then $P \in \mathcal{P}_n$ if and only if each column of P_2 contains exactly one 1. If $P \in \mathcal{P}_n$, then $Q_n(P)$ consists of all $n \times n$ matrices of the form

$$Q = \begin{bmatrix} P_2^t & Q_2 \\ 0 & \end{bmatrix}$$

where each column of Q_2 contains exactly one 1.

Proof. Again this theorem is readily verified using the conditions of (3.6). The fact that $v > 0$ guarantees that P has no zero rows; the condition on Q_2 is needed only to guarantee that Q has no zero columns; the last v rows of Q can contain only 0's for otherwise we could replace a 1 in the submatrix $J_{n,v}$ of P by 0, contradicting $P \in \mathcal{P}_n$.

When we assume $v = 0$, then the matrix P in standard form (3.11) does not automatically have the property that it has no zero rows and becomes

$$(3.12) \quad P = [P_1 \quad O_{n,w}]$$

where P_1 is an $n \times (n - w)$ matrix having at least one 0 and at least one 1 in each column. For P to satisfy (3.6), P_1 must have a 1 in each row. In this case the structure of P when $P \in \mathcal{P}_n$ is more complicated.

LEMMA 3.11. *Let P be an $n \times n$ matrix of 0's and 1's in standard form (3.12). If $P \in \mathcal{P}_n$, then P has no 2×2 submatrix of all 1's and no 2×3 submatrix whose columns can be permuted to give the form*

$$(3.13) \quad \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

Proof. Suppose $P \in \mathcal{P}_n$. Then there exists a matrix $Q \in Q_n(P)$ such that P, Q satisfy (3.6). Let $\alpha_1, \dots, \alpha_n$ be the columns of P and β_1, \dots, β_n the rows of Q . For $i = 1, \dots, n - w$, α_i contains both a 0 and a 1. Hence (3.6)(ii) is satisfied if and only if $\beta_{n-w+1} = \dots = \beta_n = e^t$ and $\beta_i \alpha_i > 0$ for $i = 1, \dots, n - w$. It follows that if a 2×2 submatrix of P contains all 1's then one of these 1's can be replaced by 0 to yield an $n \times n$ matrix P' such that P', Q satisfy (3.6). Since (P, Q) is a minimal pair for irreducibility, this is a contradiction. Hence P has no 2×2 submatrix of all 1's. It now follows that every 2×3 submatrix of P contains at most four 1's with equality if and

only if the columns of the submatrix can be permuted to give (3.13). If P has such a 2×3 submatrix, one of its 1's can be replaced by 0 to again contradict the fact that (P, Q) is a minimal pair. Hence the lemma holds.

Let $P = [P_{ij}]$ be an $n \times n$ matrix of 0's and 1's in standard form (3.12) such that $P \in \mathcal{P}_n$. The bipartite graph $G(P)$ associated with P has vertices $x_1, \dots, x_n, y_1, \dots, y_n$ with an edge $[x_i, y_j]$ in $G(P)$ if and only if $p_{ij} = 1$. It follows from Lemma 3.10, that $G(P)$ has no cycles of length 4, and a path beginning at a vertex in $Y = \{y_1, \dots, y_n\}$ has length at most 3. Consequently, $G(P)$ has no cycles of any length and a path beginning at a vertex in $X = \{x_1, \dots, x_n\}$ has length at most 4. In particular, the connected components of $G(P)$ are trees with the maximum length l of a path at most 4. It follows readily that the connected components of $G(P)$ correspond to $s \times t$ submatrices of P whose rows and columns can be permuted to one of the forms:

$$(3.14) \quad l \cong 3: \quad \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & & & \\ \vdots & & & \\ 1 & & 0 & \end{bmatrix},$$

where $s \cong 1$ and $t \cong 1$, or

$$(3.15) \quad l = 4: \quad \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & 0 & & 0 & & & \\ \vdots & \vdots & & \vdots & & & \\ 1 & 0 & & \vdots & & & \\ 0 & 1 & & \vdots & & & \\ \vdots & \vdots & \cdots & & & O_{s-1,r} & \\ \vdots & 1 & & 0 & & & \\ & 0 & & 1 & & & \\ & \vdots & & \vdots & & & \\ 0 & 0 & & 1 & & & \end{bmatrix},$$

where $s \cong 3$ and $t - r \cong 2$.

THEOREM 3.12. *Let P be an $n \times n$ matrix of 0's and 1's in standard form (3.12). Then $P \in \mathcal{P}_n$ if and only if there are permutation matrices U and V such that*

$$(3.16) \quad UPV = \left[\begin{array}{cccc|c} R_1 & 0 & \cdots & 0 & \\ 0 & R_2 & \cdots & 0 & \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & R_m & \end{array} O_{n,w} \right]$$

where $m \cong 1$ and for $i = 1, \dots, m$, R_i has one of the forms (3.14) and (3.15).

Proof. It follows from the discussion preceding the theorem, that if $P \in \mathcal{P}_n$, there exist permutation matrices U and V such that (3.16) holds. Conversely, suppose (3.16) holds. Without loss of generality, we may assume $U = V = I_n$. We exhibit a matrix Q such that (P, Q) is a minimal pair for irreducibility.

First suppose $w = 0$, so that in (3.16), $O_{n,w}$ is vacuous. Let

$$Q^t = \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Q_m \end{bmatrix}$$

where the partitioning of the rows and columns of Q^t is the same as for P and where

$$(i) \quad Q_i = \begin{bmatrix} 1 & \cdots & 1 \\ & & \vdots \\ & 0 & \\ & & 1 \end{bmatrix},$$

if R_i has the form (3.14) or R_i has the form (3.15) with $r > 0$, and

$$(ii) \quad Q_i = \left[\begin{array}{c|c|c} 1 & 1 & 1 \cdots 1 \\ \hline a & b & O_{s-1,t-2} \end{array} \right],$$

if R_i has the form (3.15) with $r = 0$, where the column vector a is obtained from the corresponding column vector of P_i by interchanging 0's and 1's and b has 1's exactly in those rows where a has 0's.

It is readily checked using Theorem 3.2 that (P, Q) is a minimal pair for irreducibility.

Now suppose $w > 0$. Let

$$Q^t = \left[\begin{array}{cccc|c} Q_1 & 0 & \cdots & 0 & \\ 0 & Q_2 & \cdots & 0 & \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & Q_n & \end{array} \right] J_{n,w}$$

where the partitioning of the rows and columns of Q^t is the same as for P and where, for $i = 1, \dots, m$,

$$Q_i = \begin{bmatrix} 1 & \cdots & 1 \\ & & 0 \end{bmatrix}.$$

Again it follows in a straightforward manner that (P, Q) is a minimal pair for irreducibility, and the theorem is proved.

When P has the standard form (3.12) and $P \in \mathcal{P}_n$, it seems that the set $Q_n(P)$ does not admit a compact characterization. However, we do have the following:

THEOREM 3.13. *Let $P = [p_{ij}]$ be an $n \times n$ matrix of 0's and 1's such that $P \in \mathcal{P}_n$, P has no column of all 1's and no column of all 0's. Then (P, Q) is a minimal pair for irreducibility if and only if there exist matrices $X = [x_{ij}]$ and $Y = [y_{ij}]$ such that $Q^t = X + Y$ and*

- (i) $X \leq P$ and every column of X has exactly one 1.
- (ii) For $i = 1, \dots, n$, row i of Y contains all 0's if row i of X contains at least one 1, while row i of Y has exactly one 1 if row i of X contains all 0's; in the latter case $y_{ij} = 1$ implies $p_{ij} = 0$.

Proof. This theorem follows from Theorem 3.2 in view of the minimality assumptions on P and Q .

4. Full indecomposability. In this section we consider linear transformations of M_n of the form $A \rightarrow PAQ$ satisfying:

- (i) PAQ is nonnegative whenever A is,
- (ii) PAQ is fully indecomposable (or nearly decomposable or decomposable) whenever the nonnegative matrix A is.

As in the previous section we may assume that P and Q are matrices of 0's and 1's, and we need only consider in (ii) matrices A of 0's and 1's. We then have the following result:

THEOREM 4.1. *Let $P = [p_{ij}]$ and $Q = [q_{ij}]$ be $n \times n$ matrices of 0's and 1's. Then the following are equivalent:*

- (4.1) *For each $n \times n$ fully indecomposable matrix A of 0's and 1's, PAQ is fully indecomposable.*
- (4.2)
 - (i) *P and Q^t have no zero rows.*
 - (ii) *There do not exist nonempty proper subsets J, K, L, M of $\{1, \dots, n\}$ such that $P[J, K] = 0$, $Q[L, M] = 0$, $|J| + |M| = n$, and $|K| + |L| = n + 1$.*

Proof. First suppose that (4.1) holds. Since a fully indecomposable matrix can have no zero rows or columns, it follows that (4.2)(i) holds. Suppose (4.2)(ii) does not hold. Then there exist subsets J, K, L, M of $\{1, \dots, n\}$ with $\emptyset \neq J, K, L, M \neq \{1, \dots, n\}$ and with $|J| + |M| = n$ and $|K| + |L| = n + 1$, such that

$$P[J, K] = 0, \quad Q[L, M] = 0.$$

Let A be the $n \times n$ matrix of 0's and 1's such that $A[\bar{K}, \bar{L}] = 0$ and all other entries of A equal 1. Since $|\bar{K}| + |\bar{L}| = 2n - |K| - |L| \leq n - 1$, the matrix A is fully indecomposable. Let $B = PAQ$. Then

$$B[J, M] = P[J, \cdot]AQ[\cdot, M] = 0.$$

Since $J, M \neq \emptyset$ and $|J| + |M| = n$, it follows that B is not fully indecomposable. Hence (4.2)(ii) holds.

Now suppose that (4.2) is satisfied. Let $A = [a_{ij}]$ be an $n \times n$ fully indecomposable matrix and let $B = PAQ = [b_{ij}]$. Consider nonempty subsets J and M of $\{1, \dots, n\}$ with $|J| + |M| = n$. We show that $B[J, M] \neq 0$, from which it follows that B is fully indecomposable. Since by (4.2)(i) P and Q^t have no zero rows, it follows from (4.2)(ii) that there exist nonempty subsets K and L of $\{1, \dots, n\}$ with $|K| + |L| = n$ such that $P[J, K]$ has at least one 1 in each column and $Q[L, M]$ has at least one 1 in each row. Since A is fully indecomposable, we conclude that $A[K, L] \neq 0$. Let $k \in K$ and $l \in L$ be determined so that $a_{kl} = 1$. Let $j \in J$ be determined so that $p_{jk} = 1$, and let $m \in M$ be determined so that $q_{lm} = 1$. Then it follows that

$$b_{jm} \geq p_{jk}a_{kl}q_{lm} = 1.$$

Hence $B[J, M] \neq 0$. Thus (4.1) holds and the theorem is proved.

If P and Q are permutation matrices then (4.2) clearly holds. But there need not even be permutation matrices P', Q' with $P \geq P'$ or $Q \geq Q'$ in order that (4.2) hold. An example is given by

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

For a matrix X , let $X(i, j)$ denote the submatrix of X obtained by deleting row i and column j . We then have the following.

COROLLARY 4.2. *Let P be an $n \times n$ matrix of 0's and 1's and let Q be a permutation matrix. Then the following are equivalent:*

- (4.3) *PAQ is fully indecomposable for each $n \times n$ fully indecomposable matrix A of 0's and 1's.*

(4.4) *Either (i) there exists a permutation matrix P' with $P \geq P'$ or (ii) for some integer r with $1 \leq r \leq n$ and all integers j with $1 \leq j \leq n$, there exists an $(n-1) \times (n-1)$ permutation matrix P_j such that $P(j,r) \geq P_j$.*

Proof. We apply Theorem 4.1 with Q a permutation matrix. For $j = 1, \dots, n-1$, Q contains an $(n-j) \times j$ zero submatrix but no $(n-j) \times (j+1)$ zero submatrix. It follows that (4.2) is equivalent to the statement that for $j = 1, \dots, n-1$, P has no $(n-j) \times (j+1)$ zero submatrix. Now the well known Frobenius-König theorem [10, p. 189] asserts that there exists a permutation matrix P' with $P \geq P'$ if and only if P has no $(n-j) \times (j+1)$ zero submatrix for $j = 0, 1, \dots, n-1$. If P has no zero column, it now follows that (4.2) is equivalent to the existence of a permutation matrix P' with $P \geq P'$. Now suppose column r of P contains only 0's. It then follows from the Frobenius-König theorem, that (4.2) is equivalent to there existing an $(n-1) \times (n-1)$ permutation matrix P_j with $P_{(j,r)} \geq P_j$ for $j = 1, \dots, n$. Hence the corollary follows.

We now consider the property of near decomposability of $n \times n$ matrices. For $n = 2$, the only fully indecomposable matrix of 0's and 1's is the matrix $J_{2,2}$ of all 1's and it is nearly decomposable. Let P and Q be 2×2 matrices of 0's and 1's. Then it is readily checked that $PJ_{2,2}Q$ is nearly decomposable (has no 0's) if and only if P has no zero rows and Q has no zero columns. For $n > 2$ we have the following.

THEOREM 4.3. *Let $n > 2$, and let $P = [p_{ij}]$ and $Q = [q_{ij}]$ be $n \times n$ matrices of 0's and 1's. Then PAQ is nearly decomposable for each $n \times n$ nearly decomposable matrix A of 0's and 1's if and only if P and Q are permutation matrices.*

Proof. Suppose that PAQ is nearly decomposable for each nearly decomposable matrix A of 0's and 1's. Then P has no zero rows and Q has no zero columns. From this it follows that if there is no permutation matrix P' with $P \geq P'$, then P has a column with at least two 1's; an analogous conclusion holds for Q . We distinguish several cases.

Case 1. There exist permutation matrices P', Q' such that $P \geq P', Q \geq Q'$. Let A be nearly decomposable. Then A has at least two 1's in each row and each column. Suppose $P \neq P'$. Then there exist i, j, r with $i \neq j$ such that $p_{ir} = p_{jr} = 1$. Since row r of A contains two 1's, it follows that within rows i and j of PAQ there is a 2×2 submatrix with no 0's. By Lemma 2.1, PAQ is not nearly decomposable, a contradiction. Thus $P = P'$ and, similarly, $Q = Q'$, so that P and Q are permutation matrices.

Case 2. There exists a permutation matrix P' with $P \geq P'$ or a permutation matrix Q' with $Q \geq Q'$. Suppose there is a permutation matrix Q' with $Q \geq Q'$. If there is no permutation matrix P' with $P \geq P'$, then some column of P has at least two 1's. Arguing as above we conclude that PAQ is not nearly decomposable. Hence $P \geq P'$ for some permutation matrix P' , and Case 1 applies.

Case 3. There is no permutation matrix P' with $P \geq P'$ and no permutation matrix Q' with $Q \geq Q'$. In this case there exist i, j, r with $i \neq j$ and $p_{ir} = p_{jr} = 1$, and t, u, v with $u \neq v$ and $q_{tu} = q_{tv} = 1$. Choose a nearly decomposable matrix $A = [a_{kl}]$ with $a_{rt} = 1$. Then column t of PA contains nonzero entries in positions (i, t) and (j, t) . It then follows that the 2×2 submatrix of PAQ determined by rows i and j and columns u and v contains no 0's. By Lemma 2.1, PAQ is not nearly decomposable. Thus this case cannot occur.

It follows that P and Q are permutation matrices. The converse clearly holds.

For completeness we include the following two theorems which follow immediately from Lemma 3.6.

THEOREM 4.4. *Let P and Q be $n \times n$ matrices of 0's and 1's. Then PAQ is decomposable for each $n \times n$ decomposable matrix A of 0's and 1's if and only if one*

of the following holds:

(4.5) P has a zero row or Q has a zero column.

(4.6) P and Q are permutation matrices.

THEOREM 4.5. *Let P and Q be $n \times n$ matrices of 0's and 1's. Then the following are equivalent:*

(4.7) *For each $n \times n$ matrix A of 0's and 1's, PAQ is fully indecomposable if and only if A is.*

(4.8) P and Q are permutation matrices.

To conclude, we discuss some interesting problems which arise out of the criterion (4.2) in Theorem 4.1. First we reformulate that criterion. Let P be an $n \times n$ matrix of 0's and 1's. For $k = 1, 2, \dots, n$, define $c_k(P)$ to be the largest j such that P has a $k \times j$ zero submatrix. It follows that

$$n \geq c_1(P) \geq \dots \geq c_n(P) \geq 0$$

with $c_1(P) = n$ if and only if P has a zero row, and $c_n(P) = 0$ if and only if P has no zero column. We put $c(P) = (c_1(P), \dots, c_n(P))$. We define another n -tuple $d(P) = (d_1(P), \dots, d_n(P))$ by $d(P) = c(P')$. Before proceeding we record the following observation. Recall that if $r = (r_1, r_2, \dots, r_n)$ is a monotone decreasing sequence of nonnegative integers, the *conjugate sequence* $r^* = (r_1^*, r_2^*, \dots, r_n^*)$ is defined by

$$r_j^* = \max \{k : r_k \geq j, k = 1, \dots, n\}.$$

Note that $\sum_{j=1}^n r_j = \sum_{j=1}^n r_j^*$.

LEMMA 4.6. *If P is an $n \times n$ matrix of 0's and 1's, then $d(P) = c(P)^*$.*

Proof. It follows from definition that $c_k(P) \geq j$ if and only if $d_j(P) \geq k$. By taking $k = c_j(P)^*$, we conclude that $d_j(P) \geq c_j(P)^*$ for $j = 1, \dots, n$. Similarly, $c_j(P) \geq d_j(P)^*$ for $j = 1, \dots, n$. Since

$$\sum_{j=1}^n c_j(P) = \sum_{j=1}^n c_j(P)^*, \quad \sum_{j=1}^n d_j(P) = \sum_{j=1}^n d_j(P)^*,$$

it follows that $d(P) = c(P)^*$.

Let P and Q be $n \times n$ matrices of 0's and 1's, then condition (4.2) of Theorem 4.1 is equivalent to

(4.9)
$$c_1(P) < n, d_1(Q) < n \text{ and for } k = 1, \dots, n - 1, \\ c_k(P) + d_{n-k}(Q) \leq n.$$

Note that $c_n(P)$ and $d_n(P)$ do not enter into this condition. We say (P, Q) is a *minimal pair for full indecomposability* provided P, Q satisfy (4.9), but the replacement of a 1 by a 0 (in either P or Q) always results in matrices P' and Q' not satisfying (4.9). An example of a minimal pair for full indecomposability when $n = 4$ is

(4.10)
$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Here $c_1(P) = 3, c_2(P) = 3, c_3(P) = 2$ and $d_1(Q) = 2, d_2(Q) = 1, d_3(Q) = 1$ so that $c_k(P) + d_{4-k}(Q) = 4$ for $k = 1, 2, 3$.

It seems an interesting but difficult question to characterize the minimal pairs for full indecomposability. The following property does hold.

LEMMA 4.7. *If (P, Q) is a minimal pair for full indecomposability, then there exists k with $1 \leq k \leq n - 1$ such that $c_k(P) + d_{n-k}(Q) = n$.*

Proof. Assume to the contrary that $c_k(P) + d_{n-k}(Q) < n$ for $k = 1, \dots, n - 1$. First suppose that there exists a row of P containing at least two 1's. Let P' be the matrix obtained from P by replacing one of these 1's by 0. Then $c_1(P') < n$ and $c_k(P') + d_{n-k}(Q) \leq n$ for $k = 1, \dots, n - 1$. Hence (P, Q) is not a minimal pair for full indecomposability. Now suppose each row of P contains exactly one 1. Then not every column of P can contain two or more 1's so that $c_{n-1}(P) \geq 1$. Hence, $d_1(Q) < n - 1$, so that some column of Q contains at least two 1's. An argument similar to the first leads to a contradiction again, and the lemma follows.

The converse of the above lemma is not true, as is seen by taking Q as in (4.10) and

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

Nor is the stronger statement that $c_k(P) + d_{n-k}(Q) = n$ for $k = 1, \dots, n - 1$ whenever (P, Q) is a minimal pair for full indecomposability. A counterexample is provided by the following minimal pair

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

where $c_1(P) = 3, c_2(P) = 3, c_3(P) = 2$ and $d_1(Q) = 2, d_2(Q) = 1, d_3(Q) = 0$. Neither is it true that (P, Q) is a minimal pair for full indecomposability when $c_1(P) < n, d_1(Q) < n$ and $c_k(P) + d_{n-k}(Q) = n$ for $k = 1, \dots, n - 1$. A counterexample is furnished by

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Consideration of the n -tuple $c(P)$ for an $n \times n$ matrix P of 0's and 1's leads to some questions independent of the condition (4.9) and its relation to full indecomposability. Let $c = (c_1, \dots, c_n)$ be an integer n -tuple with $n \geq c_1 \geq \dots \geq c_n \geq 0$, and let $\mathcal{P}(c)$ denote the collection of all $n \times n$ matrices P of 0's and 1's with $c(P) = c$. By taking \bar{P} to be the matrix whose k th row consists of c_k 0's followed by $(n - c_k)$ 1's ($k = 1, \dots, n$), we see that $\bar{P} \in \mathcal{P}(c)$ so that $\mathcal{P}(c) \neq \emptyset$. (If $d = (d_1, \dots, d_n)$ is an integer n -tuple with $n \geq d_1 \geq \dots \geq d_n \geq 0$ and we define $\mathcal{P}(c, d)$ to consist of all $n \times n$ matrices P of 0's and 1's with $c(P) = c$ and $d(P) = d$, then it follows from Lemma 4.6 that $\mathcal{P}(c, d) \neq \emptyset$ if and only if $d = c^*$.)

The following observation is easy to prove. Let $\sigma(P)$ denote the number of 1's in the $n \times n$ matrix P of 0's and 1's.

LEMMA 4.7. *Let $P \in \mathcal{P}(c)$. Then $\sigma(P) \leq \sigma(\bar{P}) = \sum_{i=1}^n (n - c_i)$.*

The following problems appear difficult.

Problem 4.8. Determine the minimum value σ_c of $\sigma(P)$ for $P \in \mathcal{P}(c)$ (or for P satisfying $c_j(P) \leq c_j$ for $j = 1, \dots, n$).

An easy lower bound is: $\sigma_c \geq \max \{n(n - c_1), (n(n - c_1^*))\}$.

Problem 4.9. Characterize those $P \in \mathcal{P}(c)$ for which $\sigma(P) = \sigma_c$.

Call P a *minimal matrix* of $\mathcal{P}(c)$ if $P \in \mathcal{P}(c)$, but for each matrix P' obtained from P by replacing a 1 by a 0, $P' \notin \mathcal{P}(c)$.

Problem 4.10. Characterize the minimal matrices of $\mathcal{P}(c)$.

Let $c = (2, 1, 0, 0)$. Then the matrix

$$P = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

is in $\mathcal{P}(c)$ and $\sigma(P) = 8$. Suppose there were a matrix $P' \in \mathcal{P}(c)$ with $\sigma(P') < 8$. Then some row of P' has at least 3 0's so that $c_1(P') \geq 3$. It follows that $\sigma_c = 8$. Now consider the matrix in $\mathcal{P}(c)$,

$$P_1 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Then it is easy to check that P_1 is a minimal matrix of $\mathcal{P}(c)$, but $\sigma(P_1) = 9 > \sigma_c$. It follows that Problems 4.9 and 4.10 are, in general, different.

Problem 4.8 is related to the well-known problem of Zarankiewicz [12]. For $n \times n$ matrices this problem asks for the minimum positive integer $z(n, j, k)$ such that every $n \times n$ matrix of 0's and 1's containing at least $z(n, j, k)$ 1's has a $j \times k$ submatrix of all 1's. If we interchange 0's with 1's, the problem can be reformulated as: Determine the maximum number $\sigma(n, j, k)$ such that every $n \times n$ matrix P of 0's and 1's with $\sigma(P) \leq \sigma(n, j, k)$ satisfies $c_j(P) \geq k$. A generalization of this problem is then the following:

Problem 4.11 (Generalization of Zarankiewicz's problem). Let $c = (c_1, \dots, c_n)$ be an integer n -tuple with $n \geq c_1 \geq \dots \geq c_n \geq 0$. Determine the maximum value τ_c of $\sigma(P)$ for $P \in \mathcal{P}(c)$ (or for P satisfying $c_j(P) \geq c_j$ for $j = 1, \dots, n$).

REFERENCES

- [1] R. A. BRUALDI AND M. B. HEDRICK, *A unified treatment of nearly reducible and nearly decomposable matrices*, Linear Algebra and Appl., 24 (1979), pp. 51-73.
- [2] R. A. BRUALDI, S. V. PARTER AND H. SCHNEIDER, *The diagonal equivalence of a matrix to a stochastic matrix*, J. Math. Anal. Appl., 16 (1966), pp. 31-50.
- [3] G. FROBENIUS, *Über die Darstellung der endlichen Gruppen durch lineare Substitutionen*, S.-B. Preuss. Akad. Wiss. Berlin, 1897, pp. 994-1015.
- [4] D. J. HARTFIEL, *On constructing nearly decomposable matrices*, Proc. Amer. Math. Soc., 27 (1971), pp. 222-228.
- [5] L. LOVÁSZ AND M. D. PLUMMER, *On minimal elementary bipartite graphs*, J. Combin. Theory Ser. B, 23 (1977), pp. 127-138.
- [6] M. MARCUS, *Linear operations on matrices*, Amer. Math. Monthly, 69 (1962), 837-847.
- [7] ———, *Linear transformations on matrices*, J. Res. Nat. Bur. Standards, 75B (1971), pp. 107-113.
- [8] M. MARCUS AND B. N. MOYLS, *Linear transformations on algebras of matrices*, Canad. J. Math., 11 (1959), pp. 61-66.

- [9] H. MINC, *Linear transformations on nonnegative matrices*, Linear Algebra and Appl., 9 (1974), pp. 149–153.
- [10] L. MIRSKY, *Transversal Theory*, Academic Press, New York, 1971.
- [11] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
- [12] K. ZARANKIEWICZ, *Problem P 101*, Colloq. Math., 2 (1951), p. 301.

MINIMIZING SETUPS FOR ORDERED SETS: A LINEAR ALGEBRAIC APPROACH*

GERHARD GIERZ† AND WERNER POGUNTKE‡

Abstract. The purpose of this paper is to give a lower bound for the setup number of an arbitrary finite ordered set P in terms of the rank of the incidence matrix of P . This bound turns out to equal the setup number for a rather wide family of ordered sets. For the subfamily of cycle-series-parallel ordered sets, a recognition algorithm is presented which produces a (setup) optimal linear extension for every ordered set which is recognized to be cycle-series-parallel.

1. Introduction and basic definitions. Scheduling problems and their complexity have received attention by many authors in the past few years. We refer the reader to J. K. Lenstra–A. H. G. Rinnooy Kan [7] for a survey.

This paper deals with the following very special scheduling problem: A single machine is to perform a set of jobs, one at a time. Certain precedence constraints imply that some jobs cannot be started unless certain other jobs have been completed. However, any time a job is performed immediately after a job which is not constrained to precede it, there has to be a “setup” which causes some fixed additional cost. The problem is to find a schedule which minimizes the number of setups.

The above situation is reflected by the notions of an *ordered set* (precedence constraints) and a *linear extension of an ordered set* (schedule) (cf. [3], [5]).

By an *ordered set*, we mean a set P with a binary relation \preceq which is reflexive, antisymmetric, and transitive. (Throughout this paper, we will only deal with finite ordered sets.) An order-preserving bijection $f: P \rightarrow \{1, 2, \dots, |P|\}$ is called a *linear extension of P* . We say that f has a *setup at $f^{-1}(i)$* ($1 \leq i < |P|$) if $f^{-1}(i) \not\preceq f^{-1}(i+1)$.

Finally, we define

$$s(f) := |\{i: f \text{ has a setup at } f^{-1}(i)\}|,$$

and the *setup number of P* ,

$$s(P) = \min \{s(f): f \text{ is a linear extension of } P\}.$$

As is customary, we will represent ordered sets by certain diagrams. To illustrate the above definitions, let us look at the ordered set N described in Fig. 1(a). Figure 1(b) and (c) describe two different linear extensions f_1 and f_2 of N (with the obvious interpretation: $f_1(a) = 1, f_1(b) = 2$, etc.). Since $s(f_1) = 2$ and $s(f_2) = 1$, and since at least one setup is needed for any linear extension of N , $S(N) = 1$, i.e., f_2 is “optimal.”

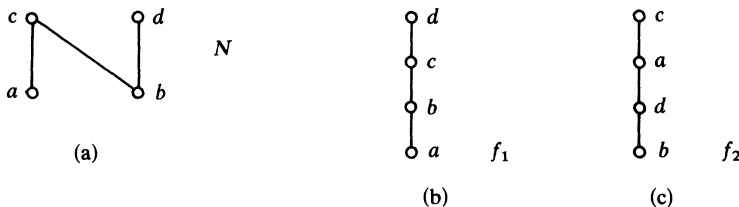


FIG. 1

* Received by the editors January 11, 1982, and in revised form March 15, 1982.

† University of California, Riverside, California 92521.

‡ Fachbereich Mathematik, Technische Hochschule Darmstadt, 6100 Darmstadt, Schlossgartenstrasse 7, West Germany.

The problem of determining the setup number $s(P)$ and producing an optimal linear extension for any given ordered set P has been considered by a number of authors. In most of the papers (cf. [1], [2], [4], [8], [9]), the problem has been looked at within the framework of graph theory. While good algorithms have been found for certain restricted classes of ordered sets, it has been shown by W. R. Pulleyblank [8] that even if attention is restricted to ordered sets not containing a three-element totally ordered set, then finding the setup number still is an NP-complete problem. (For the terminology regarding computational complexity, the reader is referred to M. R. Garey–D. S. Johnson [6].)

2. Further notation and an easy lemma. As usual, for elements a, b of an ordered set P , we write $a < b$ if $a \leq b$ and $a \neq b$. If $a \in P$, then $\downarrow a := \{p \in P \mid p \leq a\}$ is called the *order ideal generated by a* ; if $A \subseteq P$, the *order ideal generated by A* is $\downarrow A := \bigcup_{a \in A} \downarrow a$; a subset B of P is said to be an *order ideal* if $\downarrow B = B$.

A subset C of P which is totally ordered (i.e., $x \leq y$ or $y \leq x$ holds for any $x, y \in C$) is called a *chain*. An *antichain* A is a totally unordered set: $x \not\leq y$ and $y \not\leq x$ for all $x, y \in A$; the *width of P* , $w(P)$, is the maximum size of a subset of P which is an antichain.

The following is easy to show:

LEMMA 2.1. *Let P be an ordered set, $A \subseteq P$ with $\downarrow A = A$, and $E := P \setminus A$. If f is a linear extension of A and g is a linear extension of E , then*

$$f \square g : P \rightarrow \{1, 2, \dots, |P|\}$$

defined by

$$f \square g(x) := \begin{cases} f(x) & \text{if } x \in A, \\ |A| + g(x) & \text{if } x \in E \end{cases}$$

is a linear extension of P , and

$$s(f) + s(g) \leq s(f \square g) \leq s(f) + s(g) + 1.$$

Furthermore, $s(f) + s(g) = s(f \square g)$ holds precisely in the case where $f^{-1}(|A|) \leq g^{-1}(1)$.

3. Independent sequences and the setup number. In this section, we give a characterization of the setup number in terms of the sequence of elements below which setups do not occur. We assume that an arbitrary finite ordered set P is given throughout the section.

DEFINITION 3.1. A sequence $a_1, \dots, a_n \in P$ of elements of P is called an *independent sequence* if the following are satisfied:

- (i) $\downarrow a_1 \neq \{a_1\}$;
- (ii) for each $k \in \{1, \dots, n-1\}$, $\downarrow a_{k+1} \setminus \{a_{k+1}\} \not\subseteq \bigcup_{i \leq k} \downarrow a_i \setminus \{a_k\}$.

LEMMA 3.2. a) *If $a_1, \dots, a_n \in P$ is an independent sequence, then there is a linear extension f of P with $s(f) \leq |P| - n - 1$.*

b) *If f is a linear extension of P with $s(f) = m$, then there is an independent sequence $a_1, \dots, a_{|P|-m-1} \in P$.*

Proof. a) By induction, we define linear extensions f_k of $\bigcup_{i \leq k} \downarrow a_i$ with the following properties:

$$(i) \quad f_k(a_k) = \left| \bigcup_{i \leq k} \downarrow a_i \right|,$$

$$(ii) \quad f_{k+1} \Big|_{\bigcup_{i \leq k} \downarrow a_i} = f_k,$$

$$(iii) \quad s(f_k) \leq \left| \bigcup_{i \leq k} \downarrow a_i \right| - k - 1.$$

Let f_1 be a linear extension of $\downarrow a_1$. Since a_1 is the largest element of $\downarrow a_1$, $f(a_1) = |\downarrow a_1|$ follows. Furthermore, f cannot have a setup at $f^{-1}(|\downarrow a_1| - 1)$, i.e., $s(f_1) \leq |\downarrow a_1| - 2$.

Assume f_k has already been defined. From 2.1 (ii), we know that $\downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i \neq \emptyset$. Let g_k be a linear extension of $\downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i$ and $f_{k+1} := f_k \square g_k \cdot f_{k+1}$, obviously, has the properties (i) and (ii). To verify (iii), we have to consider two cases:

Case 1. $|\downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i| = 1$. One gets from Definition 3.1 (ii) that $a_k \leq a_{k+1}$. Then $f_k(a_k) = |\bigcup_{i \leq k} \downarrow a_i|$ and Lemma 2.1 now imply

$$\begin{aligned} s(f_{k+1}) &= s(f_k) + s(g_k) = s(f_k) + 0 \leq \left| \bigcup_{i \leq k} \downarrow a_i \right| - k - 1 \\ &= \left| \bigcup_{i \leq k} \downarrow a_i \right| + \left| \downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i \right| - k - 2 \\ &= \left| \bigcup_{i \leq k+1} \downarrow a_i \right| - (k+1) - 1. \end{aligned}$$

Case 2. $|\downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i| \geq 2$. In this case, a_{k+1} is the largest element in $\downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i$, hence $s(g_k) \leq |\downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i| - 2$. Thus, one gets

$$\begin{aligned} s(f_{k+1}) &\leq s(f_k) + (g_k) + 1 \leq \left| \bigcup_{i \leq k} \downarrow a_i \right| - k - 1 + \left| \downarrow a_{k+1} \setminus \bigcup_{i \leq k} \downarrow a_i \right| - 2 + 1 \\ &= \left| \bigcup_{i \leq k+1} \downarrow a_i \right| - (k+1) - 1. \end{aligned}$$

This shows that the linear extensions f_k have the desired properties. If $P = \bigcup_{i \leq n} \downarrow a_i$, letting $f = f_n$, we are done. Otherwise, let g be any linear extension of $P \setminus \bigcup_{i \leq n} \downarrow a_i$. Obviously, $s(g) \leq |P \setminus \bigcup_{i \leq n} \downarrow a_i| - 1 = |P| - |\bigcup_{i \leq n} \downarrow a_i| - 1$.

With $f = f_n \square g$, one has

$$s(f) \leq s(f_n) + s(g) + 1 \leq \left| \bigcup_{i \leq n} \downarrow a_i \right| - n - 1 + |P| - \left| \bigcup_{i \leq n} \downarrow a_i \right| - 1 + 1 = |P| - n - 1.$$

b) Let $f: P \rightarrow \{1, 2, \dots, |P|\}$ be a linear extension. We set $R = \{i : i < |P| \text{ and } f^{-1}(i) \leq f^{-1}(i+1)\}$, i.e., R is the set of places where f has no setup. Obviously, $|P| = s(f) + |R| + 1$. We now label the elements of R in such a way that $R = \{i_k : 1 \leq k \leq |P| - s(f) - 2\}$ and $i_k \leq i_{k+1}$ for each $1 \leq k \leq |P| - s(f) - 2$.

Defining $a_k := f^{-1}(i_k + 1)$ for $1 \leq k \leq |P| - s(f) - 1$, our claim is that the elements a_k form an independent sequence. Since $f^{-1}(i_1) \leq f^{-1}(i_1 + 1) = a_1$, condition (i) of Definition 3.1 is satisfied. To show that (ii) holds as well, we set

$$b_{k+1} := f^{-1}(i_{k+1}) \quad \text{for } 1 \leq k \leq |P| - s(f) - 2.$$

We first observe that $b_{k+1} \in \downarrow a_{k+1} \setminus \{a_{k+1}\}$. Let us assume $b_{k+1} \in \bigcup_{i \leq k} \downarrow a_i$, and choose an $l \leq k$ with $b_{k+1} \in \downarrow a_l$.

Since f is isotone, it follows that

$$i_{k+1} = f(b_{k+1}) \leq f(a_l) = i_l + 1.$$

The assumption $l < k$ now implies $i_{k+1} \leq i_l + 1 < i_k + 1 \leq i_{k+1}$, hence $l = k$. But now, $i_k < i_{k+1}$ implies $i_{k+1} = i_k + 1$ which means $b_{k+1} = a_k$. This shows that $\downarrow a_{k+1} \setminus \{a_{k+1}\} \not\subseteq \bigcup_{i \leq k} \downarrow a_i \setminus \{a_k\}$. \square

COROLLARY 3.3. *There is an independent sequence $a_1, \dots, a_n \in P$ if and only if $s(P) \leq |P| - n - 1$.*

4. The defect and a lower bound for the setup number.

DEFINITION 4.1. Let P be an ordered set, and let K be a field. By $\varphi_{P,K}$ we denote the linear mapping from K^P to itself given by

$$\varphi_{P,K}(f)(a) := \sum_{b>a} f(b).$$

The rank and defect of P over K are defined by

$$\begin{aligned} \text{rk}_K(P) &:= \dim(\text{im } \varphi_{P,K}), \\ \text{def}_K(P) &:= \dim(\text{ker } \varphi_{P,K}). \end{aligned}$$

We list some immediate consequences of Definition 4.1:

Fact 4.2. Let $P = \{p_1, \dots, p_m\}$. Obviously, the mappings $\delta_i : P \rightarrow K$ defined by

$$\delta_i(p_k) := \begin{cases} 1 & \text{if } k = i, \\ 0 & \text{if } k \neq i \end{cases}$$

form a basis of K^P . With respect to this basis $\{\delta_1, \dots, \delta_m\}$, $\varphi_{P,K}$ is described by the matrix $A_{P,K} = (a_{i,k})_{i,k}$ with

$$a_{i,k} = \begin{cases} 1 & \text{if } p_i < p_k, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $\text{rk}_K(P) = \text{rk}(A_{P,K})$ and $\text{def}_K(P) = \text{def}(A_{P,K})$.

Fact 4.3. By induction, it is easy to show that

$$\varphi_{P,K}^l(f)(a) = \sum_{b_1>a} \sum_{b_2>b_1} \dots \sum_{b_l>b_{l-1}} f(b_l)$$

for every integer l . This means that if $c^l(p, q)$ denotes the number of $(l + 1)$ -element chains $p = p_1 < p_2 < \dots < p_l < p_{l+1} = q$ in P , then

$$A_{P,K}^l = (c^l(p_i, p_k) \cdot 1)_{i,k}.$$

For each $a \in P$, we define $\chi_a \in K^P$ by

$$\chi_a(p) := \begin{cases} 1 & \text{if } a > p, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $\chi_{p_i} = \varphi_{P,K}(\delta_i)$.

The following proposition is of central importance in this paper. It links independent sequences (cf. the preceding section) to linearly independent elements of K^P :

PROPOSITION 4.4. *Let P be an ordered set, and let K be a field. If the elements $a_1, \dots, a_n \in P$ form an independent sequence, then $\chi_{a_1}, \dots, \chi_{a_n}$ are linearly independent in K^P ; in particular, $\text{rk}_K(P) \geq n$.*

Proof. Let $\sum_{i=1}^n r_i \chi_{a_i} = 0$, $r_i \in K$. Since b_1, \dots, b_{m-1} is an independent sequence whenever b_1, \dots, b_m is, it is enough to show that $r_n = 0$. By condition (ii) in Definition 3.1, there is a $b \in \downarrow a_n \setminus \{a_n\}$ which is not an element of $\cup_{i \leq n-1} \downarrow a_i \setminus \{a_{n-1}\}$. It follows that $\chi_{a_n}(b) = 1$, and $\chi_{a_i}(b) = 0$ for any $i < n$. Consequently,

$$0 = \left(\sum_{i=1}^n r_i \chi_{a_i} \right)(b) = r_n \cdot 1 = r_n. \quad \square$$

Corollary 3.3 and Proposition 4.4 now together imply:

THEOREM 4.5. *If P is an ordered set and K is a field, then $\text{def}_K(P) - 1 \leq s(P)$.*

It turns out that even if one chooses $K = F_2$ to be the two-element field, the lower bound for $s(P)$ given in Theorem 4.5 is sharp for a suprisingly rich class of ordered sets. The remainder of this section is devoted to studying properties of this class. (Fig. 2 shows a small ordered set Q for which $\text{def}_{F_2}(Q) - 1 < s(P)$.)

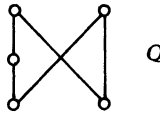
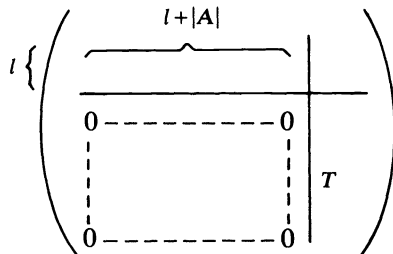


FIG. 2

It is very easy to see that for any ordered set P , $w(P) - 1 \leq s(P)$. The following lemma shows that Theorem 4.5 improves this inequality:

LEMMA 4.6. *If P is an ordered set and K is a field, then $w(P) \leq \text{def}_K(P)$.*

Proof. Let A be an antichain in P and $|P| = m$. We may assume that the elements of $P = \{p_1, p_2, \dots, p_m\}$ are labelled in such a way that $\{p_1, p_2, \dots, p_l\} = P \uparrow A$, $\{p_{l+1}, \dots, p_{l+|A|}\} = A$, and $\{p_{l+|A|+1}, \dots, p_m\} = \uparrow A \setminus A$. Let $\delta_1, \dots, \delta_m$ be the basis of K^P as defined in Fact 4.2. With respect to this basis, $\varphi_{P,K}$ is described by the matrix $A_{P,K}$, having a shape like this:



Hence, it follows that $\text{rk}_K(P) \leq l + r_K(T) \leq l + (|P| - (|A| + l)) = |P| - |A|$ and $\text{def}_K(P) \geq |A|$. □

We continue with a list of ordered sets for which $\text{def}_K(P) - 1 = s(P)$ (with a suitable K) holds.

Example 4.7. If P is a chain with n elements, then $\text{def}_K(P) = 1$, hence $\text{def}_K(P) - 1 = 0 = s(P)$. If P is an antichain with n elements, then $\text{def}_K(P) = n$, i.e., $\text{def}_K(P) - 1 = s(P)$.

Example 4.8. An ordered set $\{x_1, y_1, x_2, y_2, \dots, x_n, y_n\}$ of size $2n$ ($n \geq 3$) with the comparabilities

$$y_1 < x_1, \quad x_1 > y_2, \quad y_2 < x_2, \quad x_2 > y_3, \quad \dots, \quad x_{n-1} > y_n, \quad y_n < x_n, \quad x_n > y_1$$

(and no others) is called a $2n$ -cycle (see Fig. 3(a)). If P is a $2n$ -cycle, then it is easy to check that $s(P) = n$. Furthermore, $\text{def}_{F_2}(P) = n + 1$ and hence $\text{def}_{F_2}(P) - 1 = s(P)$, although $w(P) - 1 = n - 1 < s(P)$.

Example 4.9. (This contains Example 4.7 as a special case.) It has been shown in D. Duffus-I. Rival-P. Winkler [5] that if P is cycle-free, i.e., if P contains no subset isomorphic to any $2n$ -cycle or to the ordered set described in Fig. 3(b), then $w(P) - 1 = s(P)$. By Lemma 4.6, $\text{def}_K(P) - 1 = s(P)$ for any such P (and any field K).

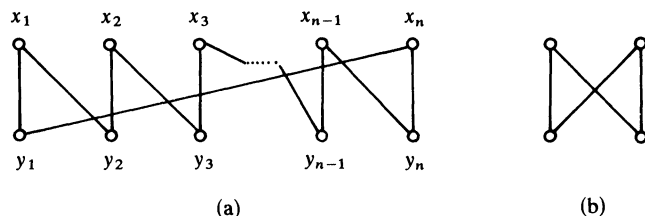


FIG. 3

We next show that the property of an ordered set P that $\text{def}_K(P) - 1$ is the setup number is preserved by a very general construction, the *lexicographic sum*.

DEFINITION 4.10. Let S be an ordered set, and let P_s be an ordered set for each $s \in S$. On the disjoint union $\dot{\bigcup}_{s \in S} P_s$, we define an order structure \sqsubseteq by saying that $p \sqsubseteq q$ holds if and only if either there are $s, s' \in S$ with $s < s'$, $p \in P_s$, and $q \in P_{s'}$, or there is an $s \in S$ with $p, q \in P_s$ and $p \leq q$ in P_s . The ordered set $\mathbf{L}_{s \in S} P_s := (\dot{\bigcup}_{s \in S} P_s, \sqsubseteq)$ is called the *lexicographic sum* of the family $(P_s)_{s \in S}$ over S .

We first need the following:

PROPOSITION 4.11. Let $P = \mathbf{L}_{s \in S} P_s$ be the lexicographic sum of $(P_s)_{s \in S}$ over S , and let K be any field. Then

$$\text{def}_K(P) - 1 \cong \sum_{s \in S} (\text{def}_K(P_s) - 1) + \text{def}_K(S) - 1.$$

Proof. We identify K^P with $\prod_{s \in S} K^{P_s}$. Let $(f_s)_{s \in S} \in \prod_{s \in S} K^{P_s}$ and $\varphi_{P,K}((f_s)_{s \in S}) = (g_s)_{s \in S}$. For any $s \in S$ and $p \in P_s$,

$$g_s(p) = \sum_{q > p} f_s(q) + \sum_{t > s} \left(\sum_{a \in P_t} f_t(a) \right) = \varphi_{P_s, K}(f_s)(p) + \sum_{t > s} \left(\sum_{a \in P_t} f_t(a) \right).$$

We define $\pi : P \rightarrow S$ by $\pi(p) = s$ if $p \in P_s$, π induces a linear mapping $\hat{\pi} : K^S \rightarrow K^P$ with $\hat{\pi}(f) := f \circ \pi$. Furthermore, we have a linear mapping $\delta : K^P \rightarrow K^S$ which maps $(f_s)_{s \in S} = f$ to $\delta(f)$ with $\delta(f)(s) = \sum_{a \in P_s} f_s(a)$. Now, for any $f = (f_s)_{s \in S}$, $s \in S$ and $p \in P_s$, one has

$$\begin{aligned} \hat{\pi} \circ \varphi_{S,K} \circ \delta(f)(p) &= [(\varphi_{S,K} \circ \delta)(f)](\pi(p)) = (\varphi_{S,K}(\delta(f)))(s) \\ &= \sum_{t > s} \delta(f)(t) = \sum_{t > s} \sum_{a \in P_t} f_t(a). \end{aligned}$$

Altogether, one has

$$\varphi_{P,K} = \prod_{s \in S} \varphi_{P_s, K} + \hat{\pi} \varphi_{S,K} \circ \delta,$$

which implies

$$\text{rk}_K(P) = \text{rk}(\varphi_{P,K}) \leq \sum_{s \in S} \text{rk}(\varphi_{P_s, K}) + \text{rk}(\varphi_{S,K}),$$

and the desired inequality follows. \square

Proposition 4.11 now enables us to prove:

THEOREM 4.12. Let K be a field, let S be an ordered set, and let $(P_s)_{s \in S}$ be a family of ordered sets. If $s(S) = \text{def}_K(S) - 1$, and if $s(P_s) = \text{def}_K(P_s) - 1$ for each $s \in S$, then

$$s\left(\mathbf{L}_{s \in S} P_s\right) = \text{def}_K\left(\mathbf{L}_{s \in S} P_s\right) - 1.$$

Proof. Let f be an optimal linear extension of S , and for each $s \in S$, let f_s be an optimal linear extension of P_s . We define

$$g: \prod_{s \in S} P_s \rightarrow \left\{ 1, 2, \dots, \left| \prod_{s \in S} P_s \right| \right\}$$

by

$$g(p) := \sum_{f(t) < f(s)} |P_t| + f_s(p)$$

with s being the unique $s \in S$ such that $p \in P_s$.

Obviously, g is a linear extension, and

$$s(g) = \sum_{s \in S} s(f_s) + s(f).$$

Using Proposition 4.11, one now gets

$$\begin{aligned} s\left(\prod_{s \in S} P_s\right) &\leq s(g) = \sum_{s \in S} s(f_s) + s(f) = \sum_{s \in S} (\text{def}_K(P_s) - 1) + \text{def}_K(S) - 1 \\ &\leq \text{def}_K\left(\prod_{s \in S} P_s\right) - 1 \leq s\left(\prod_{s \in S} P_s\right), \end{aligned}$$

and the proof is finished. \square

If P_1 and P_2 are ordered sets, then the lexicographic sum of (P_1, P_2) over the two-element chain $\{1, 2\}$ is also called the *linear sum* of P_1 and P_2 ; the lexicographic sum over the two-element antichain is also known as the *disjoint sum* of P_1 and P_2 . The smallest class of ordered sets containing the one-element ordered set and being closed under taking linear sums and disjoint sums is the class of all *series-parallel* ordered sets. Theorem 4.12 now immediately gives:

COROLLARY 4.13. *If P is a series-parallel ordered set, then $s(P) = \text{def}_K(P) - 1$ for an arbitrary field K .*

For series-parallel ordered sets, the proof of Theorem 4.12 shows that an “obvious” linear extension is optimal. Several authors have dealt with this class of ordered sets (cf. [4], [8], [9]). In W. R. Pulleyblank [8], a recognition algorithm for this class is presented that finds a decomposition (and hence an optimal linear extension) for every series-parallel ordered set in polynomial time.

We want to consider a slightly larger class in the next section. As a preparation, we formulate one more consequence of Example 4.8 and Theorem 4.12; the class of ordered sets P with $s(P) = \text{def}_{F_2}(P) - 1$ will be denoted by \mathcal{C}_2 .

COROLLARY 4.14. *\mathcal{C}_2 contains all chains and all cycles and is closed under taking linear sums and disjoint sums.*

5. Cycle-series-parallel ordered sets. By \mathcal{CSP} , we denote the class of all (finite) ordered sets P with the following properties:

- (a) If P contains a subset S isomorphic to N (see Fig. 1(a)), then there is precisely one $2n$ -cycle C in P such that $S \subseteq C$.
- (b) If C is a $2n$ -cycle, C' is a $2m$ -cycle, and $C \cap C' \neq \emptyset$, then $C = C'$.
- (c) If $C \subseteq P$ is a cycle and $c \in C$ is minimal in P , then each minimal element of C is minimal in P .

THEOREM 5.1. a) *The class \mathcal{CSP} contains all chains and all cycles and is closed under taking linear and disjoint sums.*

b) *\mathcal{CSP} is the smallest class of ordered sets with the properties described in a).*

Because of Theorem 5.1, we call the elements of \mathcal{CSP} *cycle-series-parallel* ordered sets.

Proof. a) is obvious.

To show b), let $P \in \mathcal{CSP}$. If P is not connected (i.e., P is a disjoint sum of several components), then we can treat the components separately. Hence, we may now assume that P is connected and $|P| > 2$. We have to show that P is either a cycle or a linear sum of two ordered sets Q and R ; by induction, this will conclude the proof. Let M be the set of all minimal elements of P . If $|M| = 1$, then P is the linear sum of M and $P \setminus M$. Thus, we may assume that $|M| \geq 2$ and set $R := \{q \in P : q \geq p \text{ for each } p \in M\}$.

Case 1. $R \neq \emptyset$. We show that P is the linear sum of $Q := P \setminus R$ and R . Let $q \in Q$ and $r \in R$, and assume $q \not\leq r$. Since $q \notin R$, we may choose a $p \in M$ with $p \not\leq q$. Let $s \in M$ with $s \leq q$. One can now easily check that the subset $\{p, q, r, s\}$ is isomorphic to N . Since $P \in \mathcal{CSP}$, there is a $2n$ -cycle C with $\{p, q, r, s\} \subseteq C$, and each minimal element of C is minimal in P . Furthermore, since $|C| > 4$, there is an $m \in M \cap C$ with $m \notin \{p, q, r, s\}$. But now, since $s, p < r$, and r has only two lower neighbours in C , it follows that $m \not\leq r$, contradicting $r \in R$. This shows that $q \leq r$ and P is the linear sum of Q and R .

Case 2. $R = \emptyset$. We show that P is a cycle. The *first step* will be to prove that if $m \in M$ and $p > m$, then there is a cycle C with $\{p, m\} \subseteq C$: Since $R = \emptyset$, there is an $m' \in M$ with $m' \not\leq p$. As P is connected, there is a smallest $k \geq 0$ and elements $p_0, \dots, p_{k+1}, q_0, \dots, q_k \in P$ with $m = p_0 < q_0 > p_1 < \dots < q_k > p_{k+1} = m'$. It is easy to see that $\{p_0, \dots, p_{k+1}\}$ and $\{q_0, \dots, q_k\}$ are antichains. Let $i = \max\{j : p_j \leq p\}$. If $p_i = p_0 = m$, then $\{m, p, q_i, p_{i+1}\} \cong N$; if $p_i \neq p_0 = m$, then $\{m, p, p_i, q_i\} \cong N$; in any case, there has to be a cycle C with $\{p, m\} \subseteq C$.

In the *second step*, we show that if $m, m' \in M$ and $m \neq m'$, there is a cycle C with $\{m, m'\} \subseteq C$: Since P is connected, there is a smallest $k \geq 1$ and elements $p_0, p_1, \dots, p_{2k} \in P$ with $p_0 = m < p_1 > p_2 < \dots < p_{2k-1} > p_{2k} = m'$. It follows that for each even $i \leq 2k - 4$, $p_i \not\leq p_{i+3}$ and $p_{i+1} \not\leq p_{i+4}$.

Let us assume that $k \geq 2$. For each $i \leq 2k - 3$, the subset $\{p_i, p_{i+1}, p_{i+2}, p_{i+3}\}$ is isomorphic to N . Hence, for each such i , there is a cycle C_i with $\{p_i, p_{i+1}, p_{i+2}, p_{i+3}\} \subseteq C_i$. By condition (b) on \mathcal{CSP} , we can conclude from $\{p_{i+1}, p_{i+2}, p_{i+3}\} \subseteq C_i \cap C_{i+1}$ that $C_i = C_{i+1}$ for $0 \leq i \leq 2k - 4$. Thus, we get $p_0 = m \in C_0 = C_1 = \dots = C_{2k-3} \ni p_{2k} = m'$, and m and m' belong to a common cycle.

If $k = 1$, then there is a p_1 with $m < p_1 > m'$. Using the above first step, we get cycles C and C' with $\{m, p_1\} \subseteq C$ and $\{m', p_1\} \subseteq C'$. Again, since $p_1 \in C \cap C'$, $C = C'$ follows. This proves the second step.

Using the second step and condition (b) on \mathcal{CSP} repeatedly, it is now easy to conclude the existence of one cycle C with $M \subseteq C$. But this fact, together with step one and (b), implies $P = C$. \square

Implicit in the proof of Theorem 5.1 is a polynomial algorithm that recognizes the elements of \mathcal{CSP} and constructs an optimal linear extension:

ALGORITHM

Input: A finite ordered set P .

Output: The information if $P \in \mathcal{CSP}$ and, if the answer is positive, an optimal linear extension of P .

Description: We present a procedure which encodes each element of P by a string of integers if $P \in \mathcal{CSP}$ and which, if $P \notin \mathcal{CSP}$, discovers this fact. In the positive case, the lexicographic ordering of the strings obtained will describe an optimal linear extension of P :

Step 1: A string consisting only of "0" is attributed to each element of P .

- Step 2:* Decompose P into the disjoint sum of connected ordered sets, and label these arbitrarily P_1, \dots, P_m . For any $1 \leq i \leq m$ and $x \in P_i$, add an i to the string that was already attributed to x before.
- Step 3:* Set $J = \{i: 1 \leq i \leq m, |P_i| > 1, \text{ and } P_i \text{ is not a cycle}\}$. If $J = \emptyset$, we go to step 4. Otherwise, for each $i \in J$, let M_i be the set of all minimal elements of P_i , $R_i := \{q \in P_i : p < q \text{ for each } p \in M_i\}$, and $Q_i := P_i \setminus R_i$. It has to be checked whether $q < p$ for each $q \in Q_i$ and $p \in R_i$. If this is not the case, then $P \notin \mathcal{CSP}$, and the algorithm STOPS. If it is true, we add a “0” to the string belonging to each element of Q_i and “1” to the string belonging to each element of R_i . Then we return to step 2 with P replaced by Q_i and by R_i .
- Step 4:* It follows that $P \in \mathcal{CSP}$. For each $1 \leq i \leq m$ with $|P_i| > 1$, we label the elements of the $2n$ -cycle P_i optimally by $\{1, 2, \dots, 2n\}$ and add the respective integer to the string of each element.

The algorithm is illustrated in Fig. 4 for a particular cycle-series-parallel ordered set S . The reader should observe that the optimality of the constructed linear extension follows from $\mathcal{CSP} \subseteq \mathcal{C}_2$, which is a consequence of Corollary 4.14 and Theorem 5.1.

6. Some more results—and counterexamples. By Corollary 4.14, the class \mathcal{C}_2 (i.e., all ordered sets P with $s(P) = \text{def}_{F_2}(P) - 1$) is quite big. In particular, it contains the class of all cycle-series-parallel ordered sets which we investigated in the preceding section. At present, we are not able to give a good characterization of \mathcal{C}_2 . It is the purpose of this section to present some more—positive as well as negative—results on \mathcal{C}_2 (and on the corresponding classes \mathcal{C}_p with an arbitrary prime p).

As usual, when we consider the product $P \times Q$ of ordered sets P and Q , the order is taken componentwise. We will mainly deal with products of chains. The following proposition is needed first:

PROPOSITION 6.1. *If C is a chain and P is any ordered set, then $s(P \times C) \leq |C| \cdot s(P) + |C| - 1$.*

Proof. We proceed by induction on the cardinality of C . The case $|C| = 1$ is trivial. Now, let us assume that $|C| = n + 1$, and let u be the greatest element of C . Obviously, the set $P \times C = (P \times (C \setminus \{u\})) \cup (P \times \{u\})$. Let f and g be optimal linear extensions of $P \times (C \setminus \{u\})$ and of $P \times \{u\}$, respectively. (Observe that $P \times (C \setminus \{u\})$ is an order ideal of $P \times C$.) We get

$$\begin{aligned}
 s(P \times C) &\leq s(f \square g) \leq s(f) + s(g) + 1 \\
 &= s(P \times (C \setminus \{u\})) + s(P) + 1 \\
 &\leq n \cdot s(P) + n - 1 + s(P) + 1 \\
 &= (n + 1)s(P) + (n + 1) - 1 \\
 &= |C| \cdot s(P) + |C| - 1. \quad \square
 \end{aligned}$$

If P and Q are ordered sets and K is a field, then $K^{P \times Q}$ is isomorphic to the tensor product over K

$$K^P \otimes K^Q;$$

a canonical isomorphism $K^P \otimes K^Q \rightarrow K^{P \times Q}$ is induced by $f \otimes g \rightarrow (f, g)$ with

$$(f, g)(p, q) = f(p) \cdot g(q).$$

This isomorphism induces an isomorphism between the endomorphism rings as well.

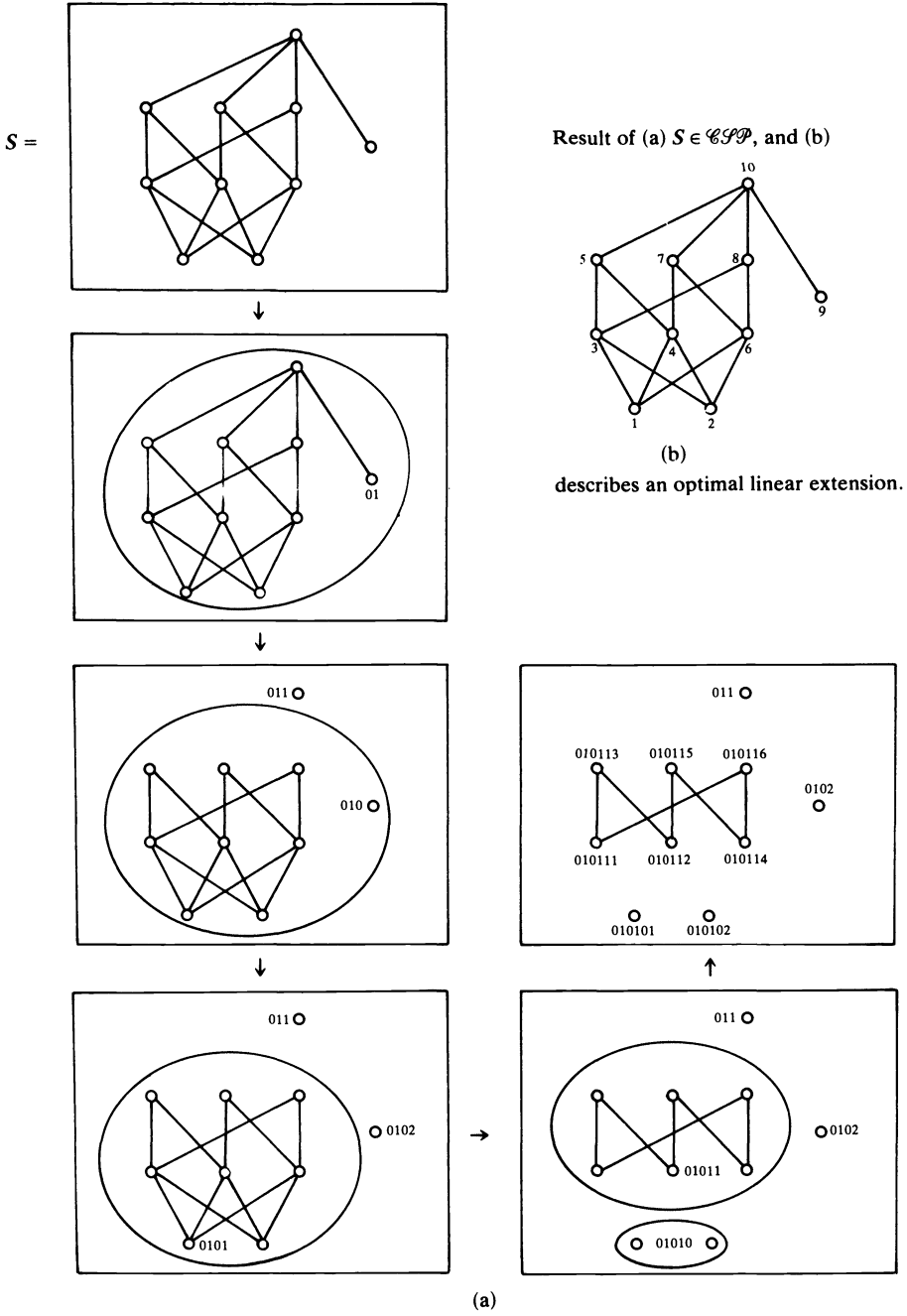


FIG. 4

In what follows, to avoid too-complicated formulas, we will identify $K^P \otimes K^Q$ and $K^{P \times Q}$.

LEMMA 6.2. *If P and Q are ordered sets and K is a field, then*

$$(\varphi_{P,K} + \text{id}_{K^P}) \otimes (\varphi_{Q,K} + \text{id}_{K^Q}) = \varphi_{P \times Q, K} + \text{id}_{K^{P \times Q}}.$$

Proof. For each pair $(a, b) \in P \times Q$ and each $f \otimes g \in K^P \otimes K^Q$, one has

$$\begin{aligned} & \{[(\varphi_{P,K} + \text{id}_{K^P}) \otimes (\varphi_{Q,K} + \text{id}_{K^Q})](f \otimes g)\}(a, b) \\ &= (\varphi_{P,K} + \text{id}_{K^P})(f)(a) \cdot (\varphi_{Q,K} + \text{id}_{K^Q})(g)(b) \\ &= (\varphi_{P,K}(f) + f)(a) \cdot (\varphi_{Q,K}(g) + g)(b) \\ &= \left(\sum_{u \cong a} f(u) \right) \cdot \left(\sum_{v \cong b} g(v) \right) \\ &= \sum_{u \cong a} \sum_{v \cong b} f(u) \cdot g(v) \\ &= \sum_{(u,v) \cong (a,b)} (f \otimes g)(u, v) \\ &= (\varphi_{P \times Q, K}(f \otimes g) + f \otimes g)(a, b) \\ &= \{[\varphi_{P \times Q, K} + \text{id}_{K^{P \times Q}}](f \otimes g)\}(a, b). \quad \square \end{aligned}$$

We continue with some applications of Proposition 6.1 and Lemma 6.2 to products of chains.

PROPOSITION 6.3. *Let C_r be the chain with r elements where r is prime. If the r -element field is denoted by F_r , then*

$$\text{def}_{F_r}(C_r^m) - 1 = s(C_r^m) = r^{m-1} - 1$$

for each integer m .

Proof. The inequality $s(C_r^m) \leq r^{m-1} - 1$ is obtained from Proposition 6.1. Hence, it is enough to show that $r^{m-1} \leq \text{def}_{F_r}(C_r^m)$. (We will drop the subscript F_r in the rest of this proof.) We now determine $\text{def}((\varphi_{C_r^m})^r)$ and make use of Lemma 6.2 and the fact that $\binom{i}{i} \equiv 0 \pmod r$ for each $0 < i < r$.

Since $C_r^m = C_r^{m-1} \times C_r$, we get

$$\begin{aligned} (\varphi_{C_r^m})^r &= [(\varphi_{C_r^{m-1}} + \text{id}_{C_r^{m-1}}) \otimes (\varphi_{C_r} + \text{id}_{C_r}) - \text{id}_{C_r^m}]^r \\ &= \sum_{i=0}^r \binom{r}{i} (-1)^{r-i} (\varphi_{C_r^{m-1}} + \text{id}_{C_r^{m-1}})^i \otimes (\varphi_{C_r} + \text{id}_{C_r})^i \\ &= (-1)^r (\text{id}_{C_r^{m-1}} \otimes \text{id}_{C_r}) + ((\varphi_{C_r^{m-1}} + \text{id}_{C_r^{m-1}})^r \otimes (\varphi_{C_r} + \text{id}_{C_r})^r) \\ &= (-1)^r \text{id}_{C_r^m} + \left(\sum_{i=0}^r \binom{r}{i} \varphi_{C_r^{m-1}}^i \right) \otimes \left(\sum_{i=0}^r \binom{r}{i} \varphi_{C_r}^i \right) \\ &= (-1)^r \text{id}_{C_r^m} + (\text{id}_{C_r^{m-1}} + \varphi_{C_r^{m-1}}^r) \otimes (\text{id}_{C_r} + \varphi_{C_r}^r). \end{aligned}$$

From Fact 4.3, we know that $\varphi_{C_r}^r = 0$; furthermore, $(-1)^r \text{id}_{C_r^m} = -\text{id}_{C_r^m}$. We thus get $(\varphi_{C_r^m})^r = \varphi_{C_r^{m-1}}^r \otimes \text{id}_{C_r}$, which implies by induction that

$$\text{def}(\varphi_{C_r^m}^r) = r \cdot \text{def}(\varphi_{C_r^{m-1}}^r) = r \cdot r^{m-1} = r^m.$$

One now gets $r^{m-1} \leq \text{def}(\varphi_{C_r^m})$ from $\text{def}(\varphi_{C_r^m}^r) \leq r \cdot \text{def}(\varphi_{C_r^m})$. \square

An immediate consequence of Proposition 6.3 is

PROPOSITION 6.4. *If B is a finite Boolean lattice, then $B \in \mathcal{C}_2$, i.e., $s(B) = \text{def}_{F_2}(B) - 1$.*

PROPOSITION 6.5. *If C_r and C_s are chains with r and s elements, respectively, then $\text{def}_K(C_r \times C_s) - 1 = s(C_r \times C_s) = \min\{r, s\} - 1$ for any field K .*

Proof. The inequality $s(C_r \times C_s) \leq \min\{r, s\} - 1$ easily follows from Proposition 6.1. We assume $r \leq s$ and show that $r \leq \text{def}_K(C_r \times C_s)$. To this end, we choose bases of K^{C_r} and K^{C_s} , respectively, such that with respect to these bases, $\varphi_{C_r, K}$ is described by the $r \times r$ -matrix

$$A_r = \begin{pmatrix} 0 & 1 & & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & & & & \vdots \\ 0 & & \cdots & & 1 \\ 0 & & & \cdots & 0 \end{pmatrix}$$

and $\varphi_{C_s, K}$ is described by the $s \times s$ -matrix A_s of the same shape. Lemma 6.2 now implies that, with respect to a suitable basis, $\varphi_{C_r \times C_s, K}$ is described by the $rs \times rs$ -matrix

$$A_{r,s} = \begin{pmatrix} A_s & A_s + E & \cdots & A_s + E \\ & & \ddots & \vdots \\ & 0 & & A_s + E \\ & & & A_s \end{pmatrix}$$

where E is the $s \times s$ identity matrix. Gaussian elimination leads to the matrix

$$\begin{pmatrix} A_s & E & & 0 \\ & & \ddots & \\ & & & E \\ 0 & & & A_s \end{pmatrix}$$

and it is easy to see that its defect is r . \square

Proposition 6.3 and Corollary 6.5 might lead to the conjecture that if P is a product of chains, then $s(P) = \text{def}_K(P) - 1$ holds with an arbitrary field K . The ordered set C_3^3 (where C_3 is the three-element chain) refutes this, though: $s(C_3^3) = 8$ and $\text{def}_{F_2}(C_3^3) = 7$.

We conclude this paper with examples showing that $s(P) - \text{def}_K(P) + 1$ can be arbitrarily large for an ordered set P : For every ordered set Q out of the series described in Fig. 5, $\text{def}_K(Q) = 2$ holds, but the setup numbers of these ordered sets are obviously not bounded.

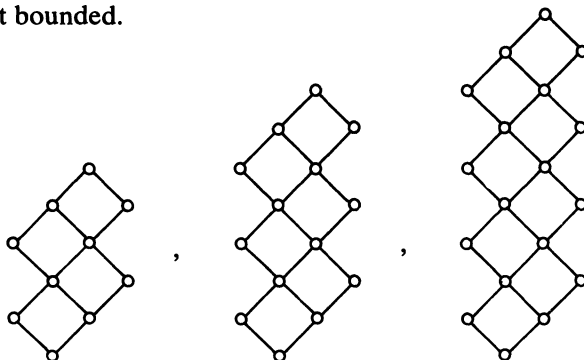


FIG. 5

REFERENCES

- [1] G. CHATY AND M. CHEIN, *Ordered matchings and matchings without alternating cycles in bipartite graphs*, *Utilitas Mathematica*, 16 (1979), pp. 183–187.
- [2] M. CHEIN AND P. MARTIN, *Sur le nombre de sauts d'une forêt*, *C.R. Acad. Sci. Paris*, 275 (1972), pp. 159–161.
- [3] M. CHEIN AND M. HABIB, *The jump number of dags and posets: An introduction*. *Ann. Discrete Math.*, 9 (1980), pp. 189–194.
- [4] O. COGIS AND M. HABIB, *Nombre de sauts et graphes série-parallèles*, *RAIRO Inform. Théor.*, 13 (1979), pp. 3–18.
- [5] D. DUFFUS, I. RIVAL AND P. WINKLER, *Minimizing setups for cycle-free ordered sets*, Emory Univ., Atlanta, GA, 1981 (Preprint).
- [6] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.
- [7] J. K. LENSTRA AND A. H. G. RINNOOY KAN, *Complexity of scheduling under precedence constraints*, *Oper. Res.* 26 (1978), pp. 22–35.
- [8] W. R. PULLEYBLANK, *On minimizing setups in precedence constrained scheduling*, Universität Bonn, 1981 (Preprint).
- [9] J. VALDES, *Parsing flow charts and series-parallel graphs*, Tech. Rep. STAN-CS-78-682, Computer Science Dept., Stanford Univ., Stanford, CA, 1978.

LATTICES OF SIMPLEX TYPE*

A. NEUMAIER†

Abstract. Lattices of simplex type provide a common setting for the Leech lattice, a lattice related to Shult and Yanushka's [Geom. Dedicata, 9 (1980), pp. 1-72] very regular line system with 256 lines and Du Val's [Proc. London Math. Soc., 42 (1937), pp. 18-51] hyperbolic version of the root system E_8 . Many other examples are given.

Introduction. The present paper discusses some observations on the borderline between a few famous topics, reflection groups (Coxeter), extremal lattices (Leech, Sloane), spherical designs (Goethals, Seidel), line systems (Shult, Yanushka) and primitive permutation groups.

After a definition of lattices of simplex type, we present a normal form, and discuss the fact that any lattice of simplex type is a refinement of a "trivial" lattice. Among the trivial lattices we find the root lattices A_p . Then a number of nontrivial examples are given, among them the Leech lattice [3], some extremal lattices and a lattice in \mathbb{R}^{16} whose set of 256 lines through the minimal norm vectors forms one of the nice tetrahedrally closed line systems of Shult and Yanushka [15].

In § 2 we define the hyperbolic transform of a lattice of simplex type. It is a generalization of a construction of Du Val [19] who studied the root system E_8 in a hyperbolic setting (see also Manin [11]). This leads to lattices of hyperbolic simplex type, which are slightly more general than hyperbolic transforms. By forming sections we obtain sequences of lattices, and some examples show that sections give a natural geometric interpretation to some well-known sequences of combinatorial configurations and associated permutation groups.

Section 3 partly explains the sporadic nature of the examples. We study integral lattices of simplex type and divide them into two classes: standard and exceptional lattices. Standard lattices turn out to be related to certain self-orthogonal codes, and they seem to be abundant. On the other hand, if the minimal norm n is given, there are only finitely many exceptional lattices generated by norm n vectors. The cases $n = 2$ and $n = 3$ are discussed in some detail, and relations to star-closed and tetrahedrally-closed line systems become apparent.

1. The Euclidean normal form. We motivate our investigation with a property of the Leech lattice, i.e., the unique even unimodular lattice Λ_{24} in \mathbb{R}^{24} with minimal norm 4 (Conway [3]). In Leech and Sloane [10], there are two constructions for the Leech lattice. The first [10, § 4.4] exhibits a set of 24 vectors in $\sqrt{2} \Lambda_{24}$ of shape $(-1\frac{1}{2}, (\frac{1}{2})^{23})$ —with respect to a suitable basis—of norm 8 and mutual inner product 4, and the second [10, § 5.7] exhibits a set of 24 vectors in $\sqrt{3} \Lambda_{24}$ of shape $(-2\frac{1}{2}, (\frac{1}{2})^{23})$ —with respect to another basis—of norm 12 and mutual inner product 3. After scaling, this leads to vectors $z_1, \dots, z_{24} \in \Lambda_{24}$ with

$$(z_i, z_j) = \begin{cases} 4 & \text{if } i = j, \\ 2 & \text{if } i \neq j \end{cases}$$

and to vectors $z'_1, \dots, z'_{24} \in \Lambda_{24}$ with

$$(z'_i, z'_j) = \begin{cases} 4 & \text{if } i = j, \\ 1 & \text{if } i \neq j. \end{cases}$$

* Received by the editors August 26, 1981, and in revised form May 21, 1982.

† Institut für Angewandte Mathematik, Universität Freiburg, West Germany.

We also note that any two distinct vectors of norm 4 in Λ_{24} have inner product $(x, y) \leq 2$. Thus Λ_{24} contains (in two essentially distinct ways) a regular simplex consisting of 24 minimal norm vectors.

Let us say that a lattice E of dimension $p \geq 2$ is of *strict (Euclidean) simplex type* if there are numbers $n > m > 0$ such that (with the inner product associated with E):

(L1) There are vectors $z_1, \dots, z_p \in E$ such that for $i, j = 1, \dots, p$,

$$(1) \quad (z_i, z_j) = \begin{cases} n & \text{if } i = j, \\ m & \text{if } i \neq j. \end{cases}$$

(L2) The minimal norm is n ; i.e., $(x, x) \geq n$ for all $x \in E \setminus \{0\}$.

(L3) If x, y are distinct norm n vectors of E then $(x, y) \leq m$.

If only (L1) and (L2) hold we say that E is of *simplex type*. (Actually, we should speak of (strict) simplex type with respect to z_1, \dots, z_p since, as shown above, Λ_{24} is of simplex type in two essentially different ways. But in order to keep the language simple we delete the reference to z_1, \dots, z_p .) Clearly, (L2) implies that E is a Euclidean lattice. Also, for $i \neq j$, the norm of $z_i - z_j$ is $2n - 2m$; hence (L2) implies the relation

$$(2) \quad n \geq 2m.$$

In the extremal case $n = 2m$, we say that E is of *strong simplex type*.

John Leech (private communication) observed the following geometric interpretation. A p -dimensional lattice is of simplex type if the vertex figure (the set of minimal norm vertices) contains a set of p equidistant points forming a regular simplex. If the lattice is of strict simplex type, this simplex forms a cell of the vertex figure. If the lattice is of strong simplex type then the origin can be joined to give a regular $(p + 1)$ -simplex, a fundamental cell of the honeycomb of the lattice. In particular, every lattice of strong simplex type is strict.

Note that a scalar multiple $cE = \{cx | x \in E\}$ of a lattice E of simplex type is again of simplex type, with $n' = c^2n, m' = c^2m$. The quotient

$$(3) \quad d = \frac{n - m}{m} \geq 1$$

is independent of scaling, and hence a useful invariant. In their construction of the Leech lattice, Leech and Sloane use a particular standard basis with respect to which the z_i take a simple form. Motivated by this, we say that a lattice E is in *Euclidean normal form* if there are numbers p and t such that in terms of the standard basis $a_1 = (1, 0, \dots, 0)^T, \dots, a_p = (0, 0, \dots, 1)^T$ of \mathbb{R}^p , the following statements hold:

(E1) E is a Euclidean lattice of dimension p .

(E2) $z_i = \sum_{l=1}^p a_l - ta_i \in E$ for $i = 1, \dots, p$.

(E3) $0 < 2t < p \leq t^2 + 2t$.

Condition (E2) implies that (1) holds with

$$(4) \quad n = p - 2t + t^2, \quad m = p - 2t, \quad d = \frac{t^2}{p - 2t},$$

and condition (E3) guarantees that $n \geq 2m > 0$, in particular $n > m > 0$.

THEOREM 1.1. (i) *A lattice in Euclidean normal form is of simplex type if and only if its minimal norm is $n = p - 2t + t^2$.*

(ii) *Every lattice of simplex type is isomorphic to a multiple of a lattice in Euclidean normal form.*

Proof. (i) This part is clear from the preceding.

(ii) Let E be a lattice of simplex type, with parameters p, n, m and d , defined by (3). The equation $d = t^2/(p - 2t)$ has a positive solution $t = -d + \sqrt{d(d + p)}$, and (E3) holds since $d \geq 1$. Moreover, with $c = t^{-1}\sqrt{n - m}$, we have $n - m = c^2 t^2 = dm$, whence $m = c^2 t^2 d^{-1} = c^2(p - 2t)$, $n = c^2 t^2 + m = c^2(p - 2t + t^2)$. Now put $z = \sum_{i=1}^p z_i$. Then $(z, z_i) = (p - 1)m + n = c^2(p - t)^2$ and $(z, z) = \sum (z, z_i) = pc^2(p - t)^2$. Now it is easy to show that the vectors a_i defined by

$$a_i = \frac{1}{ct} \left(\frac{1}{p - t} z - z_i \right), \quad i = 1, \dots, p,$$

form an orthonormal basis of \mathbb{R}^p , and we have

$$(5) \quad z_i = c \left(\sum_{i=1}^p a_i - ta_i \right), \quad i = 1, \dots, p.$$

If we identify a_1, \dots, a_p with the standard basis of \mathbb{R}^p (which amounts to an isomorphism) we find that the lattice $c^{-1}E$ is in Euclidean normal form. \square

If E is a lattice of simplex type then the sublattice E_0 generated by z_1, \dots, z_p is also of simplex type, with the same parameters. We call E_0 the *trivial part* of E , and say that E is *trivial* if $E = E_0$, i.e., if E is generated by z_1, \dots, z_p . Since every lattice of simplex type is the refinement of a trivial lattice, it is important to determine all trivial lattices.

THEOREM 1.2. (i) *For any dimension $p \geq 2$, and any pair (m, n) with $n \geq 2m > 0$ there is a trivial lattice of strict simplex type and parameters p, n, m .*

(ii) *Two trivial lattices are isomorphic if and only if they have the same parameters p, n, m .*

Proof. (i) As in the proof of Theorem 1.1, find $c > 0$ and $d \geq 1$ such that $m = c^2(p - 2t)$, $n = c^2(p - 2t + t^2)$. For the standard basis (a_i) of \mathbb{R}^p , the vectors (5) satisfy (1); so we have to show that $E = \langle z_1, \dots, z_p \rangle$ is of strict simplex type. Now for $x = \sum \alpha_i z_i$ we have $(x, x) = \sum_{i,j} \alpha_i \alpha_j (z_i, z_j) = \sum_{i,j} \alpha_i \alpha_j (m + (n - m)\delta_{ij}) = (\sum \alpha_i)^2 m + (\sum \alpha_i^2)(n - m)$. This implies that the z_i are linearly independent ($x = 0 \rightarrow (x, x) = 0 \rightarrow \sum \alpha_i^2 = 0 \rightarrow$ all $\alpha_i = 0$), hence $\dim E = p$. Moreover, if $x \in E \setminus \{0\}$ has norm $\leq n$ then the α_i are integers not all zero, and $\sum \alpha_i^2 \leq 2$ since $2(n - m) \geq n$. Hence x is one of $\pm z_i, \pm(z_i - z_j), \pm(z_i + z_j)$, where $j \neq i$. In the last case, $(x, x) = 4m + 2(n - m) > n$, in the second case $(x, x) = 2(n - m) > n$ unless $n = 2m$ when $(x, x) = n$, and in the first case $(x, x) = n$. Hence E has minimal norm n , i.e., (L2) holds, and the vectors of minimal norm are

$$\pm z_i \quad \text{if } n > 2m, \quad \pm z_i, \pm(z_i - z_j) \quad \text{if } n = 2m.$$

In both cases, (L3) is satisfied, whence E is of strict simplex type.

(ii) By Theorem 1.1, a trivial lattice is isomorphic to one of the lattices just constructed. \square

We write $T_p^{n,m}$ for a p -dimensional trivial lattice of simplex type with parameters n, m . The lattices $T_p^{2,1}$ arise in connection with extreme forms. Coxeter's [5] extreme form A_p is the symmetric bilinear form corresponding to the lattice A_p consisting of all $x \in \mathbb{Z}^{p+1}$ with $\sum x_i = 0$. If we denote by a_1, \dots, a_{p+1} the standard basis of \mathbb{Z}^{p+1} then A_p is generated by the vectors $z_i = a_i - a_{p+1}$ ($i = 1, \dots, p$) which satisfy (1) with $n = 2, m = 1$. Since the minimal norm of A_p is $n = 2$, A_p is a trivial lattice of simplex type, hence isomorphic to $T_p^{2,1}$. Note also that the vectors of minimal norm of A_p form a root system and a spherical 3-design (see [2], [9]). The lattices $T_2^{n,m}$ arise in a classification problem. It is not difficult to show that every two-dimensional lattice generated by its vectors of minimal norm is either a trivial lattice of simplex type (i.e.,

one of the $T_2^{n,m}$), or a multiple of \mathbb{Z}^2 . In dimension $p > 2$, not every lattice generated by its vectors of minimal norm is of simplex type; for $p \geq 4$, counterexamples are the lattices D_p consisting of all $x \in \mathbb{Z}^p$ with $\sum x_i \equiv 0 \pmod 2$. On the other hand, there are lattices of simplex type not generated by their vectors of minimal norm; see Example 7 below.

Now we give some examples of nontrivial lattices of simplex type. In each case, there is a t such that the vectors of shape $(1^{p-1}, 1-t)$ are minimal norm vectors of the lattice; hence all examples are in Euclidean normal form (we use a_1, \dots, a_p as standard basis of \mathbb{R}^p). We construct examples with the parameters listed in Table 1 (τ is the number of vectors of minimal norm).

TABLE 1

p	d	τ	strict?	t	n	m	
6	2	32	yes	2	6	2	C_{16}
8	1	240	yes	2	8	4	E_8
16	2	512	yes	4	24	8	S_{256}
24	1	196,560	yes	4	32	16	Λ_{24}
24	3	196,560	no	6	48	12	Λ_{24}
48	1	52,416,000	yes	6	72	36	$P48p, P48q$

Example 1. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^6$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 2$;
- (ii) the coordinate sum $\sum x_i$ is divisible by 4.

The minimal norm of E is $n = 6$; there are 32 vectors of norm 6 (forming the polytope $h\gamma_6$, cf. Coxeter [5]), namely

$$2 \times 6 \text{ of shape } \pm(1^5, -1),$$

$$20 \text{ of shape } (1^3, (-1)^3);$$

they form a spherical 3-design. Moreover the corresponding set of 16 lines is the line system C_{16} of Shult and Yanushka [15].

Example 2. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^8$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 2$;
- (ii) the coordinate sum $\sum x_i$ is congruent to $2\gamma \pmod 4$.

This is the lattice $2E_8$ defined, e.g., in Leech and Sloane [10]. The minimal norm is 8; there are 240 vectors of norm 8, namely

$$2 \times 8 \text{ of shape } \pm(1^7, -1),$$

$$2 \times 56 \text{ of shape } \pm(1^5, (-1)^3),$$

$$4 \times 28 \text{ of shape } (\pm 2, \pm 2, 0^6);$$

they form the root systems E_8 [2]. This root system is a tight spherical 7-design [8].

Example 3. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^{16}$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 2$;
- (ii) the set of indices i for which $x_i \pmod 4$ takes a given value is a \mathcal{C} -set;
- (iii) the coordinate sum $\sum x_i$ is congruent to $4\gamma \pmod 8$.

Here a \mathcal{C} -set is either \emptyset , or $\{1, \dots, 16\}$, or a hyperplane in an affine space $AG(4, 2)$ whose points are labelled $1, \dots, 16$. (There are 30 such hyperplanes, and the \mathcal{C} -sets form a self-orthogonal linear code with respect to the symmetric difference.) The

minimal norm is 24; there are 512 vectors of norm 24, namely

$$\begin{aligned} &2 \times 16 \text{ of shape } \pm(1^{15}, -3), \\ &2 \times 240 \text{ of shape } \pm(1^8, 3, (-1)^7), 1^8 \text{ on a hyperplane.} \end{aligned}$$

These vectors form a spherical 5-design (Neumaier [12]). The corresponding set of 256 lines can be shown to be isomorphic to the line system S_{256} of Shult and Yanushka [15].

Example 4. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^{24}$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 2$;
- (ii) the set of indices i for which $x_i \pmod 4$ takes a given value is a \mathcal{C} -set;
- (iii) the coordinate sum $\sum x_i$ is congruent to $4\gamma \pmod 8$;

but this time a \mathcal{C} -set is a subset of $\{1, \dots, 24\}$ whose characteristic vector belongs to the binary Golay code. This is Conway's [3] description of the Leech lattice $\sqrt{8} \Lambda_{24}$. The minimal norm of E is $n = 32$; there are 196,560 vectors of norm 32; they form a tight spherical 11-design [8].

Example 5. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^{24}$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 2$;
- (ii) $(x_1, \dots, x_{24}) \pmod 3$ is a codeword of \mathcal{C} ;
- (iii) the coordinate sum $\sum x_i$ is congruent to $2\gamma \pmod 4$.

Now \mathcal{C} is either the ternary $(24, 3^{12}, 9)$ quadratic residue code, or the $(24, 3^{12}, 9)$ symmetry code; they both contain the words $\pm(1^{24})$. Hence by Leech and Sloane [10, § 5.7], we get in both cases $\sqrt{12} \Lambda_{24}$, with minimal norm 48. As remarked in the introduction, E is not of strict Leech type.

Example 6. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^{48}$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 2$;
- (ii) $(x_1, \dots, x_{48}) \pmod 3$ is a codeword of \mathcal{C} ;
- (iii) the coordinate sum $\sum x_i$ is congruent to $2\gamma \pmod 4$;

but \mathcal{C} is the $(48, 3^{24}, 15)$ ternary quadratic residue code or symmetry code. By Leech and Sloane [10, § 5.7], we get the extremal lattices $P48q$ and $P48p$ with minimal norm 72, and 52,416,000 minimal norm vectors, cf. also Sloane [16].

Note that in all the examples given so far, the vectors of minimal norm generate the lattice. In the next example, the situation is different.

Example 7. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^{15}$ with

- (i) all x_i are integers congruent to the same value $\gamma \pmod 3$;
- (ii) the coordinate sum $\sum x_i$ is divisible by 12.

The minimal norm is 18; there are 240 vectors of norm 18, namely

$$\begin{aligned} &2 \times 15 \text{ of shape } \pm(1^{14}, -2), \\ &210 \text{ of shape } (3, -3, 0^{13}). \end{aligned}$$

But $(3, -3, 0^{13}) = (1, -2, 1^{13}) - (-2, 1, 1^{13})$; hence the 240 vectors only generate the trivial part of E , which does not contain, say, $12a_1$.

We close this section with some examples whose simplest presentation is not the Euclidean normal form.

Example 8. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^7$ such that:

- (i) all x_i are integers;
- (ii) the set of indices i for which x_i is odd is a \mathcal{C} -set;

where a \mathcal{C} -set is either \emptyset , or the complement of a line in the Fano plane $PG(2, 2)$ whose points are labelled $1, 2, \dots, 7$ (cf. Example 19; the \mathcal{C} -sets form a linear code with respect to symmetric difference). This is the lattice $\sqrt{2} E_7$. The minimal norm

is 4, and E contains 126 norm 4 vectors,

$$14 \text{ of shape } (\pm 2, 0^6),$$

$$2^4 \times 7 \text{ of shape } ((\pm 1)^4, 0^3), \text{ zeros on a line.}$$

The lattice is of strong simplex type with respect to the simplex consisting of the 7 vertices of shape $(1^4, 0^3)$.

Example 9. Let E consist of all vectors $\sum x_i a_i \in \mathbb{R}^{15}$ such that

- (i) all x_i are integers;
- (ii) the set of indices i for which x_i is odd is a \mathcal{C} -set;
- (iii) the coordinate sum $\sum x_i$ is divisible by 4,

where now a \mathcal{C} -set is either \emptyset , or the complement of a plane in the projective space $PG(3, 2)$ whose points are labelled $1, 2, \dots, 15$ (the planes can be taken as the 15 sets $\{i, i + 1, i + 2, i + 4, i + 5, i + 8, i + 10\} \bmod 15$; the \mathcal{C} -sets again form a linear code). This is the lattice Λ_{15} of Leech and Sloane [10, § 3.4]. The minimal norm is 8, and E contains 2,340 norm 8 vectors,

$$4 \times \binom{15}{2} \text{ of shape } ((\pm 2)^2, 0^{13}),$$

$$15 \times 2^7 \text{ of shape } ((\pm 1)^8, 0^7), \text{ zeros on a plane, even times +.}$$

The lattice is of strong simplex type with respect to the simplex consisting of the 15 vertices of shape $(1^8, 0^7)$.

Example 10. Similarly, the lattice obtained from Λ_{32} in Leech and Sloane [10] by equating a coordinate to zero gives a lattice of strong simplex type in \mathbb{R}^{31} with respect to the simplex whose 31 vertices are those of shape $(1^{16}, 0^{15})$ with zeros on the coordinates of a hyperplane of $PG(4, 2)$.

2. Hyperbolic transforms. Let E be a lattice in \mathbb{R}^p of simplex type, with parameters $n, m, d = (n - m)/m$. We adjoin to \mathbb{R}^p an element w orthogonal to \mathbb{R}^p and of norm $-d$. In this way we get a hyperbolic space $\mathbb{R}^p \oplus \mathbb{R}w$. This space contains the *hyperbolic transform* of E which we define as the hyperbolic lattice

$$H = c^{-1}E \oplus d^{-1}\mathbb{Z}w \quad (c = \sqrt{n - m});$$

the hyperbolic transform of an individual element $z \in E$ is defined as the element

$$x = c^{-1}z - d^{-1}w \in H.$$

PROPOSITION 2.1. *Let E be a p -dimensional lattice of simplex type. The hyperbolic transform H of E has the following properties:*

- (M1) H contains vectors e_1, \dots, e_p such that for $i, j = 1, \dots, p$,

$$(e_i, w) = (e_i, e_i) = 1, \quad (e_i, e_j) = 0 \quad \text{if } i \neq j.$$

- (M2) For all $x \in H$ linearly independent from w ,

$$(x, w) = 1 \Rightarrow (x, x) \geq 1.$$

Moreover, if E is of strict simplex type, then H also satisfies:

- (M3) If x, y are distinct vectors from the set

$$H_{1,1} = \{x \in H \mid (x, w) = (x, x) = 1\}$$

then $(x, y) \leq 0$.

Proof. (i) The vectors $e_i = c^{-1}z_i - d^{-1}w$ ($i = 1, \dots, p$) belong to H and satisfy $(e_i, w) = -d^{-1}(w, w) = 1$, $(e_i, e_j) = c^{-2}(z_i, z_j) + d^{-2}(w, w) = (n - m)^{-1}(m + (n - m)\delta_{ij}) - d^{-1} = \delta_{ij}$. Hence (M1) holds.

(ii) If $x \in H \setminus \langle w \rangle$ and $(x, w) = 1$ then $x = c^{-1}z - d^{-1}w$ with $z \in E \setminus \{0\}$, and by (L2), $(x, x) = c^{-2}(z, z) + d^{-2}(w, w) \geq c^{-2}n - d^{-1} = 1$. Hence (M2) holds.

(iii) If $x \in H_{1,1}$ then $(x, w) = 1$ implies that $x = c^{-1}z - d^{-1}w$ with $z \in E$. Now $c^{-2}(z, z) = (x + d^{-1}w, x + d^{-1}w) = (x, x) + 2d^{-1}(x, w) + d^{-2}(w, w) = 1 + 2d^{-1} - d^{-1} = (n - m)^{-1}n$, hence $(z, z) = n$. If $y \in H_{1,1}$ is distinct from x then similarly $y = c^{-1}z' - d^{-1}w$ with $z' \in E$, $(z', z') = n$, and $z' \neq z$. If E is of strict Leech type then by (L3), $(x, y) = c^{-2}(z, z') + d^{-2}(w, w) \leq c^{-2}m - d^{-1} = 0$, whence (M3) holds. \square

We denote by $\mathbb{R}^{p,1}$ the standard hyperbolic space, i.e., the real linear space of dimension $p + 1$ equipped with the indefinite inner product

$$(x, y) = -x_0y_0 + x_1y_1 + \dots + x_p y_p.$$

Let us say that a lattice $H \subseteq \mathbb{R}^{p,1}$ is of *hyperbolic simplex type* with respect to $w \in \mathbb{R}^{p,1}$ if $(w, w) = -d \leq 1$ and (M1) and (M2) hold; we say that H is *strict* if also (M3) holds. Trivial examples of lattices of strict hyperbolic simplex type are the lattices $\langle e_1, \dots, e_p, d^{-1}w \rangle$, where

$$(6) \quad w = -\sqrt{N}e_0 + e_1 + \dots + e_p, \quad N \geq p + 1$$

and $e_0 = (1; 0, \dots, 0)^T$, $e_1 = (0; 1, \dots, 0)^T, \dots, e_p = (0; 0, \dots, 1)^T$ is the standard basis of $\mathbb{R}^{p,1}$. These lattices are just the hyperbolic transforms of trivial lattices of simplex type, with $d = N - p$.

THEOREM 2.2. (i) *The hyperbolic transform of a lattice of (strict) simplex type is a lattice of (strict) hyperbolic simplex type.*

(ii) *Let H be a lattice of (strict) hyperbolic simplex type with respect to a vector w satisfying $(w, w) = -d \leq -1$. If $d^{-1}w \in H$ then $E = w^\perp \cap H = \{x \in H \mid (x, w) = 0\}$ is a lattice of (strict) simplex type with parameters $n = 1 + d^{-1}$, $m = d^{-1}$.*

Proof. Part (i) follows directly from the definition and Proposition 2.1.

(ii) The vectors $z_i = e_i + d^{-1}w$ ($i = 1, \dots, p$) satisfy $(z_i, z_j) = (e_i, e_j) + d^{-1}(e_i, w) + d^{-1}(e_j, w) + d^{-2}(w, w) = \delta_{ij} + d^{-1}$, hence E satisfies (L1) with $n = 1 + d^{-1}$, $m = d^{-1}$. Further, if $x \in E \setminus \{0\}$ then $x' = x - d^{-1}w$ is in $H \setminus \langle w \rangle$ and satisfies $(x', w) = 1$, whence by (M2), $1 \leq (x', x') = (x, x) + d^{-2}(w, w) = (x, x) - d^{-1}$, and so $(x, x) \geq 1 + d^{-1} = n$. Hence (L2) holds. Finally, if x, y are distinct norm $1 + d^{-1}$ vectors of E then $x' = x - d^{-1}w$ and $y' = y - d^{-1}w$ are distinct vectors in $H_{1,1}$. Hence, if H is strict, $0 \geq (x', y') = (x, y) + d^{-2}(w, w) = (x, y) - d^{-1}$, whence $(x, y) \leq d^{-1} = m$. So (M3) holds if H is strict. \square

Lattices of hyperbolic simplex type have the following obvious hereditary property:

PROPOSITION 2.3. *If H is a $(p + 1)$ -dimensional lattice of (strict) hyperbolic simplex type with respect to w , then every p -dimensional section $H^{(i)} = e_i^\perp \cap H = \{x \in H \mid (x, e_i) = 0\}$ ($i = 1, \dots, p$) is of (strict) hyperbolic simplex type with respect to $w^{(i)} = w - e_i$.*

Note, that if $(w, w) = -d$, then $(w^{(i)}, w^{(i)}) = -d - 1$. Hence the number

$$(7) \quad N = p + d$$

remains invariant under forming sections. In fact, the vector $e_0 = N^{-1/2}(-w + e_1 + \dots + e_p)$ satisfies $(e_0, e_0) = -1$, $(e_0, e_i) = 0$ for $i = 1, \dots, p$; hence $e_0; e_1, \dots, e_p$ can be identified with the standard basis of $\mathbb{R}^{p,1}$. In the following, we shall do this. Then w is given by formula (6) above.

Remarks. 1. Even if H is a hyperbolic transform, i.e., $d^{-1}w \in H$, then, in general, $H^{(i)}$ will not contain $(d+1)^{-1}w^{(i)}$, and hence will not be a hyperbolic transform. In particular, lattices of hyperbolic simplex type form a richer class than lattices of (Euclidean) simplex type. But it can be shown that if H is a lattice of hyperbolic simplex type with respect to w and H contains a multiple of $d^{-1}w$ then the projection

$$E = \left\{ x - \frac{(x, w)}{(w, w)} w \mid x \in H \right\}$$

of H onto w^\perp is a Euclidean lattice satisfying (L1) and:

(L2*) If $z = \sum \alpha_i z_i \in E$ then $\sum \alpha_i = 1$ implies $\sum \alpha_i^2 \geq 1$.

Conversely, if a Euclidean lattice E satisfies (L1) and (L2*) then there is a lattice H of hyperbolic simplex type (with respect to w) such that E is isomorphic to a multiple of the projection of H onto w^\perp .

2. If H is a lattice of hyperbolic simplex type with respect to w and if H contains a nonzero, integral multiple of $d^{-1}w$, then the set $H_{1,1}$ is finite. Indeed, the vectors of $H_{1,1}$ are projected to vectors $\sum \alpha_i z_i \in w^\perp$ with

$$d(\sum \alpha_i^2) + (\sum \alpha_i)^2 = d + 1.$$

Since this equation describes a bounded domain, the projection of $H_{1,1}$ —being a bounded part of a Euclidean lattice—is finite. But each vector of $H_{1,1}$ is determined by its projection.

3. If H is of strict hyperbolic simplex type then, by Andreev’s lemma (cf. Vinberg [18, p. 19]), the set $H_{1,1}$ determines a hyperbolic polyhedron with the property that hyperplanes corresponding to nonadjacent faces do not intersect.

4. If H is of (strict) hyperbolic simplex type then the sublattice $H' = \{x \in H \mid (x, w) \text{ integral}\}$ is also of strict hyperbolic simplex type and $H'_{1,1} = H_{1,1}$. Hence the following axiom can always be forced to hold:

(M4) For all $x \in H$, (x, w) is integral.

Note that (M4) is trivially satisfied if H is a hyperbolic transform, or if H is generated by $H_{1,1}$.

Example 11. The following example, considered first by Du Val [19], is studied extensively in Manin [11, Chapt. 4], in connection with cubic forms, and led me to the study of hyperbolic transforms. Let e_0, e_1, \dots, e_p be the standard basis of \mathbb{R}^{p+1} , and let $H = \mathbb{Z}^{p+1}$ be the lattice generated by e_0, e_1, \dots, e_p . For $p \leq 8$, H is of strict hyperbolic simplex type with respect to $w = -3e_0 + e_1 + \dots + e_p$. In fact, $(w, w) = 9 - p \leq -1$, and (M1) is obvious. Further if $x = \alpha e_0 - \sum_{i=1}^p \alpha_i e_i \in H$ and $(x, w) = 1$ then α, α_i are integers and $\sum \alpha_i = 3\alpha - 1$. Now consider $(x, x) = -\alpha^2 + \sum \alpha_i^2$. Modulo 2 we get $(x, x) \equiv -\alpha + \sum \alpha_i \equiv 2\alpha - 1 \equiv 1$, whence (x, x) is odd. By the Cauchy-Schwarz inequality $\sum \alpha_i^2 \geq (\sum \alpha_i)^2 / p \geq (3\alpha - 1)^2 / 8$, whence $8(x, x) \geq -8\alpha^2 + (3\alpha - 1)^2 = (\alpha - 3)^2 - 8 \geq -8$. But if $(x, x) = -1$ then we have equality throughout, whence $\alpha = 3, p = 8, \sum \alpha_i^2 = \sum \alpha_i = 8$, which implies $x = -w$. Since (x, x) is an odd integer, $(x, x) \geq 1$ for $x \neq -w$, so (M2) holds. Finally, if $x \in H_{1,1}$ then $1 = (x, x) \geq (\alpha - 3)^2 - 1$ or $3 - \sqrt{2} \leq \alpha \leq 3 + \sqrt{2}$, and since α is an integer, $-1 \leq \alpha \leq 4$. Now the equations $\sum \alpha_i = 3\alpha - 1, \sum \alpha_i^2 = \alpha^2 + 1, -1 \leq \alpha \leq 4$ have only finitely many integral solutions which lead to the list of vectors in $H_{1,1}$, given by Manin [11, Prop. 26.1] and shown in Table 2.

From this list, (M3) is easily verified. In fact, for $p = 8, H = \mathbb{Z}^{p+1}$ is the hyperbolic transform of the root lattice E_8 ; this can be seen from the fact that both E_8 (as defined above) and $E = w^\perp \cap H$ are generated by 8 special norm 2 vectors whose mutual inner products determine the Dynkin diagrams for E_8 (i.e., the inner product of two vectors is -1 if they are adjacent in the diagram, and 0 otherwise). See Figs. 1 and 2, and Coxeter [5].

TABLE 2

Type of vector	$\mathbb{R}^{3,1}$	$\mathbb{R}^{4,1}$	$\mathbb{R}^{5,1}$	$\mathbb{R}^{6,1}$	$\mathbb{R}^{7,1}$	$\mathbb{R}^{8,1}$
e_1	3	4	5	6	7	8
$e_0 - e_1 - e_2$	3	6	10	15	21	28
$2e_0 - e_1 - e_2 - e_3 - e_4 - e_5$			1	6	21	56
$3e_0 - 2e_1 - e_2 - e_3 - e_4 - e_5 - e_6 - e_7$					7	56
$4e_0 - 2e_1 - 2e_2 - 2e_3 - e_4 - e_5 - e_6 - e_7 - e_8$						56
$5e_0 - 2e_1 - 2e_2 - 2e_3 - 2e_4 - 2e_5 - 2e_6 - e_7 - e_8$						28
$6e_0 - 3e_1 - 2e_2 - 2e_3 - 2e_4 - 2e_5 - 2e_6 - 2e_7 - 2e_8$						8
Total number of vectors	6	10	16	27	56	240
Automorphism group	D_6	S_5	2^4S_5	$O_6^-(2)$	$2Sp(6, 2)$	$2O_8^-(2)$
Configuration name	prism	$t\alpha_4$	$h\gamma_5$	2_{21}	3_{21}	4_{21}

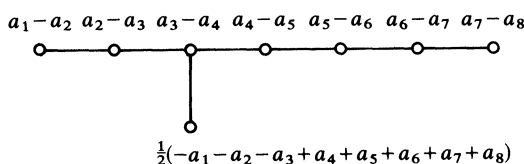


FIG. 1

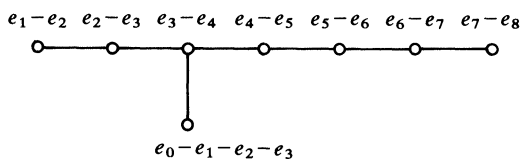


FIG. 2

The sets $H_{1,1}$ are related to interesting classical configurations. Their Euclidean projections are the successive vertex figures of the honeycomb 5_{21} ($=E_8$), the Gosset polytopes 4_{21} , 3_{21} , 2_{21} and the polytopes $h\gamma_5$ and $t\alpha_4$ (cf. Coxeter [5]). The sets $H_{1,1}$ are also related to certain famous graphs. For $p \leq 6$, the only occurring inner products of distinct vectors are 0 and -1 , and the graphs obtained by calling two vectors adjacent if they have inner product -1 turn out to be the hexagon ($p = 3$), the Peterson graph ($p = 4$), the Clebsch graph ($p = 5$) and the Schläfli graph ($p = 6$) (Seidel [13]); the latter is related to the 27 lines on a cubic surface (Baker [1]). For $p = 7$, the same construction yields a graph with 56 vertices, related to the regular twograph on 28 vertices (Taylor [17]) and to the 28 bitangents of a plane quartic curve (Dickson [6]). For $p = 8$, $H_{1,1}$ is also related to the set of 240 Cayley units (Coxeter [4]).

Example 12. Let H be the lattice generated by $\sqrt{2}e_0, e_1, \dots, e_p$. For $p \leq 7$, H is of strict hyperbolic simplex type with respect to $w = -2\sqrt{2}e_0 + e_1 + \dots + e_p$. This is proved as above and for $H_{1,1}$ we get Table 3.

For $p = 7$, we get the hyperbolic transform of the root lattice E_7 , using Fig. 3 as Dynkin diagram. The Euclidean normal form for E_7 would have $t = 2\sqrt{2} - 1$, and does not lead to a nice description (but cf. Example 8). For $p = 6$, we get the hyperbolic transform of the lattice defined in Example 1. For $p = 4, 5$, the graphs corresponding to $H_{1,1}$ are the cube and the complement of the triangular graph $T(6)$.

TABLE 3

Type of vector	$\mathbb{R}^{4,1}$	$\mathbb{R}^{5,1}$	$\mathbb{R}^{6,1}$	$\mathbb{R}^{7,1}$
e_1	4	5	6	7
$\sqrt{2}e_0 - e_1 - e_2 - e_3$	4	10	20	35
$2\sqrt{2}e_0 - 2e_1 - e_2 - e_3 - e_4 - e_5 - e_6$			6	42
$3\sqrt{2}e_0 - 2e_1 - 2e_2 - 2e_3 - 2e_4 - e_5 - e_6 - e_7$				35
$4\sqrt{2}e_0 - 3e_1 - 2e_2 - 2e_3 - 2e_4 - 2e_5 - 2e_6 - 2e_7$				7
Total number of vectors	8	15	32	126
Automorphism group	2^3S_3	S_6	2^5S_6	$2Sp(6, 2)$

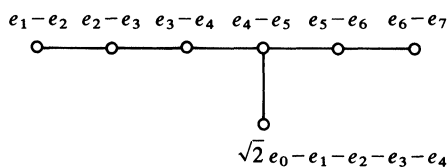


FIG. 3

Remark. The lattice E_6 is not of simplex type; cf. Theorems 3.5 and 3.6.

Example 13. Let H be the lattice generated by $\sqrt{N}e_0, e_1, \dots, e_p$. For $p \leq N - 1$, H is of strict hyperbolic simplex type with respect to $w = -\sqrt{N}e_0 + e_1 + \dots + e_p$. The table for $H_{1,1}$ is Table 4. For $p = N - 1, d = 1$, and H is also generated by w and e_1, \dots, e_p ; by a remark above, H is the hyperbolic transform of the trivial lattice $T_p^{2,1} = A_p$. Since A_p is a root lattice, we have again a Dynkin diagram (see Fig. 4). For $p = N - 2, H$ is the hyperbolic transform of \mathbb{Z}^p .

TABLE 4

Type of vector	$\mathbb{R}^{N-2,1}$	$\mathbb{R}^{N-1,1}$
e_1	$N - 2$	$N - 1$
$\sqrt{N}e_0 - 2e_1 - e_2 - \dots - e_{N-2}$	$N - 2$	$(N - 1)(N - 2)$
$2\sqrt{N}e_0 - 3e_1 - 2e_2 - \dots - 2e_{N-1}$		$N - 1$
Total number of vectors	$2(N - 2)$	$N(N - 1)$

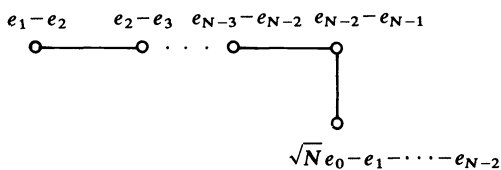


FIG. 4

Example 14. Let H be the lattice consisting of all vectors $\alpha_0\sqrt{3}e_0 - \alpha_1e_1 - \dots - \alpha_{23}e_{23} \in \mathbb{R}^{23,1}$ such that:

- (i) for $i = 0, \dots, 23$, the number $2\alpha_i$ is integral;
- (ii) the set of indices $i \in \{0, \dots, 23\}$ such that α_i is nonintegral is a \mathcal{C} -set;
- (iii) $\alpha_0 + \dots + \alpha_{23}$ is integral,

where \mathcal{C} is the binary Golay code (cf. Goethals and Seidel [7], also for the numbers quoted below). H is of strict hyperbolic simplex type with respect to $w = -3\sqrt{3}e_0 + e_1 + \dots + e_{23}$, and $H_{1,1}$ (resp. their sections) contain the vectors listed in Table 5.

Example 15. Let H be the hyperbolic transform of the Leech lattice Λ_{24} . From the work of Conway [3], it can be shown that the table for $H_{1,1}$ and its sections is Table 6.

Example 16. Similarly, Shult and Yanushka's calculations [15] for their line system S_{256} imply that the hyperbolic transform of the lattice constructed in Example 3 has Table 7 as table for $H_{1,1}$.

Remark. The projections of the sets $H_{1,1}$ given in the examples form interesting spherical t -designs; cf. Delsarte, Goethals and Seidel [8].

TABLE 5

Type of vector	$\mathbb{R}^{19,1}$	$\mathbb{R}^{20,1}$	$\mathbb{R}^{21,1}$	$\mathbb{R}^{22,1}$	$\mathbb{R}^{23,1}$
e_1	19	20	21	22	23
$\frac{1}{2}\sqrt{3}e_0 - \frac{1}{2}(e_1 + \dots + e_7)$	52	80	120	176	253
$\sqrt{3}e_0 - \frac{1}{2}(e_1 + \dots + e_{16})$	1	5	21	77	253
$\frac{3}{2}\sqrt{3}e_0 - \frac{3}{2}e_1 - \frac{1}{2}(e_2 + \dots + e_{23})$					23
Total number of vectors	72	105	162	275	2 · 276
Automorphism group	?	$L_3(4)$	$U_4(3)$	McL	2 Con. 3

TABLE 6

	$\mathbb{R}^{21,1}$	$\mathbb{R}^{22,1}$	$\mathbb{R}^{23,1}$	$\mathbb{R}^{24,1}$
Total number of vectors	336	891	4,600	196,560
Automorphism group	$2^9L_3(4)$	$U_6(2)$	2 Con. 2	Con. 0

TABLE 7

	$\mathbb{R}^{14,1}$	$\mathbb{R}^{15,1}$	$\mathbb{R}^{16,1}$
Total number of vectors	56	135	512
Automorphism group	$2^3L_3(2)?$	$Sp(6, 2)$	$2^{8+1}Sp(6, 2)$

3. Integral lattices of simplex type. In this section we classify the *integral* lattices of simplex type into standard and exceptional lattices. The aim is to determine all integral lattices of simplex type; but we are very far from achieving this.

For $(m, n) = 1$, we relate the standard lattices of simplex type to certain linear

self-orthogonal codes. From this it appears that standard lattices of simplex type are very abundant. On the other hand, restricting ourselves to integral lattices generated by their minimal norm vectors, we are able to show that for given minimal norm n there are only finitely many such *exceptional* lattices. For $n = 2$ and $n = 3$ we push the analysis further, and find close relations to star-closed and tetrahedrally-closed line systems (Cameron et al. [2]. Shult, Yanushka [15]).

Let E be an integral lattice of simplex type with (integral) parameters n, m . We call E *standard* if $(n - m)E$ is contained in the trivial part E_0 of E , and *exceptional* otherwise. Note that this definition is no longer scaling invariant. In particular, a suitable multiple cE of any integral lattice E of simplex type is always standard (take, e.g., for c the discriminant of E). We also mention that an exceptional lattice of simplex type contains a unique maximal standard sublattice, namely $E' = \{x \in E | (n - m)x \in E_0\}$.

We treat the standard case first. If $n - m = 1$ then E must be trivial. But since n, m are integers with $n \geq 2m > 0$, this happens only for $n = 2$. In particular, we have:

PROPOSITION 3.1. *A standard integral lattice of simplex type with minimal norm 2 is isomorphic to some A_p .*

If $n - m \geq 2$ then let us define a code C over the integers mod $n - m$, consisting of those p -tuples $(\beta_1, \dots, \beta_p) \pmod{n - m}$ such that

$$(8) \quad x = \frac{1}{n - m} \sum \beta_i z_i, \quad \beta_i \text{ integral}$$

is in E . Since E is standard and $z_1, \dots, z_p \in E$, this code describes E completely.

PROPOSITION 3.2. *C is a linear code. Moreover, if $(m, n) = 1$ then C is self-orthogonal, and orthogonal to $(1, \dots, 1)$.*

Proof. If $\beta = (\beta_1, \dots, \beta_p), \beta' = (\beta'_1, \dots, \beta'_p) \in C$, and if x (as in (8)) and $x' = (n - m)^{-1} \sum \beta'_i z_i$ are corresponding vectors of E then $\beta + \beta' = (\beta_1 + \beta'_1, \dots, \beta_p + \beta'_p)$ is the codeword belonging to $x + x'$. Hence C is linear. Since $z_i \in E$, the inner product

$$(x, z_i) = \frac{1}{n - m} \sum \beta_l (z_l, z_i) = \frac{m}{n - m} \sum \beta_l + \beta_i$$

is integral, say, $= \alpha + \beta_i$, and we have

$$(9) \quad \sum \beta_l = \frac{n - m}{m} \alpha, \quad \alpha \text{ integral.}$$

If $(m, n) = 1$, (9) implies $\sum \beta_l \equiv 0 \pmod{n - m}$, whence C is orthogonal to $(1, \dots, 1)$. Finally, if x and x' are as above then their inner product is given by

$$(10) \quad (x, x') = \frac{m}{(n - m)^2} (\sum \beta_l) (\sum \beta'_l) + \frac{1}{n - m} \sum \beta_l \beta'_l.$$

Now (x, x') is integral, and for $(m, n) = 1$, the first term on the right-hand side of (10) is integral. Hence the second term is also integral, i.e., $\sum \beta_l \beta'_l \equiv 0 \pmod{n - m}$, and C is self-orthogonal. \square

We are now ready to describe all standard integral lattices of simplex type with minimal norm 3.

THEOREM 3.3. *The standard integral lattices of simplex type with minimal norm 3 are just the lattices E consisting of all vectors $\frac{1}{2}(\beta_1 z_1 + \dots + \beta_p z_p)$ with integers β_i such that $(\beta_i, \dots, \beta_p) \pmod 2$ is in a given binary even self-orthogonal code C of minimum weight at least 6. Such a lattice is generated by its norm 3 vectors if and only if C is generated by its words of weight 6.*

Proof. Since $n \geq 2m > 0$ and m is integral we have $n = 3, m = 1$. Since E is standard, Proposition 3.2 implies that E consists of all vectors $x = \frac{1}{2}(\beta_1 z_1 + \dots + \beta_p z_p)$ with $\beta = (\beta_1, \dots, \beta_p) \bmod 2 \in C$, where C is binary ($n - m = 2$), even (here = orthogonal to $(1, \dots, 1)$), and self-orthogonal. Clearly any such lattice L contains z_1, \dots, z_p , and it is easy to see from (10) that L is an integral lattice. So the only question is whether E contains vectors of norm smaller than $n = 3$. Now x has norm $\mu = \frac{1}{4}(\sum \beta_i)^2 + \frac{1}{2}(\sum \beta_i^2)$. A straightforward calculation shows that $\mu \leq 2$ if and only if β is of type $\pm(1^2, 0^{p-2}), (1, -1, 0^{p-2})$ or $(1^2, -1^2, 0^{p-4})$. Hence E has minimal norm 3 (and is of simplex type) if and only if C has minimal weight ≥ 6 . Finally, $\mu = 3$ if and only if β is of type $(2, 0^{p-1})$ or $(1^3, -1^3, 0^{p-6})$; two other possibilities $\pm(1^3, -1, 0^{p-4})$ and $\pm(1^2, -2, 0^{p-3})$ cannot occur if C has minimal weight ≥ 6 . Hence E is generated by norm 3 vectors if and only if C is generated by words of weight 6. \square

Example 17. Let $p = 6, C = \{(0^6), (1^6)\}$. This gives us again the lattice of Example 1.

Example 18. Replace in the Pasch configuration shown in Fig. 5 each point by two, and let C be the code of length 12 generated by the characteristic vectors of the resulting 4 sets of size 6. The corresponding lattice is of simplex type, has dimension 12, and contains 104 norm 3 vectors, namely

$$2 \times 12 \text{ from } \beta_i = \pm(2, 0^{11}),$$

$$20 \times 4 \text{ from } \beta_i = (1^3, -1^3, 0^6).$$

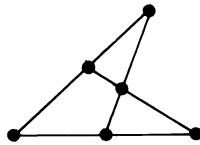


FIG. 5

Example 19. In the same way, the Fano plane in Fig. 6 gives a 14-dimensional lattice of simplex type with $2 \times 14 + 20 \times 7 = 168$ minimal norm vectors.

Example 20. Let \mathcal{C} be the code generated by the 16 characteristic vectors of sets of the shape shown in the 4×4 -grid in Fig. 7 (this is a well-known biplane). The corresponding 16-dimensional lattice is of simplex type and there are $2 \times 16 + 20 \times 16 = 352$ minimal norm vectors.

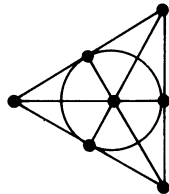


FIG. 6

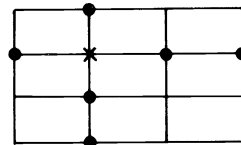


FIG. 7

Now let us consider exceptional integral lattices of simplex type.

THEOREM 3.4. *For given minimal norm n , there are only finitely many exceptional integral lattices of simplex type generated by norm n vectors.*

Proof. Suppose that E is exceptional, and generated by norm n vectors. Then there is a norm n vector $x \in E$ such that $(n - m)x \notin \langle z_1, \dots, z_p \rangle$. Put

$$(11) \quad x = \frac{1}{n - m} \sum \alpha_i z_i.$$

Then

$$(12) \quad (x, z_i) = \frac{m}{n-m} \sum \alpha_i + \alpha_i \in \mathbb{Z},$$

$$(13) \quad (x, x) = \frac{m}{(n-m)^2} (\sum \alpha_i)^2 + \frac{1}{n-m} (\sum a_i^2) = n.$$

Now not all α_i are integral, hence (12) implies that

$$(14) \quad \frac{m}{n-m} \sum \alpha_i = \alpha + \varepsilon, \quad \alpha \text{ integral}, \quad 0 < \varepsilon < 1.$$

From this, we get

$$(15) \quad \alpha_i = \beta_i - \varepsilon, \quad \beta_i \text{ integral}.$$

Substituting (15) into (14) and (13) gives, after some simplification,

$$(16) \quad \sum \beta_i = \frac{1}{m} \{ \alpha(n-m) + \varepsilon(n+m(p-1)) \},$$

$$(17) \quad \sum \beta_i(\beta_i - 1) = n(n-m) - \frac{1}{m} \{ \alpha(\alpha+1)(n-m) + \varepsilon(1-\varepsilon)(n+m(p-1)) \}.$$

Now the left-hand side of (17) is a nonnegative integer, and $\alpha(\alpha+1)(n-m) \geq 0$. Hence we find

$$(*) \quad \varepsilon(1-\varepsilon)(n+m(p-1)) \text{ is an integer } \leq mn(n-m) < n^3.$$

Also, (16) implies that ε is rational, hence $\varepsilon = u/v$ with coprime integers u, v and $0 < u < v$ since $0 < \varepsilon < 1$. Now (*) implies that $n+m(p-1) = v^2w$ with an integer $w > 0$, and the resulting inequality is $u(v-u)w < n^3$. Since $u, v-u$ and w are positive integers, this leaves only finitely many choices for u, v and w . Also m is bounded by $0 < m \leq n/2$. Hence p takes only finitely many values. So the theorem is proved if we show that for given m, n, p there are only finitely many integral lattices of simplex type with these parameters. But, in fact, m, n, p determine a unique trivial lattice of simplex type $T_p^{n,m}$, and any integral lattice has only finitely many integral refinements. Since every lattice of simplex type is a refinement of its trivial part, the proof is completed. \square

Remark. If $d = (n-m)/m$ is integral then (16), (17) can be written as

$$(16') \quad \sum \beta_i - d\alpha = \varepsilon(p+d),$$

$$(17') \quad \sum \binom{\beta_i}{2} + d \binom{\alpha+1}{2} = \frac{1}{2}(m^2d(d+1) - \varepsilon(1-\varepsilon)(p+d)).$$

Since the left-hand side is a nonnegative integer and $d(d+1)$ is even, we have

$$(**) \quad \varepsilon(1-\varepsilon)(p+d) \text{ is an even integer } \leq m^2d(d+1).$$

Proceeding as before, we find that $\varepsilon = u/v, p+d = v^2w$ where u, v, w are positive integers such that $u < v$ and $u(v-u)w$ is an even integer $\leq m^2d(d+1)$. We shall use equations (16') and (17') to determine the exceptional integral lattices of simplex type with $n = 2$.

THEOREM 3.5. *The only exceptional integral lattices of simplex type with minimal norm 2 are the root lattices $E_7 = \langle z_1, \dots, z_7, \frac{1}{2}(z_1 + \dots + z_7) \rangle$ in \mathbb{R}^7 and $E_8 =$*

$\langle z_1, \dots, z_8, \frac{1}{3}(z_1 + \dots + z_8) \rangle$ in \mathbb{R}^8 ; here

$$(z_i, z_j) = \begin{cases} 2 & \text{if } i = j, \\ 1 & \text{if } i \neq j. \end{cases}$$

Proof. We have $n = 2, m = 1$, hence $d = 1$. The numbers u, v, w of the remark above must satisfy $u(v - u)w = 2$, leaving $(u, v, w) = (1, 2, 2), (1, 3, 1)$ or $(2, 3, 1)$. Hence either $p = 7, \varepsilon = \frac{1}{2}$ or $p = 8, \varepsilon \in \{\frac{1}{3}, \frac{2}{3}\}$. To find the exceptional vectors it is sufficient to solve (16'), (17') for $\alpha \geq 0$ (otherwise replace x by $-x$). For $p = 7$, (16') and (17') hold if and only if $\alpha = 0, \beta = (1^4, 0^3)$, or equivalently if and only if $\alpha_i = \frac{1}{2}(1^4, (-1)^3)$; hence the shape of exceptional norm 2 vectors is $\pm \frac{1}{2}(z_1 + z_2 + z_3 + z_4 - z_5 - z_6 - z_7)$. Each such vector with z_1, \dots, z_7 generates E_7 ; therefore $E_7 \subseteq E$. Similarly, for $p = 8$ we find $\alpha = 0, \beta = (1^3, 0^5)$ or $(1^6, 0^2)$, whence $\alpha_i = \frac{1}{3}(2^3, (-1)^5)$ or $\frac{1}{3}(1^6, (-2)^2)$. Therefore the shape of an exceptional vector is $\pm \frac{1}{3}(2z_1 + 2z_2 + 2z_3 - z_4 - z_5 - z_6 - z_7 - z_8)$ or $\pm \frac{1}{3}(z_1 + z_2 + z_3 + z_4 + z_5 + z_6 - 2z_7 - 2z_8)$, and $E_8 \subseteq E$. Now any vector of integral norm in \mathbb{R}^7 (resp. \mathbb{R}^8) which has integral inner product with all vectors of E_7 (resp. E_8) is itself in E_7 (resp. E_8); this can be shown in a similar way as we found the norm 2 vectors. Hence $E = E_7$ or $E = E_8$. \square

Proposition 3.1 and Theorem 3.5 are related to the following theorem which is implicitly in Cameron, et al. [2].

THEOREM 3.6 (Cameron, Goethals, Seidel, Shult). *An integral lattice generated by norm 2 vectors is isomorphic to one of the root lattices $A_p = \{x \in \mathbb{Z}^{p+1} | \sum x_i = 0\}$, $D_p = \{x \in \mathbb{Z}^p | \sum x_i \text{ even}\}$, $E_8 = A_8 + \frac{1}{3}\mathbb{Z}(1^6, (-2)^3)$, $E_7 = A_7 + \frac{1}{2}\mathbb{Z}(1^4, -1^4)$, or $E_6 = E_7 \cap (0^6, 1^2)^\perp$.*

Proof. Any two norm 2 vectors x, y ($y \neq \pm x$) have inner product $\in \{0, \pm 1\}$, whence the corresponding set S of lines has angles 60° or 90° . Since $(x, y) = -1$ implies that $z = -x - y$ also has norm 2, S is star-closed in the sense of Cameron et al. [2]. Hence by [2, Thm. 3.5], the norm 2 vectors generate one of the lattices mentioned. \square

For minimal norm 3, we give only a partial result:

THEOREM 3.7. *An exceptional integral lattice of simplex type with minimal norm 3, generated by norm 3 vectors, can exist only in dimensions 6, 7, 14, 16, 22, 23, 25, 30 or 47.*

Proof. We have $n = 3, m = 1, d = 2$, and (**) requires that $\varepsilon(1 - \varepsilon)(p + 2)$ be an even integer ≤ 6 . With ε rational, $0 < \varepsilon < 1$, this leaves the values in Table 8 for p and ε :

TABLE 8

p	6	7	14	16	22	23	25	30	47
ε	$\frac{1}{2}$	$\frac{1}{3}, \frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{3}, \frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$	$\frac{1}{3}, \frac{2}{3}$	$\frac{1}{4}, \frac{3}{4}$	$\frac{1}{7}, \frac{6}{7}$

Remark. Let E be an arbitrary integral lattice of minimal norm 3. If $x, y \in E$ have norm 3, $y \neq \pm x$ then $(x, y)^2 < (x, x)(y, y) = 9$, whence $-2 \leq (x, y) \leq 2$. But if $(x, y) = \pm 2$ then $(x \mp y, x \mp y) = 3 - 2 - 2 + 3 = 2$, a contradiction. Hence $(x, y) \in \{0, \pm 1\}$, and the set of norm 3 vectors of E form a set Σ of vectors of type $\{0, \frac{1}{3}\}$ in the sense of Shult and Yanushka [15]. Moreover, if x, y, z are norm 3 vectors with $(x, y) = (x, z) = (y, z) = -1$ then $w = -x - y - z$ has also norm 3, and $(x, w) = (y, w) = (z, w) = -1$. Hence Σ is tetrahedrally-closed.

This explains the occurrence of some of Shult and Yanushka's line systems in the present context. In fact, at least four of their examples occur: C_{16} ($p = 6$; Example 1), C_{28} (case $p = 7$ of Example 11; projection to w^\perp), S_{256} ($p = 16$; Example 3) and a system of 2300 lines (case $p = 23$ of Example 15; projection to w^\perp). In fact, the last three examples come from exceptional lattices.

Acknowledgments. I am indebted to Prof. J. Seidel who interested me in the relations between combinatorial configurations and hyperbolic spaces. I recall with pleasure the discussions I have had with him on this subject. Further thanks go to John Leech and the referee for useful comments.

REFERENCES

- [1] H. F. BAKER, *A locus with 25920 linear self-transformations*, Cambridge Univ. Press, London, 1946.
- [2] P. J. CAMERON, J. M. GOETHALS, J. J. SEIDEL AND E. E. SHULT, *Line graphs, root systems, and elliptic geometry*, *J. Algebra*, 43 (1976), pp. 305–327.
- [3] J. H. CONWAY, *Three lectures on exceptional groups*, in *Finite Simple Groups*, Academic Press, New York, Chapt. 7, London 1971.
- [4] H. S. M. COXETER, *Integral Cayley numbers*, *Duke Math. J.*, 13 (1946), pp. 561–578.
- [5] ———, *Extreme forms*, *Canad. J. Math.*, 3 (1951), pp. 391–441.
- [6] L. E. DICKSON, Part III of: G. A. Miller, H. F. Blichfeldt and L. E. Dickson, *Theory and Applications of Finite Groups*, Dover reprint, New York, 1961, pp. 351–377.
- [7] J. M. GOETHALS AND J. J. SEIDEL, *Strongly regular graphs derived from combinatorial designs*, *Canad. J. Math.*, 22 (1970), pp. 597–614.
- [8] P. DELSARTE, J. M. GOETHALS AND J. J. SEIDEL, *Spherical codes and designs*, *Geom. Dedicata*, 6 (1977), pp. 363–388.
- [9] J. M. GOETHALS AND J. J. SEIDEL, *Spherical designs*, *Proc. Symp. Pure Math.*, Vol. 34, American Mathematical Society, Providence, RI, 1979, pp. 255–272.
- [10] J. LEECH AND N. J. A. SLOANE, *Sphere packings and error-correcting codes*, *Canad. J. Math.*, 23 (1971), pp. 718–745.
- [11] Y. I. MANIN, *Cubic Forms*, North-Holland, Amsterdam–London, 1974.
- [12] A. NEUMAIER *Combinatorial configurations in terms of distances*, Memorandum 81-09-Wiskunde; Techn., Univ. Eindhoven, the Netherlands, 1981.
- [13] J. J. SEIDEL, *Strongly regular graphs with $(-1, 1, 0)$ —adjacency matrix having eigenvalue 3*, *Linear Alg. Appl.*, 1 (1968), pp. 281–298.
- [14] J.-P. SERRE, *A Course in Arithmetic*, Springer, New York–Heidelberg–Berlin, 1973.
- [15] E. SHULT AND A. YANUSHKA, *Near n -gons and line systems*, *Geom. Dedicata*, 9 (1980), pp. 1–72.
- [16] N. J. A. SLOANE, *Binary codes, lattices, and sphere-packings*, in *Combinatorial Surveys*, P. J. Cameron, ed., Academic Press, London, 1977, pp. 117–164.
- [17] D. E. TAYLOR, *Regular twographs*, *Proc. London Math. Soc.*, (3) 35 (1977), pp. 257–274.
- [18] E. B. VINBERG, *On groups of unit elements of certain quadratic forms*, *Math. USSR Sb.*, 16 (1972), pp. 17–35.
- [19] P. DU VAL, *The Kantor group of a set of points*, *Proc. London Math. Soc.*, 42 (1937), pp. 18–51.

GENERATING NONISOMORPHIC MAPS WITHOUT STORING THEM*

T. R. WALSH†

Abstract. In 1964, A. B. Lehman found a code for rooted planar maps which he later generalized to rooted maps of arbitrary orientable genus. By generating all the code words of a given length and keeping only those which are maximal, with respect to a predefined linear order on the set of code words, among all those coding different rootings of the same map, one can generate nonisomorphic maps of a given genus with a given number of edges. The necessary algorithms are described, along with a brief discussion of the generation of planar maps which have certain prescribed properties, such as 2-connectedness.

Introduction. Since Edmonds [6] reduced maps of orientable surfaces to combinatorial objects, it has been possible to store maps in a computer and to decide quickly whether two maps are isomorphic [9], and thus to generate by computer exactly one representative from each isomorphism class of maps satisfying a given set of conditions (such as planarity or 3-connectedness). If one can somehow produce *at least* one representative from each isomorphism class of such maps, one can store *exactly* one representative from each class (if they fit in the memory!) by storing each new map as it is generated if and only if it is not isomorphic to any of the maps already stored. Tutte described such a scheme for generating 3-connected planar maps [20, p. 454], and Duijvestijn thus generated all the nonisomorphic 3-connected planar maps with up to 22 edges in order to investigate squared squares [5]. A similar scheme for generating 0-, 1-, 2- and 3- connected planar graphs appears in [12].

Here we present a method for generating nonisomorphic maps without having to store them. We apply to maps an idea used by Faradzhev [7], and independently by Read [17], to generate nonisomorphic graphs. Like Heap [8], who generated the 8-vertex graphs, and Baker *et al.* [2], who generated the 9-vertex graphs, Faradzhev and Read defined a linear order on the set of $n \times n$ 0-1 matrices, which code vertex-labelled graphs, and chose the largest matrix from among those which code the same graphs but with different labellings—the *canonical* matrix for that graph. But whereas Heap and Baker converted each matrix into the canonical matrix for the same graph and then compared it with the canonical matrices already stored, Faradzhev and Read tested each matrix to see if it were *already* canonical, and thus avoided storing any matrices. We apply that idea of “testing for canonicity” to A. B. Lehman’s code for rooted maps [13]: we generate all the code words which satisfy a given set of conditions and throw away all those words which can be made “bigger” by coding the same map but with a different rooting. In this way we have generated nonisomorphic maps with a given number of vertices, edges and faces, and also several sets of nonisomorphic planar maps, including 2-connected maps and 1- and 2-connected plane graphs.

The first section of this article describes Lehman’s code and some of its properties and presents an $O(m^2)$ algorithm which uses this code to construct the multiplication table for the automorphism group of an m -edge map of arbitrary orientable genus. The basic scheme for generating nonisomorphic maps is presented in § 2 and applied to maps of arbitrary orientable genus in § 3 and to planar maps with certain prescribed properties in § 4. Table 1 gives the number of nonisomorphic maps of genus g with $m \leq 6$ edges and $n \leq m + 1$ vertices. Table 3 gives the types of planar maps whose

* Received by the editors June 20, 1980, and in revised form June 1, 1982.

† Department of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B9.

generation we have programmed and Table 4 gives the number of nonisomorphic maps of each type for small m . Two definitions of isomorphism are considered: one in which an isomorphism is assumed to preserve the orientation of the imbedding surface, and one without this restriction. The computations were done on the BESM-6 computer in the Computing Centre of the U.S.S.R. Academy of Sciences in Moscow, which executes 10^6 operations/second. The average time taken to do the computations described here are given in Tables 2 and 3.

1. Lehman's code for rooted maps. Edmonds' theorem can be worded as follows:

PROPOSITION 1 [6]. *For every connected graph G (loops and multiple edges allowed), and for every cyclic order of the edge-ends incident on each vertex, there is a topologically unique 2-cell imbedding of G in an oriented surface such that the clockwise order of the edge-ends around each vertex are as specified.*

A. Jacques ([11], [4, p. 13]) and Lehman ([13], [24, p. 193]) independently defined a map as an object which is essentially a connected graph with a cyclic order defined on the edge-ends at each vertex, and the definition we give is a hybrid of their definitions. Let X be a finite set of *darts*, let σ be a permutation on X , and let α be a fixed-point-free involution on X . This defines a graph whose *edges* are the cycles of α and whose *vertices* are the cycles of σ . An edge e and a vertex v are *incident* if they have at least one common dart d , so that d can be regarded as an end of the edge e incident to v , and σ imposes a cyclic order on the edge-ends incident to v . If the group generated by σ and α is transitive on X , so that the graph is connected, then the triplet (X, σ, α) is called a *map*. In agreement (up to conjugacy) with Edmonds's scheme, the *dual* of (X, σ, α) is the map $(X, \sigma\alpha, \alpha)$ (where permutations are multiplied from right to left), and so the *faces* of (X, σ, α) are the cycles of the permutation $\sigma\alpha$. If a map has n vertices, m edges and f faces, then its *genus* g is defined by the Euler-Poincaré formula

$$(1) \quad n - m + f = 2(1 - g).$$

A *planar* map is a map of genus 0.

An *isomorphism* from (X, σ, α) onto (X', σ', α') is essentially a graph isomorphism which preserves cyclic order. More precisely, it is a 1-1 correspondence $\phi: X \rightarrow X'$ such that for all d in X , $\phi(\sigma(d)) = \sigma'(\phi(d))$ and $\phi(\alpha(d)) = \alpha'(\phi(d))$, so that an *automorphism* of (X, σ, α) is a permutation on X which commutes with σ and α . An isomorphism is the combinatorial equivalent of an orientation-preserving homeomorphism between two maps on oriented surfaces. The equivalent of an orientation-reversing homeomorphism is a *reflection*: an isomorphism from (X, σ, α) onto $(X', \sigma'^{-1}, \alpha')$.

A *rooted map* (X, σ, α, d) is a map (X, σ, α) with a distinguished dart d , its *root*. It is easy to prove either topologically [21, p. 252], [3, p. 16] or combinatorially [24, p. 208]:

PROPOSITION 2. *Every automorphism of a map except the trivial automorphism is fixed-point-free.*

So if an isomorphism from a rooted map (X, σ, α, d) onto another one $(X', \sigma', \alpha', d')$ is defined as an isomorphism ϕ from (X, σ, α) onto (X', σ', α') such that $\phi(d) = d'$, then a rooted map has only the trivial automorphism. This makes it possible to code nonisomorphic rooted maps without considering their symmetries. Lehman [13] and R. Cori [4] each found codes for rooted planar maps. Either code could be used to generate planar maps (indeed, some of the same results have been obtained with both codes, such as the enumeration of various classes of planar

maps—see [4], [25] and [26]) and Cori’s code has been published, but we used Lehman’s code because it is simpler and has been generalized to nonplanar maps. Lehman’s code was described in [4, p. 83], and his generalization to rooted maps of arbitrary genus appears in [13] and [22, p. 92], but since [4] is in French and neither [13] (notes for a graduate course) nor [22] (a Ph.D. thesis) has been published, we summarize the results here.

Lehman’s code is a generalization of the classical code for rooted plane trees in terms of parenthesis systems. A *parenthesis system on p pairs* is a word $p = s_1, s_2, \dots, s_{2p}$ consisting of p left parentheses and p right parentheses such that the function $E(i)$, defined as the number of left parentheses minus the number of right parentheses among the first i symbols, is nonnegative for all $i = 1, 2, \dots, 2p$. The *mate* to a left parenthesis s_i is the right parenthesis s_j , where j is the smallest integer greater than i such that $E(j) = E(i) - 1$. The mate to s_j is s_i , and s_i and s_j together constitute a *parenthesis pair*. Define the rooted plane tree $l_r(P) = (X, \sigma, \alpha, d_1)$ with p edges as follows:

$$\begin{aligned}
 X &= \{d_1, d_2, \dots, d_{2p}\}, \quad \text{and for } i = 1, 2, \dots, 2p, \\
 \alpha(d_i) &= d_j, \quad \text{where } s_j \text{ is the mate to } s_i, \quad \text{and} \\
 \sigma\alpha(d_i) &= d_{i+1}, \quad \text{where } d_{2p+1} \text{ means } d_1.
 \end{aligned}
 \tag{2}$$

The parenthesis system P is the usual code for $l_r(P)$.

An *integer system on c pairs* is a word $I = s_1, s_2, \dots, s_{2c}$ consisting of 2 copies each of the integers $1, 2, \dots, c$ such that the first occurrences come in increasing order: if $s_i = s_j, i < j, s_k = s_h, k < h$, and $s_i < s_k$, then $i < k$. The *mate* to each symbol is the other copy of the same integer. Define the rooted map $l_r(I) = (X, \sigma, \alpha, d_1)$ with c edges and 1 vertex as follows:

$$\begin{aligned}
 X &= \{d_1, d_2, \dots, d_{2c}\}, \quad \text{and for all } i = 1, 2, \dots, 2c, \\
 \alpha(d_i) &= d_j, \quad \text{where } s_j \text{ is the mate to } s_i, \quad \text{and} \\
 \sigma(d_i) &= d_{i+1}, \quad \text{where } d_{2c+1} \text{ means } d_1.
 \end{aligned}
 \tag{3}$$

The genus of (X, σ, α) is given by (1), where $n = 1, m = c$, and f is the number of cycles of $\sigma\alpha$, and the genus of I is defined to be the genus of $l_r(I)$.

A *parenthesis-integer system on m pairs* is a word $S = s_1, \dots, s_{2m}$ constructed by merging a parenthesis system P on p pairs with an integer system I on $c = m - p$ pairs. Define $l_r(S)$ to be the rooted map $(M, d_1) = (X, \sigma, \alpha, d_1)$ with m edges and $p + 1$ vertices, where

$$\begin{aligned}
 X &= \{d_1, d_2, \dots, d_{2m}\}, \quad \text{and for all } i = 1, 2, \dots, 2m, \\
 \alpha(d_i) &= d_j, \quad \text{where } s_j \text{ is the mate to } s_i, \quad \text{and} \\
 d_{i+1}(d_{2m+1} \equiv d_1) &= \begin{cases} \sigma(d_i) & \text{if } s_i \text{ is an integer,} \\ \sigma\alpha(d_i) & \text{if } s_i \text{ is a parenthesis,} \end{cases}
 \end{aligned}
 \tag{4}$$

and let $T(S)$ be the spanning tree of M whose edges are the cycles (d_i, d_j) of α such that (s_i, s_j) is a parenthesis pair. Then $l_r(P) = (T(S), d_i)$, where s_i is the first parenthesis in S , and $l_r(I)$ is the rooted 1-vertex map $(M/T(S), d_1)$ formed by contracting the edges of $T(S)$ and shifting the root to d_j , where s_j is the first integer in S . The *genus* of S is defined to be the genus of the (rooted) map $l_r(S)$. It is easy to show that S and I have the same genus: removing a parenthesis pair (s_i, s_j) from S merely deletes the

darts d_i and d_j from the cycle(s) of $\sigma\alpha$ containing them, so that M and $M/T(S)$ have the same number of faces and thus the same genus.

Lehman has shown that for every rooted map (M, d_1) and for every spanning tree T of M there exists a unique parenthesis-integer system S such that $l_r(S)$ is isomorphic to (M, d_1) and $T = T(S)$ ([13], [22, p. 93]). To find a unique system $S = l_r^{-1}(M, d_1)$ which codes (M, d_1) (with no spanning tree) it is necessary to choose a *canonical spanning tree*. He chose the spanning tree $T(M, d_1)$ constructed by executing a depth-first search [18, p. 147] of M starting with the vertex containing d_1 , with the darts in each vertex explored in an order determined by σ and by the dart of entry to the vertex. The system $S = l_r^{-1}(X, \sigma, \alpha, d_1)$, which we call the *Lehman code* for (X, σ, α, d_1) , is constructed during the search, one symbol for each new dart reached. His coding algorithm, a modification of the Trémaux–Tarry maze algorithm [19], is given as Algorithm 1 in Fig. 1. When $l_r(S)$ is constructed according to (4), the darts are renumbered d_1, d_2, \dots, d_{2m} in the order in which Algorithm 1 reaches them, and this renumbering defines the isomorphism taking (M, d_1) onto $l_r(S)$.

For example, suppose

$$X = \{1, 2, \dots, 12\}, \quad \sigma = (1, 2, 3)(4, 5, 6)(7, 8, 9)(10, 11, 12)$$

ALGORITHM 1. Coding the rooted map (X, σ, α, r) with m edges

BEGIN {Algorithm 1}

$d \leftarrow r$; { d is the current dart}

mark the vertex containing d ; {it has now been reached}

$k \leftarrow 0$; { k counts the fronds}

FOR $i \leftarrow 1$ to $2*m$ DO BEGIN {for i }

IF the edge e containing d is unmarked THEN BEGIN {outer then} { e has not yet been explored}

IF the vertex containing $\alpha(d)$ is marked THEN BEGIN {inner then}

{the vertex at the other end of e has already been reached}

$k \leftarrow k + 1$;

mark e with k ; {this excludes e from the canonical tree}

$s_i \leftarrow k$;

$d \leftarrow \sigma(d)$

END {inner then}

ELSE BEGIN {inner else}

{the vertex at the other end of e has not yet been reached}

mark e with -1 (or a heavy line); {to include e in the tree}

$s_i \leftarrow$ left parenthesis;

$d \leftarrow \sigma\alpha(d)$;

mark the vertex containing d {it has now been reached}

END {inner else}

END {outer then}

ELSE IF e is marked with a positive integer j THEN BEGIN {else if}

{ e has already been excluded from the tree and numbered j }

$s_i \leftarrow j$;

$d \leftarrow \sigma(d)$

END {else if}

ELSE BEGIN {outer else}

{ e has already been included in the tree}

$s_i \leftarrow$ right parenthesis;

$d \leftarrow \sigma\alpha(d)$

END {outer else}

END {for i }

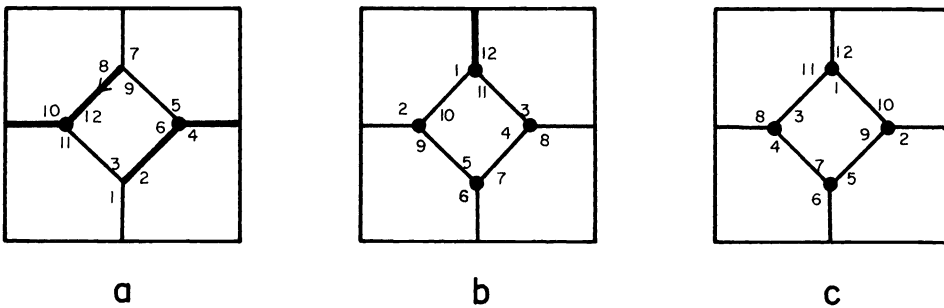
END. {Algorithm 1}

FIG. 1

and

$$\alpha = (1, 7)(2, 6)(3, 11)(4, 10)(5, 9)(8, 12).$$

Then $\sigma\alpha = (1, 8, 10, 5, 7, 2, 4, 11)(3, 12, 9, 6)$; so $n = 4$, $m = 6$, $f = 2$ and $g = 1$. If dart 8 is the root, then $T = \{(8, 12), (10, 4), (6, 2)\}$ and $S = ((1(23))2)13$. A drawing of this map, with its root represented by an arrow and its tree-edges drawn in heavy lines, and with the torus represented by a square whose opposite sides are identified in pairs, is shown in Fig. 2a. In this drawing σ is represented by counterclockwise rotation. The darts are reached in the order (8, 10, 5, 6, 3, 1, 2, 4, 11, 12, 9, 7). Renumbering the darts so that dart 8 is called 1, dart 10 is called 2, etc., yields $l_r(S)$ (see Fig. 2b).



$$S = ((1(23))2)13$$

FIG. 2

We summarize the foregoing in

PROPOSITION 3. *Two rooted maps (X, σ, α, d_1) and $(X', \sigma', \alpha', d'_1)$ are isomorphic if and only if they have the same Lehman code. If they are isomorphic, then the unique isomorphism $\phi: X \rightarrow X'$ takes d_i onto d'_i , where the bits of X and X' are numbered in the order in which they are reached during the execution of Algorithm 1 or, equivalently, according to formula (4).*

A parenthesis-integer system S is called *correct* if it has no subword $i(i)$, whose integers form a pair and whose parentheses form a pair, but whose symbols are not necessarily adjacent in S . A proof of the following theorem, also due to Lehman, can be found in [22, p. 98].

PROPOSITION 4. *A parenthesis-integer system is constructed by Algorithm 1 as the code for some rooted map if and only if it is correct.*

In terms of the canonical spanning tree T , this means that if a *frond* e —that is, an edge not in T —joins two distinct vertices u and v , then one of these, say v , lies on the path in T from u to the root-vertex of T (this was proved in [18, p. 148] for a graph with a depth-first-search spanning tree), and that if d_1, d_2 , and d_3 are the darts in v such that d_3 is in e and d_1 and d_2 are on the path, with d_1 on the edge closer to the root-vertex (if v is the root-vertex, make d_1 the root of T) then σ restricted to these three darts is (d_1, d_2, d_3) .

By Propositions 3 and 4, l_r is a 1-1 correspondence from $L(g, p, c)$ onto $K_r(g, p, c)$, where $L(g, p, c)$ is the set of correct parenthesis-integer systems of genus g with p parenthesis pairs and c integer pairs, and $K_r(g, p, c)$ is the set of (isomorphism classes of) rooted genus- g maps with $p + 1$ vertices and $p + c$ edges.

A rooted planar 1-vertex map is the dual of a rooted plane tree, and so can be coded by a parenthesis system, composed of square brackets to distinguish it from the code for a rooted plane tree. The previous discussion can thus be specialized to rooted planar maps. Rooted planar maps with distinguished spanning trees are coded by *parenthesis-bracket systems*, constructed by merging two parenthesis systems, one made of parentheses and the other of brackets, and rooted planar maps are coded by *correct parenthesis-bracket systems* in which there are no subwords $[(\])$ composed of two pairs. Algorithm 1 can be adapted to code rooted planar maps by initializing k to 1 and keeping it fixed (deleting the line “ $k \leftarrow k + 1$ ”), and changing the line “ $s_i \leftarrow k$ ” to “ $s_i \leftarrow$ left bracket” and the line “ $s_i \leftarrow j$ ” to “ $s_i \leftarrow$ right bracket”.

It should now be easier to read [25] and [26], where this code was used to count rooted maps by genus.

Proposition 3 leads to the following test for isomorphism between two maps.

PROPOSITION 5. *Given two maps (X, σ, α) and (X', σ', α') , fix one dart d in X . Then (X, σ, α) is isomorphic to (X', σ', α') if and only if there exists a dart d' in X' such that the rooted maps (X, σ, α, d) and $(X', \sigma', \alpha', d')$ have the same Lehman code.*

Since Algorithm 1 can be executed in $O(m)$ operations, this test can be done in $O(m^2)$. Better algorithms have been found for planar maps [9] since Lehman obtained his code, but for nonplanar maps this is the only isomorphism test of which we are aware.

Similarly, the number of automorphisms of the map (X, σ, α) is equal to the number of darts d' in X for which $l_r^{-1}(X, \sigma, \alpha, d') = l_r^{-1}(X, \sigma, \alpha, d)$ for a fixed d in X . The automorphisms of a map can be counted in $O(m^2)$, and without increasing the complexity by more than a multiplicative constant, one can construct the multiplication table for its automorphism group. Given a map $M = (X, \sigma, \alpha)$, fix a dart d_1 in X , find $S = s_1, s_2, \dots, s_{2m} = l_r^{-1}(X, \sigma, \alpha, d_1)$ using Algorithm 1, and order the darts of X accordingly. For every dart d_i such that $l_r^{-1}(X, \sigma, \alpha, d_i) = S$, let $a_{i,j}$ be the number k such that d_k is the j th dart reached during the coding of (X, σ, α, d_i) . Then by Proposition 3, the automorphism ϕ_i taking d_1 into d_i takes d_j into d_k . So if an automorphism ϕ_j taking d_1 into d_j exists, then $\phi_i \phi_j(d_1) = \phi_i(d_j) = d_k$ and so, by Proposition 2, $\phi_i \phi_j = \phi_k$. The matrix $(a_{i,j})$, restricted to those columns j for which an automorphism ϕ_j exists, is a multiplication table for the automorphism group of M .

In our example, S has already been found and the darts renumbered as in Fig. 2b. Continuing the above procedure, we find that

$$(5) \quad (a_{i,j}) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 3 & 12 & 7 & 5 & 10 & 2 & 9 & 6 & 1 & 11 & 4 & 8 \\ 7 & 8 & 9 & 10 & 11 & 12 & 1 & 2 & 3 & 4 & 5 & 6 \\ 9 & 6 & 1 & 11 & 4 & 8 & 3 & 12 & 7 & 5 & 10 & 2 \end{pmatrix}.$$

Restricting $(a_{i,j})$ to the columns beginning with 1, 3, 7 and 9 we get the required multiplication table.

To see if (X, σ, α) has a reflection onto itself, we first construct (X, σ^{-1}, α) (Fig. 2c) and apply Proposition 5. In our example, $l_r^{-1}(X, \sigma^{-1}, \alpha, 4) = S$ so there is a reflection ϕ_{-4} which takes dart 1 into dart 4, and in the process of coding, the darts are reached in the order

$$(6) \quad (4, 6, 11, 1, 9, 2, 10, 12, 5, 7, 3, 8).$$

We can now write the multiplication table for the group of automorphisms and

reflections:

$$(7) \begin{matrix} 1 & 3 & 7 & 9 & -4 & -11 & -10 & -5 \\ 3 & 7 & 9 & 1 & -5 & -4 & -11 & -10 \\ 7 & 9 & 1 & 3 & -10 & -5 & -4 & -11 \\ 9 & 1 & 3 & 7 & -11 & -10 & -5 & -4 \\ -4 & -11 & -10 & -5 & 1 & 3 & 7 & 9 \\ -11 & -10 & -5 & -4 & 9 & 1 & 3 & 7 \\ -10 & -5 & -4 & -11 & 7 & 9 & 1 & 3 \\ -5 & -4 & -11 & -10 & 3 & 7 & 9 & 1 \end{matrix}$$

where reflections are represented by negative numbers to avoid possible duplication of entries. The numbers 4, 11, 10 and 5 in the first row of (7) are the 1st, 3rd, 7th and 9th entries of (6), since they represent the reflections $\phi_{-4}\phi_1$, $\phi_{-4}\phi_3$, $\phi_{-4}\phi_7$ and $\phi_{-4}\phi_9$. The first five rows of (7) are (5) and (6) restricted to columns 1, 3, 7, 9, 4, 11, 10, 5, with appropriate sign changes. The remaining rows can be filled in using the associative law; for example $a_{-11,7} = \phi_{-11}\phi_7(1) = (\phi_{-4}\phi_3)\phi_7(1) = \phi_{-4}(\phi_3\phi_7)(1) = \phi_{-4}\phi_9(1) = \phi_{-4}(9) = -5$.

2. Generating nonisomorphic maps. We present some algorithms for generating $K(g, p, c)$, the set of nonisomorphic maps (or, more precisely, a set of nonisomorphic maps which includes one representative from each isomorphism class of maps) of genus g with $p + 1$ vertices and $p + c$ edges.

Let $M = (X, \sigma, \alpha)$ be a map in $K(g, p, c)$ and consider the set of rooted maps (M, d) , $d \in X$, and their Lehman codes $l_r^{-1}(M, d)$. In general, M will give rise to several code words, one for each isomorphism class of rooted maps (M, d) . To get a unique code word for M , we impose a linear order (defined later) on the set $L(g, p, c)$ (of correct genus- g parenthesis-integer systems on p parenthesis pairs and c integer pairs) and define the *canonical code* $l^{-1}(M)$ for M to be the largest code word derived from M by choosing some root d and coding the rooted map (M, d) .

Now suppose we generate the whole set $L(g, p, c)$ exactly once (an algorithm which does this will be presented in § 3). We will generate every code word to which M gives rise, and in particular, we are sure to generate $l^{-1}(M)$, so we can safely throw away all the other code words for M and keep only $l^{-1}(M)$, and in this way we generate M exactly once.

Having generated a system S , we decode it to get the rooted map $lr(S) = (M, d)$ which S codes, and then we test to see if S is the canonical code for $M = (X, \sigma, \alpha)$. For every dart d' in X , we use Algorithm 1 to construct the system $S' = l_r^{-1}(M, d')$. If ever we construct a code word S' which is larger than S , we know (without trying any more darts) that S is not canonical and we throw S away; if we exhaust every dart d' in X without ever constructing a code word S' larger than S , then we know that S is canonical and we count M .

The linear order we chose for $L(g, p, c)$ is lexicographical, with $) = 0$, $(= 1$, and each integer increased by 1. The set $L_0(p, c)$ of correct parenthesis-bracket systems is also ordered lexicographically, putting $) = 0$, $(= 1$, $] = 2$, $[= 3$. The advantage of a lexicographical order is that it will often be unnecessary to construct the whole system S' in order to decide if S' is greater than S . As soon as the i th symbol s'_i in S' is found, it is compared with the i th symbol s_i in S . If $s'_i > s_i$, then $S' > S$ and so S is not canonical. If $s'_i < s_i$, then $S' < S$, and again the construction of S' can be stopped. Only if $s'_i = s_i$ do we need to construct the next symbol of S' unless $i = 2m$, and in this case

$S' = S$ and we have found an automorphism of M . If S is canonical, we find all the automorphisms of M .

We leave it to the reader to write a description of the above nonisomorphic map-generation algorithms in GOTO-free pseudo-Pascal. Suffice it to say that the worst-case estimate for the number of comparisons required to test a word with $2m$ symbols for canonicity is $4m^2$, assuming that every dart d' is tried and that every symbol of S' is constructed. Then since every map with $2m$ edges will give rise to at most $2m$ code words, its generation requires at most $8m^3$ comparisons, and in general $O(m^3)$ operations since testing systems for canonicity is much slower than generating them. The following average-case time estimate seems to correspond more closely than the pessimistic worst-case estimates with the time-trials we conducted. Assume that M has only the trivial automorphism—this is true for almost all maps—so that as the darts d_1, d_2, \dots, d_{2m} are used as roots, the corresponding code words S_1, S_2, \dots, S_{2m} are all distinct. Assume also that every linear order of these code words is equally likely, and that we stop at S_i if $S_i > S_1$. It is easy to show that the expected value of i is not $2m$ but $\ln(m) + O(1)$, so that the average time to test S is $O(m \ln m)$ and the average time to generate M is $O(m^2 \ln m)$.

For planar maps the partitioning algorithm of [9, p. 329] can be modified to a canonicity test for a rooted planar map (X, σ, α, d) . If d is not placed in the first block $B(1)$ according to the degrees of the vertices and the faces containing d and $\alpha(d)$, or if d is ever moved out of $B(1)$, then (X, σ, α, d) is not canonical; if the algorithm terminates with d still in $B(1)$, then (X, σ, α, d) is defined to be canonical. Care must be taken that the order in which the darts are moved and the blocks into which they are put does not depend on their names. So in line H all the darts $d = f(e, D)$, $e \in B(i)$, are placed in MOVE in such a way that if $d_1 \in B(j_1)$ and $d_2 \in B(j_2)$ with $j_1 < j_2$, then d_1 comes before d_2 in MOVE (this can be arranged using a bucket sort [1, p. 77]). Since all the darts in a given block will be moved in a body, their order in MOVE is irrelevant. If the algorithm terminates and declares (X, σ, α, d) to be canonical, it still runs in $O(m \log m)$; if not, it finishes “prematurely”, so the average-case time estimate may be less. However, see Table 2.

Suppose we want to generate nonisomorphic maps no two of which are related by a reflection. Once a word S has been generated and found to be the canonical code for $M = (X, \sigma, \alpha)$, we construct $M^{-1} = (X, \sigma^{-1}, \alpha)$, the reflection of M , and test every dart d' in X to see if $l_r^{-1}(M^{-1}, d') \geq S$. If a dart d' is found such that $l_r^{-1}(M^{-1}, d') > S$, then S can be thrown away, since $l^{-1}(M^{-1})$ will at some time be generated. If a dart d' is found such that $l_r^{-1}(M^{-1}, d') = S$, then M^{-1} is isomorphic to M , and so M is counted. If $l_r^{-1}(M^{-1}, d') < S$ for every d' in X , then M is counted, and if we want to count all nonisomorphic maps as well, then we count M^{-1} as well as M .

The number of nonisomorphic planar maps with $m \leq 6$ edges and the number “up to reflection” (with two maps considered equivalent if they are related by a reflection) generated by this method appear in Table 1. These numbers agree with Liskovets’ formula [15] and Wormald’s table [27, p. 34], respectively, and the maps of genus 1 and 2 with ≤ 4 edges agree with those drawn by hand in [22, p. 204–210] (there was one disagreement but the computer turned out to be right). So the problem of generating $K(g, p, c)$ can be considered to be reduced to that of generating $L(g, p, c)$. This will be taken up in § 3, and the generation of subsets of $L_0(p, c)$ is the topic of § 4.

3. Generating rooted maps of a given genus. The set $L_0(p, 0)$ of parenthesis systems on p pairs is easily generated with the help of the function $E(i)$, defined earlier. The first system is $((\dots((\dots)))\dots))$; for this system $E(i) = \min(i, 2p - i)$. Given

TABLE 1

The number # of nonisomorphic maps of genus g with m edges and n vertices, and the number @ of such maps no two of which are related by a reflection

m	n	$g = 0$		$g = 1$		$g = 2$		$g = 3$	
		#	@	#	@	#	@	#	@
0	1	1	1						
0	sum	1	1						
1	1	1	1						
1	2	1	1						
1	sum	2	2						
2	1	1	1	1	1				
2	2	2	2						
2	3	1	1						
2	sum	4	4	1	1				
3	1	2	2	3	3				
3	2	5	5	3	3				
3	3	5	5						
3	4	2	2						
3	sum	14	14	6	6				
4	1	3	3	11	10	4	4		
4	2	14	13	24	20				
4	3	23	20	11	10				
4	4	14	13						
4	5	3	3						
4	sum	57	52	46	40	4	4		
5	1	6	6	46	35	53	38		
5	2	42	35	180	125	53	38		
5	3	108	83	180	125				
5	4	108	83	46	35				
5	5	42	35						
5	6	6	6						
5	sum	312	248	452	320	106	76		
6	1	14	12	204	132	553	328	131	82
6	2	140	104	1198	728	1276	739		
6	3	501	340	2048	1226	553	328		
6	4	761	504	1198	728				
6	5	501	340	204	132				
6	6	140	104						
6	7	14	12						
6	sum	2071	1416	4852	2946	2382	1395	131	82

a system $P = s_1, s_2, \dots, s_{2p}$ in $L_0(p, 0)$, we construct the next system as follows. Let i be the largest number such that $s_i = '('$ and such that $E(i) > 1$; if no such i exists, then P is the last system $()() \dots ()$ and we stop. Change s_i to ')', reducing $E(i)$ by 2. Now complete the string with $((\dots (()) \dots))$, updating the function E accordingly; the index k of the rightmost left parenthesis is defined by $E(k) = 2p - k$. Although a few of the systems may take as many as p operations to generate, it can be shown that the average number of operations is bounded above by a number independent of p .

TABLE 2

The average time in milliseconds on the BESM-6 to do certain computations on a map

Computation	Time
Testing a 6-edge rooted planar map for canonicity	
-with the Lehman code	5.0
-with the $V \log V$ isomorphism test	12.5
Testing a rooted map for "canonicity up to reflection"	
-7 edges	6.7
-11 edges	11.0
Generating a parenthesis system of arbitrary length	0.33
Generating a parenthesis-integer system of arbitrary length with at least as many integers as parentheses	0.33
Generating an integer system on 6 pairs and finding its genus	1.8
Generating a genus-3 integer system on 6 pairs by eliminating systems on 6 pairs which are not of genus 3	12.7
Generating a genus-3 integer system on 6 pairs from a genus-2 integer system on 4 pairs	3.0

The set $\cup_g L(g, 0, c)$ of integer systems on c pairs can be generated as follows. The first system is $112233 \cdots cc$. Given some system, the next one is found as follows. If the second 1 is not the rightmost symbol, exchange it with its right-hand neighbor and return the resulting system; otherwise move the second 1 back to the second position and try the second 2: if possible exchange the second 2 with its right-hand neighbor and return the resulting system, otherwise move the second 2 back to the fourth position and try the second 3, and so on. If we are starting with some system other than $123 \cdots cc \cdots 321$, it will be possible to exchange the second occurrence of some number $i = 1, 2, \cdots, c - 1$ with its right-hand neighbor and return the next system; otherwise we will convert $123 \cdots cc \cdots 321$ back to $112233 \cdots cc$ and then try to move the second c . So if we stop when we are required to move the second c , we will have generated all the integer systems. The average time taken to generate an integer system is easily shown to be independent of c .

Now suppose we have an integer system or a bracket system $I = s_1, s_2, \cdots, s_{2c}$ and a parenthesis system $P = t_1, t_2, \cdots, t_{2p}$, and we want to merge P with I in all possible ways so as to create a correct parenthesis-integer system (or parenthesis-bracket system). For the first system we put all the parentheses to the left of all the integers. Then the parentheses move to the right in the following increasing order of priority: the first left parenthesis, its mate, the second left parenthesis, its mate, and so on. Algorithm 2 of Fig. 3 takes a given system and produces the next one, if there is one.

Figure 4 shows the 35 systems formed by merging $P = (())$ with $I = 1212$ in the order in which they are produced by successive applications of Algorithm 2. We pick up the action after system $15 = (1(0)2)2$ has been produced. The highest priority parenthesis is the third. It can be moved, so we get system $16 = (1(2)1)2$. Now the third parenthesis cannot be moved without creating the forbidden subword $1(1)$; so it is deleted to leave $(1(2)1)2$ and the next highest priority parenthesis—the second—is tried. Since its mate is now gone, it can be moved until it hits the fourth parenthesis. It is moved one step to yield $(12(1)2)$. Then the fourth parenthesis is inserted in the leftmost spot consistent with the order of P to yield system 17: $(12(1)1)2$. Again, the third parenthesis cannot be moved without creating the forbidden subword $1(1)$; so it is deleted, the second parenthesis moved, and the third parenthesis replaced again

ALGORITHM 2. Given a correct parenthesis-integer system S whose integer subsystem is $I = s_1, s_2, \dots, s_{2c}$ and whose parenthesis subsystem is $P = t_1, t_2, \dots, t_{2p}$, $p > 0$, change S to the next correct parenthesis-integer system. PR is an array of dimension $2p$ giving the increasing order of priority in which the parentheses are to be moved: $PR(2i - 1)$ and $PR(2i)$ are the indices in P of the i th left parenthesis and its mate. IPR points to PR .

```

BEGIN {Algorithm 2}
NEXT ← FALSE; {NEXT will become true when the next system has been found}
IPR ← 2*p + 1;
REPEAT
    IPR ← IPR - 1 {back-track down one level of priority}
    i ← PR(IPR); {get the index of the parenthesis to be moved}
    IF there is a symbol  $s_k$  in  $I$  to the immediate right of  $t_i$ , and if the mate to  $t_i$  is not to the right of the mate to  $s_k$  in  $S$ 
    {so that  $t_i$  can be moved without creating a forbidden subword}
    THEN BEGIN {then}
        exchange  $s_k$  and  $t_i$ ;
        WHILE IPR < 2*p DO BEGIN {while}
            {replace all the deleted symbols of  $P$  as far left as possible in  $S$ }
            IPR ← IPR + 1; {advance up one level of priority}
            i ← PR(IPR); {get the index of the parenthesis to be replaced}
            insert  $t_i$  immediately to the right of the nearest parenthesis to the left of  $t_i$  in  $P$  which is currently in  $S$ 
        END; {while}
        NEXT ← TRUE { $S$  is now the next system}
    END {then}
    ELSE delete  $t_i$  from  $S$  {since it cannot be moved}
UNTIL NEXT OR (IPR ≤ 1); {otherwise backtrack}
IF IPR ≤ 1 THEN WRITE ('ALL DONE')
END. {Algorithm 2}
    
```

FIG. 3

- | | | | |
|---------------|---------------|---------------|---------------|
| 1: (())1212 | 11: ((0)121)2 | 19: ((0)1212) | 29: 1((0)2)12 |
| 2: ((0)1)212 | 12: ((1)2)12 | 20: ((1)2)12 | 30: 1((0)2)12 |
| 3: ((1))212 | 13: ((12)1)2 | 21: ((12)12) | 31: 1((2))12 |
| 4: (1(0))212 | 14: ((121))2 | 22: ((121)2) | 32: 1(2(0))12 |
| 5: (0(12))12 | 15: (1(02)1)2 | 23: ((1212)) | 33: 12((0))12 |
| 6: ((12)1)2 | 16: (1(2)1)2 | 24: (1(02)12) | 34: 121((0))2 |
| 7: ((12))12 | 17: (12(0)1)2 | 25: (1(2)12) | 35: 1212((0)) |
| 8: (1(02))12 | 18: (121(0))2 | 26: (12(0)12) | ALL DONE |
| 9: (1(2))12 | | 27: (121(0)2) | |
| 10: (12(0))12 | | 28: (1212(0)) | |

FIG. 4. The 35 correct parenthesis-integer systems formed by merging $P = (())$, with $I = 1212$. The array PR is (1, 4, 2, 3).

to yield system 18: (121(0))2. Now parenthesis 3 cannot be moved because parenthesis 4 is in the way; similarly parenthesis 2 cannot be moved. So both are deleted and parenthesis 4 is moved: (1212). Parentheses 2 and 3 are now placed as far left as possible to yield system 19 = ((0)1212). Now parenthesis 3 can be moved until it hits parenthesis 4 to yield systems 20 through 23 in quick succession. The reader is invited to follow this algorithm until system 35 = 1212((0)) has been produced. Then all the parentheses are deleted and the message "ALL DONE" is written.

To generate $L_0(p, c)$ where $p > 0$ and $c > 0$, we generate all the bracket systems B on c pairs, and for each of these we generate all the parenthesis systems P on p pairs, and for each of these we insert P into B in all correct ways, using Algorithm 2.

To generate $\cup_g L(g, p, c)$, we generate all the integer systems I on c pairs and sort them by genus, and for each of these we generate all the parenthesis systems P on p pairs, and for each of these we insert P into I in all correct ways.

Now suppose we want to generate $L(g, p, c)$ for a fixed $g > 0$. We could generate all the integer systems on c pairs and then eliminate all those not of genus g . But if a very small fraction of them are of genus g it is more economical to proceed as follows. We find all the integer systems I_0 in $L(g, 0, 2g)$ coding rooted genus- g maps with 1 vertex, 1 face and $2g$ edges. For each of these, we generate all the parenthesis systems P_0 on $c-g$ pairs. Each correct insertion of P_0 into I_0 codes a rooted genus- g map (X, σ, α, d) with 1 face and c edges. We then find the dual $(X, \sigma\alpha, \alpha, d)$ of this rooted map, which has 1 vertex and c edges and is also of genus g , and we use Algorithm 1 to find the integer system I which codes it. We then proceed as before, inserting parenthesis systems on p pairs into all such integer systems I .

To generate $L(g, 0, 2g)$ we could generate all the integer systems on $2g$ pairs and then eliminate those not of genus g . It is shown [24, p.213] that of the $(4g)!/2^{2g}(2g)!$ rooted maps with $2g$ edges and 1 vertex, just $(4g)!/2^{2g}(2g+1)!$ —that is, 1 out of $2g+1$ —also have only 1 face and are therefore of genus g , and a natural $(2g+1)$ -to-1 correspondence was later found by Lehman [14]. Since it takes $O(m)$ operations to find the number of faces in an m -edge map, for each integer system in $L(g, 0, 2g)$ we waste $O(g^2)$ operations in generating $2g$ other systems on $2g$ pairs and discovering that they are not of genus g . It is more economical to proceed as follows.

Lehman has shown (see [13] and [22, p. 71]) that the genus of the integer system $I = aibjcidje$, where a, b, c, d, e are segments of I , is one greater than the genus of the system $I_0 = adcbe$. Of course I may have many such representations, so we seek a canonical one and then produce all the genus- $(g+1)$ systems $aibjcidje$ exactly once by dividing each of the genus- g systems canonically into 5 parts as $adcbe$. One such choice is to let the first i of $aibjcidje$ be the leftmost integer forming a subword $ijij$ with some j and the first j of $aibjcidje$ be the leftmost integer forming a word $ijij$ with this i . It turns out that we can set $i = 1$:

PROPOSITION 6. *For any system $I = s_1, s_2, \dots, s_{2c}$ coding a rooted 1-vertex 1-face map, there exists a subword $1j1j$ such that I can be represented as $1bjc1dje$.*

Proof. We identify d_i with s_i in (3), so that $\sigma(s_i) = s_{i+1}$, $\sigma(s_{2c}) = s_1$, and $\alpha(s_i)$ is the mate to s_i . Then the permutation $\sigma\alpha$ is cyclic on the symbols of I . Clearly the second 1 cannot be adjacent to the first 1 (otherwise the cycle of $\sigma\alpha$ containing the second 1 would be of length 1) or at the end of I (otherwise the cycle of $\sigma\alpha$ containing the first 1 would be of length 1). So there must be at least one symbol between the 1s and at least one symbol after the second 1. Starting from any symbol between the 1s and applying $\sigma\alpha$ often enough we must be able to get to some symbol after the second 1, because $\sigma\alpha$ is cyclic, so there exists a symbol s_i between the 1s such that $\sigma\alpha(s_i) = j$, say, comes after the second 1. The other copy of j is $\sigma(s_i)$, but since s_i is between the 1s and $j \neq 1$, $\sigma(s_i)$ must also be between the 1s. So I can be expressed as $1bjc1dje$, Q.E.D. \square

Now if $I = 1bjc1dje$ is a canonical representation, then no symbol in b can have its mate in d or in e . This holds also for $I_0 = dcbe$. So we have the following algorithm for creating all possible systems I from a given I_0 . Three dividing lines H_1, H_2, H_3 are to be placed in the $4g+1$ slots formed by the $4g$ symbols of I_0 . The order and direction in which H_1 and H_3 move are immaterial—we suppose them to move from right to left, with H_3 laid down first and H_1 moving from H_3 to the first slot. For any position of H_1 and H_3 , H_2 begins at H_3 and moves to the left, stopping when it comes to a symbol whose mate is either to the left of H_1 or to the right of H_3 . The divided

system $dH_1cH_2bH_3e$ is changed to $I = g + 1, b, g + 2, c, g + 1, d, g + 2, e$, and if desired the numbers in I are changed to put the first occurrences in ascending order.

For example, in Fig. 5 we show the 21 ways of dividing the one system 1212 in $L(1, 0, 2)$, and the 21 corresponding systems in $L(2, 0, 4)$. Each of these systems can be divided to generate $L(3, 0, 6)$, and so on.

divided I_0	unordered I	ordered I
1: 1212↓↓↓	34312124	12134342
2: 121↓2↓↓	34231214	12314342
3: 12↓12↓↓	34123124	12341342
4: 1↓212↓↓	34212314	12343142
5: 1↓21↓2↓	32421314	12324143
6: ↓1212↓↓	34121234	12343412
7: ↓121↓2↓	32412134	12342413
8: ↓12↓12↓	31241234	12342314
9: ↓1↓212↓	32124134	12324314
10: ↓↓1212↓	31212434	12323414
11: 121↓↓↓2	34312142	12134324
12: 12↓1↓↓2	34131242	12313424
13: 1↓21↓↓2	34213142	12341423
14: ↓121↓↓2	34121342	12343124
15: ↓12↓1↓2	31412342	12324134
16: 12↓↓↓12	34312412	12134234
17: 1↓2↓↓12	34231412	12314243
18: ↓12↓↓12	34123412	12341234
19: 1↓↓↓212	34314212	12132434
20: ↓1↓↓212	34134212	12312434
21: ↓↓↓1212	34341212	12123434

FIG. 5. The 21 genus-2 integer systems I on 4 pairs formed by dividing the 1 genus-1 integer system on 2 pairs $I_0 = 1212$.

4. Generating rooted planar maps with prescribed properties. The *degree* of a face or a vertex of a map is the number of darts it contains. An edge is called a *loop* (an *isthmus*) if both its darts belong to the same vertex (face). A vertex and an edge are called *incident* if they have at least one dart in common. Two or more edges which are not loops are called *parallel edges* if they are all incident with the same pair of

TABLE 3
The types of planar maps generated and the average time in milliseconds taken to generate a rooted map of each type

Code	Type of planar map	Time
1	Without faces of degree 1	0.33
2	Without loops	0.37
3	Without faces of degree 1 or 2	0.43
4	Connected plane graphs (no loops or parallel edges)	1.0
5	Without faces or vertices of degree 1	0.56
6	Without loops or isthmuses	0.94
7	Connected plane graphs without vertices of degree 1	2.1
8	2-connected maps	0.51
9	2-connected maps without faces of degree 2	0.85
10	2-connected plane graphs	1.2

TABLE 4

The # of nonisomorphic planar maps with m edges and n vertices and the number @ of such maps no two of which are related by a reflection. (The codes for the map types are given in Table 3.)

m	n	Code=1 #	Code=2 @ #	Code=3 @ #	Code=4 @ #	Code=5 @ #	Code=6 @ #	Code=7 @ #	Code=8 @ #	Code=9 @ #	Code=10 @ #
1	2	1	1	1	1	1					
1	1	1	1	1	1	1					
2	2	1	1	1	0	1	1		1		
2	3	1	1	1	1	1					
2	sum	2	2	2	1	1	1		1		
3	2	1	1	0	0	1	1	0	1	0	0
3	3	3	2	2	1	1	1	1	1	1	1
3	4	2	2	2	2	2	1	1	1	1	1
3	sum	6	5	4	3	2	2	1	2	1	1
4	2	1	1	0	0	1	1	0	1	0	0
4	3	5	3	0	0	2	2	0	1	0	0
4	4	9	7	6	2	1	1	0	1	0	0
4	5	3	3	3	2	1	1	1	1	1	1
4	sum	18	14	9	5	4	4	1	3	1	1
5	2	1	1	0	0	4	4	0	1	0	0
5	3	8	4	0	0	4	3	0	1	0	0
5	4	31	18	7	1	4	3	1	2	2	1
5	5	28	20	21	8	4	3	1	2	1	1
5	6	6	6	6	6	1	1	1	1	1	1
5	sum	74	49	34	15	10	8	2	6	2	2
6	2	1	1	0	0	1	1	0	1	0	0
6	3	12	6	0	0	7	5	0	3	0	0
6	4	93	76	6	1	20	17	1	8	2	2
6	5	175	134	65	8	7	5	3	3	2	2
6	6	98	78	76	29	7	5	3	3	2	2
6	7	14	12	14	12	1	1	1	1	1	1
6	sum	393	313	161	52	36	29	5	16	5	4
7	2	1	1	0	0	1	1	0	1	0	0
7	3	16	7	0	0	10	6	0	4	0	0
7	4	212	163	0	0	55	39	0	16	0	0
7	5	774	533	103	6	55	39	5	16	6	5
7	6	904	613	418	60	10	6	5	4	3	3
7	7	341	239	275	113	1	1	1	1	1	1
7	8	34	27	34	34	1	1	1	1	1	1
7	sum	2282	1592	830	213	132	92	11	42	10	9

TABLE 4.—Continued

<i>m</i>	<i>n</i>	Code = 1 @ #	Code = 2 @ #	Code = 3 @ #	Code = 4 @ #	Code = 5 @ #	Code = 6 @ #	Code = 7 @ #	Code = 8 @ #	Code = 9 @ #	Code = 10 @ #
8	2				0	1	1	0	1	0	0
8	3		0		0	14	8	0	5	0	0
8	4		0		0	145	114	0	38	0	0
8	5		2		2	296	217	2	29	0	0
8	6		73		49	145	114	22	63	4	2
8	7		388		237	14	14	22	38	19	17
8	8		444		271	8	8	8	5	4	4
8	9		95		65	1	1	1	1	1	1
8	sum		1002		624	616	475	33	151	29	24
9	2						403	28	114	24	24
9	3							0	1	0	0
9	4							0	7	0	0
9	5							0	72	0	0
9	6							1	218	4	1
9	7							32	218	41	21
9	8							71	72	44	42
9	9							12	51	6	6
9	sum							117	596	96	81
10	2							0	1	0	0
10	3							0	8	0	0
10	4							0	134	0	0
10	5							0	622	0	0
10	6							22	1075	59	0
10	7							216	622	182	22
10	8							176	370	111	157
10	9							16	134	90	87
10	10							1	8	7	7
10	sum							431	2605	339	274
								285	1565	222	177

vertices. Clearly, a map with a face (vertex) of degree 1 has a loop (an isthmus), and a loopless map with at least 2 edges which has a face of degree 2 has parallel edges. If the edge-set of a planar map M can be partitioned into two disjoint nonnull subsets S and T so that there is just one vertex v incident with both a member of S and a member of T , then v is called a *cut-vertex* of M . Clearly, a planar map with at least 2 edges which has either a loop or an isthmus has a cut-vertex. A planar map without cut-vertices is called *2-connected*. If the edge-set of a 2-connected planar map M can be partitioned into two disjoint subsets S and T with at least 2 edges apiece so that there are just 2 vertices u and v incident with both a member of S and a member of T , then $\{u, v\}$ is called a *cut-pair* of M . Clearly a 2-connected planar map with at least 4 edges which has either vertices of degree 2 or parallel edges has a cut-pair. A 2-connected planar map without cut-pairs is called *3-connected*.

This section deals with the effective generation of rooted planar maps with (or rather, without) the above-mentioned properties. A list of the types of planar maps we have generated appears in Table 3. See also Table 4.

Certain properties of a rooted planar map (M, d) are easily interpreted in terms of its code $l_r^{-1}(M, d)$. Lehman has proved the following two propositions [13], [22], [26].

PROPOSITION 7. *If $S = s_1, s_2, \dots, s_{2m} = l_r^{-1}(X, \sigma, \alpha, d)$, then $S^* = s_{2m}^*, s_{2m-1}^*, \dots, s_2^*, s_1^* = l_r^{-1}(X, \alpha\sigma^{-1}, \alpha, \alpha\sigma^{-1}(d))$, where $(^* =), [^* =),]^* = [,]^* = (:$ to get from S to S^* , change all the brackets to parentheses and vice versa and turn the whole system backwards [22, p. 106].*

PROPOSITION 8. *For each property R listed in Table 5, the rooted planar map (M, d) has property R if and only if $S = l_r^{-1}(M, d) = s_1, s_2, \dots, s_{2m}$ has one of the configurations in the corresponding set $l_r^{-1}(R)$.*

We prolong this list in

PROPOSITION 9. *The conclusions of Proposition 8 hold for the pairs $R, l_r^{-1}(R)$ in Table 6.*

Parallel edges are larger to express in terms of the code, but one can test for them by means of the adjacency matrix.

TABLE 5

R	$l_r^{-1}(R)$
a face of degree 1 [22, p. 133]	a pair of adjacent brackets, or s_1 and s_{2m} are a bracket pair;
a vertex of degree 1 [22, p. 133]	a pair of adjacent parentheses, or s_1 and s_{2m} are a parenthesis pair;
a loop [22, p. 151]	a bracket pair surrounding a parenthesis-bracket system;
an isthmus [22, p. 151]	a parenthesis pair surrounding a parenthesis-bracket system;
a cut-vertex [22, p. 145]	a nonnull proper segment which is a parenthesis-bracket system.

TABLE 6

R	$l_r^{-1}(R)$
a face of degree 2	a bracket pair surrounding a single right parenthesis, or two bracket pairs, one immediately inside the other, or s_2, s_{2m} is a bracket pair and s_1 is a left parenthesis, or two bracket pairs s_1, s_i and s_{i+1}, s_{2m} ;
a vertex of degree 2	S^* has one of the above 4 configurations;
a cut-pair	two nonnull segments a and b with a total of at least 4 and at most $2m - 4$ symbols such that the mate in S of any symbol in a or b is also either in a or in b .

Evidently one could generate any of the subsets of rooted planar maps listed in Table 3 by generating all the correct parenthesis-bracket systems and testing each one for the presence of the appropriate configurations listed in Propositions 8 and 9. Some idea of the inefficiency of this procedure can be obtained from the observation that the proportion of m -edge rooted planar maps which are 2-connected is asymptotic to $2(9/16)^m$, whereas we have generated the rooted 2-connected planar maps in an average time per map of only 50% greater than the average time taken by Algorithm 2 to generate an arbitrary planar map.

We spare the reader the tedious details of the algorithm used for each class of maps—the descriptions and/or listings (in FORTRAN!) are available on request. Suffice it to say that instead of inserting parenthesis systems into bracket systems we insert individual pairs of brackets into parenthesis systems (which is like inserting fronds into rooted trees) and we avoid insertions which will make forbidden configurations inevitable. In most cases this involves straightforward branching and bounding, but in the case of 2-connected planar maps we also use

PROPOSITION 10. *If S is a parenthesis-bracket system with no subword $(\)$, then S has no nonnull proper segments which are also parenthesis-bracket systems if and only if the following 3 conditions are satisfied:*

1. *Every bracket pair encloses at least one parenthesis and no bracket pair encloses all the parentheses (that is, there are no faces of degree 1).*

2. *The first and last parentheses form a pair—the outside pair (that is, the root-vertex of the canonical spanning tree T is of degree 1).*

3. *Given any parenthesis pair s_i, s_j in S except the outside pair, let s_k be the right parenthesis of the pair immediately surrounding s_i, s_j (k is the smallest number greater than j such that $E(k) < E(j)$). Then there is a bracket pair which forms the word $(\)$ with s_i, s_j and s_k . (That is, for every vertex v of T except the root and the leaves and for every son s of v there is a frond joining s or one of its descendants to some proper ancestor of v .)*

The number of nonisomorphic 2-connected maps with $m \leq 10$ edges was checked against the counting formula in [16]. The corresponding numbers for the other 9 classes of planar maps in Table 3 were checked using polynomial-time counting algorithms we obtained in [23].

The problems of generating 2-connected planar maps with faces of degree 2 or vertices of degree 2 and of generating 3-connected planar maps by these methods are considerably more difficult and are as yet unsolved. In [10, p. 142] there is a test for the 3-connectivity of graphs with depth-first-search spanning trees which could be applied to Lehman's code. A program based on that principle could generate 3-connected planar maps, no two of which are related by either isomorphism or reflection, without using as much storage space as Tutte's scheme [20]. But it is not at all clear which program would run faster.

Acknowledgment. The author is grateful for the support and facilities provided by the Computing Centre of the U.S.S.R. Academy of Sciences in Moscow, and in particular wishes to thank its vice-director N. N. Moiseev.

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] H. H. BAKER, A. K. DEWDNEY AND A. L. SZILARD, *Generating the nine-point graphs*, Math. Comp., 28 (1974), pp. 833–838.

- [3] W. G. BROWN, *On the enumeration of non-planar maps*, Mem. Amer. Math. Soc., 65 (1966), pp. 1–42.
- [4] R. CORI, *Un code pour les graphes planaires et ses applications*, Astérisque, 27, Paris, 1975.
- [5] A. J. W. DUIJVESTIJN, *Simple perfect squared squares of lowest order*, J. Combin. Theory Ser. B, (1978), pp. 240–245.
- [6] J. R. EDMONDS, *A combinatorial representation for oriented polyhedral surfaces*, M. A. Thesis, University of Maryland, College Park, 1960. See also Notices Amer. Math. Soc., 7 (1960), p. 646.
- [7] I. A. FARADZHEV, *Generation of non-isomorphic graphs with a given distribution of vertex-degrees*, Algorithmic Investigations in Combinatorics, Nauka, Moscow, 1976, (Russian).
- [8] B. R. HEAP, *The production of graphs by computer*, in Graph Theory and Computing, Academic Press, New York and London, 1972, pp. 47–62.
- [9] J. E. HOPCROFT AND R. E. TARJAN, *A $V \log V$ algorithm for isomorphism of triconnected planar graphs*, J. Comput. Systems Sci., 7 (1973), pp. 321–331.
- [10] ———, *Dividing a graph into triconnected components*, SIAM J. Comput., 3 (1973), pp. 135–158.
- [11] A. JACQUES, *Constellations et graphes topologiques*, in Combinatorial Theory and its Applications, P. Erdős et al., eds., Colloq. Math. Soc. Janos Bolyai, North-Holland, Amsterdam, 1970, pp. 657–672.
- [12] P. LÄUCHLI, *Generating all planar 0-, 1-, 2-, 3-connected graphs*, Lecture Notes in Computer Science 100, Springer-Verlag, New York, 1981, pp. 379–382.
- [13] A. B. LEHMAN, *Maps and their encoding*, unpublished manuscript, 1965.
- [14] ———, *Full genus maps and hypermaps*, unpublished manuscript, 1973.
- [15] V. A. LISKOVETS, *On the enumeration of nonisomorphic planar maps*, in Proc. International Colloquium on Algebraic Methods in Graph Theory, Szeged, Hungary, 1978, pp. 479–494.
- [16] V. A. LISKOVETS AND T. R. S. WALSH, *Counting non-isomorphic 2-connected planar maps*, Canad. J. Math., submitted. See also Graph Theory Newslett. 9, no. 6 (1980), p. 3.
- [17] R. C. READ, *Every one a winner or how to avoid isomorphism search when cataloguing combinatorial configurations*, Ann. Discrete Math., 2 (1978), pp. 107–120.
- [18] R. Tarjan, *Depth-first search and linear graph algorithms*, SIAM J. Comput., 1 (1972), 146–160.
- [19] G. Tarry, *Le problème des labyrinthes*, Nouvelles Annales de Mathématiques, 3 (1895), pp. 187–190.
- [20] W. T. TUTTE, *A theory of 3-connected graphs*, Indag. Math., 23 (1961), pp. 441–455.
- [21] ———, *A census of planar maps*, Canad. J. Math., 15 (1963), pp. 249–271.
- [22] T. R. S. WALSH, *Combinatorial enumeration of non-planar maps*, Ph.D. Thesis, Univ. of Toronto, Toronto, Ontario, Canada, 1971.
- [23] ———, *Counting non-isomorphic three-connected planar maps*, J. Combin. Theory Ser. B, 32 (1982), pp. 33–44.
- [24] T. R. S. WALSH AND A. B. LEHMAN, *Counting rooted maps by genus, I*, J. Combin. Theory Ser. B 13 (1972), pp. 192–218.
- [25] ———, *Counting rooted maps by genus, II*, J. Combin. Theory Ser. B, 13 (1972), pp. 122–141; Erratum, 14 (1973), p. 185.
- [26] ———, *Counting rooted maps by genus, III: Non-separable maps*, J. Combin. Theory Ser. B, 18 (1975), pp. 222–259.
- [27] N. C. WORMALD, *Counting unrooted planar maps*, Research Report Corr. 79–32, Univ. of Waterloo, Waterloo, Ontario, Canada, 1979.

THE HAMMOND SERIES OF A SYMMETRIC FUNCTION AND ITS APPLICATION TO P -RECURSIVENESS*

I. P. GOULDEN,† D. M. JACKSON† AND J. W. REILLY‡

Abstract. We give a method for determining the exponential generating function for the coefficient of $x_1^p \cdots x_n^p$ in a symmetric function S in the indeterminates x_1, \dots, x_n . This generating function is called the Hammond series of S , and we use it to show that the counting series for certain combinatorial problems satisfy linear recurrence equations with polynomial coefficients. These problems include p -regular labelled graphs and square matrices with row and column sums equal to p .

1. Introduction. Let $[(a_1^p \cdots a_\alpha^p)(b_1^q \cdots b_\beta^q) \cdots]T$ denote the coefficient of $(a_1^p \cdots a_\alpha^p)(b_1^q \cdots b_\beta^q) \cdots$ in the formal power series T which is a symmetric function in each of the sets $\{a_1, \dots, a_\alpha\}, \{b_1, \dots, b_\beta\} \cdots$ of commutative indeterminates. We call such coefficients the *regular* coefficients of T . In this paper we present a method for calculating the exponential generating function for regular coefficients, where p, q, \dots are fixed. We call this power series the Hammond series (or H -series) of T , because of its connection to the Hammond operators.

In the later sections of this paper we use the H -series to determine whether certain sequences of regular coefficients satisfy a linear recurrence equation of fixed order, with polynomial coefficients. Such sequences are called *polynomially-recursive* (or P -recursive). This term is of considerable importance computationally since it means that the n th term of such a sequence may be computed in an amount of time which is linear in n and space which is independent of n (assuming that the time taken to multiply two integers is independent of their size).

We establish P -recursiveness for a sequence by deriving a linear differential equation, with polynomial coefficients, for its H -series. Power series with this property are called *differentially-finite* (or D -finite). The equivalence of D -finiteness and P -recursiveness is discussed in Stanley [6].

Regular coefficients arise in a variety of contexts and the problem of calculating them is a classical one which has been considered by MacMahon [3] in his combinatorial work on symmetric functions. We use the H -series to study two combinatorial configurations, namely

- a) p -regular labelled graphs and simple graphs on n vertices for $n = 0, 1, 2, \dots$ and
- b) $n \times n$ matrices with row and column sums p over the nonnegative integers for $n = 0, 1, 2, \dots$.

This enables us to establish the P -recursiveness of (a) for $p = 4$ and (b) for $p = 3$, an open problem cited by Stanley [6].

The following notation is used. Let $\mathbf{x} = (x_1, x_2, \dots)$ and $\mathbf{y} = (y_1, y_2, \dots)$ be sets of indeterminates. If $\mathbf{i} = (i_1, i_2, \dots)$, then $\mathbf{x}^{\mathbf{i}}$ denotes $x_1^{i_1} x_2^{i_2} \cdots$ and $[\mathbf{x}^{\mathbf{i}}]f(\mathbf{x})$ denotes the coefficient of $\mathbf{x}^{\mathbf{i}}$ in the formal power series $f(\mathbf{x})$. Let $\partial/\partial\mathbf{y}$ denote $(\partial/\partial y_1, \partial/\partial y_2, \dots)$. We say that $\mathbf{i} \geq \mathbf{j}$, where $\mathbf{j} = (j_1, j_2, \dots)$, if $i_1 \geq j_1, i_2 \geq j_2, \dots$.

We begin by considering an arbitrary symmetric formal power series T in the single set $\mathbf{t} = (t_1, t_2, \dots)$ of commutative indeterminates, since the extension to the multisymmetric case is straightforward. Let $\tau(i_1, i_2, \dots) = (j_1, j_2, \dots) = \mathbf{j}$, where j_k is

* Received by the editors March 12, 1982, and in final form June 7, 1982. The work of these authors was supported (in part) by the Natural Sciences and Engineering Research Council of Canada under Grants U0073 and A8235.

† Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

‡ Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A7.

the number of occurrences of k in (i_1, i_2, \dots) . We define the H -series of $T(\mathbf{t})$ to be the formal power series g , in the indeterminates $\mathbf{y} = (y_1, y_2, \dots)$, such that

$$\left[\frac{\mathbf{y}^{\mathbf{j}}}{\mathbf{j}!} \right] g(\mathbf{y}) = [\mathbf{t}^{\mathbf{j}}] T(\mathbf{t}),$$

where $\mathbf{j}! = j_1! j_2! \dots$. The H -series, g , of T is denoted by $H(T)$.

For regular coefficients, we observe that

$$[t_1^p \dots t_n^p] T(\mathbf{t}) = \left[\frac{z^n}{n!} \right] f(z),$$

where $f(z) = (H(T))(0, \dots, 0, z, 0, \dots)$, where z occurs as the p th argument. The H -series therefore enables us to obtain a univariate exponential generating function for the regular coefficients of the multivariate generating function T .

2. The H -series. Let $s_k = t_1^k + t_2^k + \dots$ and let $\mathbf{s} = (s_1, s_2, \dots)$, where s_k is called a *power sum symmetric function*. Now the symmetric power series $T(\mathbf{t})$ can be expressed uniquely in terms of the power sum symmetric functions, to give $T(\mathbf{t}) = G(\mathbf{s}(\mathbf{t}))$. We adopt the notational convention that the H -series, $H(T)(\mathbf{y})$, of $T(\mathbf{t})$ may also be denoted by $H(G)(\mathbf{y})$ without ambiguity, since G and T are used only in this context.

The next theorem enables us to express the H -series for $\partial G / \partial s_n$ and $s_n G$ in terms of the H -series for G . We shall use this theorem later to deduce a system of differential equations for $H(G)(\mathbf{y})$ from a system for $G(\mathbf{s})$. It happens that the latter system is often easy to derive.

THEOREM (H -series). *Let $g(\mathbf{y})$ be the H -series for a symmetric function $T(\mathbf{t})$ and let $T(\mathbf{t}) = G(\mathbf{s}(\mathbf{t}))$. Then*

$$1) \quad \left(H \left(\frac{\partial G}{\partial s_n} \right) \right) (\mathbf{y}) = \sum_{\mathbf{i} \geq \mathbf{0}} (-1)^{m-1} \frac{1}{\mathbf{i}!} (m-1)! \frac{\partial^{\mathbf{i}}}{\partial \mathbf{y}^{\mathbf{i}}} g(\mathbf{y}),$$

where $m = i_1 + i_2 + \dots$ and the summation is over $\mathbf{i} = (i_1, i_2, \dots)$ such that $i_1 + 2i_2 + \dots = n$.

$$2) \quad (H(s_n G))(\mathbf{y}) = \left\{ y_n + \sum_{i \geq 1} y_{n+i} \frac{\partial}{\partial y_i} \right\} g(\mathbf{y}).$$

Proof. 1) Let $A_{\mathbf{i}}(\mathbf{t})$ denote the *monomial symmetric function* defined by

$$A_{\mathbf{i}}(\mathbf{t}) = \sum_{\substack{\mathbf{j} \geq \mathbf{0} \\ \tau(\mathbf{j}) = \mathbf{i}}} \mathbf{t}^{\mathbf{j}} = [\mathbf{x}^{\mathbf{i}}] \prod_{k \geq 1} (1 + x_1 t_k + x_2 t_k^2 + \dots).$$

Since $T(\mathbf{t})$ is a symmetric function in t_1, t_2, \dots , there exist $c(\mathbf{i})$, independent of \mathbf{t} , such that

$$T(\mathbf{t}) = \sum_{\mathbf{i} \geq \mathbf{0}} c(\mathbf{i}) A_{\mathbf{i}}(\mathbf{t})$$

whence

$$H(T)(\mathbf{y}) = \sum_{\mathbf{i} \geq \mathbf{0}} c(\mathbf{i}) \frac{\mathbf{y}^{\mathbf{i}}}{\mathbf{i}!}.$$

Let $E_n(\mathbf{x}) = \sum_{i \geq 0} (-1)^{m-1} (m-1)! \mathbf{x}^i / i!$, where m and the range of summation are defined in 1). Then $\sum_{n \geq 1} E_n(\mathbf{x}) z^n = \log(1 + zx_1 + z^2 x_2 + \dots)$, so that

$$\begin{aligned} \sum_{i \geq 0} A_i(\mathbf{t}) \mathbf{x}^i &= \prod_{j \geq 1} (1 + x_1 t_j + x_2 t_j^2 + \dots) = \exp \sum_{j \geq 1} \log(1 + x_1 t_j + x_2 t_j^2 + \dots) \\ &= \exp \sum_{j \geq 1} \sum_{n \geq 1} E_n(\mathbf{x}) t_j^n = \exp \sum_{n \geq 1} E_n(\mathbf{x}) s_n(\mathbf{t}). \end{aligned}$$

From this we obtain

$$\begin{aligned} \sum_{i \geq 0} \mathbf{x}^i \frac{\partial}{\partial s_n} A_i(\mathbf{t}) &= \frac{\partial}{\partial s_n} \sum_{i \geq 0} \mathbf{x}^i A_i(\mathbf{t}) = \frac{\partial}{\partial s_n} \exp \sum_{n \geq 1} E_n(\mathbf{x}) s_n \\ &= E_n(\mathbf{x}) \exp \sum_{n \geq 1} E_n(\mathbf{x}) s_n = E_n(\mathbf{x}) \sum_{j \geq 0} A_j(\mathbf{t}) \mathbf{x}^j. \end{aligned}$$

The application of $[\mathbf{x}^i]$ to this equation yields

$$(2.1) \quad \frac{\partial}{\partial s_n} A_i(\mathbf{t}) = \sum_{0 \leq j \leq i} A_j(\mathbf{t}) [\mathbf{x}^{i-j}] E_n(\mathbf{x}).$$

Thus

$$H\left(\frac{\partial G}{\partial s_n}\right) = \sum_{i \geq 0} c(i) H\left(\frac{\partial}{\partial s_n} A_i(\mathbf{t})\right),$$

so from (2.1) we have

$$\begin{aligned} H\left(\frac{\partial G}{\partial s_n}\right) &= \sum_{i \geq 0} c(i) \sum_{j \leq i} \{[\mathbf{x}^{i-j}] E_n(\mathbf{x})\} H(A_j(\mathbf{t})) \\ &= \sum_{i \geq 0} c(i) \sum_{j \leq i} [\mathbf{x}^{i-j}] E_n(\mathbf{x}) \mathbf{y}^j / j! \\ &= \sum_{i \geq 0} c(i) \sum_{j \leq i} [\mathbf{x}^{i-j}] E_n(\mathbf{x}) \frac{\partial^{i-j}}{\partial \mathbf{y}^{i-j}} (\mathbf{y}^i / i!). \end{aligned}$$

Now let $\mathbf{i} - \mathbf{j} = \mathbf{k} \geq \mathbf{0}$ since $\mathbf{i} \geq \mathbf{j}$. Thus

$$\begin{aligned} H\left(\frac{\partial G}{\partial s_n}\right) &= \sum_{i \geq 0} c(i) \sum_{\mathbf{k} \geq \mathbf{0}} [\mathbf{x}^{\mathbf{k}}] E_n(\mathbf{x}) \frac{\partial^{\mathbf{k}}}{\partial \mathbf{y}^{\mathbf{k}}} (\mathbf{y}^i / i!) \\ &= \sum_{\mathbf{k} \geq \mathbf{0}} \frac{\partial^{\mathbf{k}}}{\partial \mathbf{y}^{\mathbf{k}}} [\mathbf{x}^{\mathbf{k}}] E_n(\mathbf{x}) \sum_{i \geq 0} c(i) \mathbf{y}^i / i! \\ &= E_n\left(\frac{\partial}{\partial \mathbf{y}}\right) H(T) \quad \text{and 1) follows.} \end{aligned}$$

2) Let $\delta_n = (0, \dots, 0, 1, 0, \dots)$, where the one appears in the n th position. Thus, by definition we have

$$\begin{aligned} s_n A_{\mathbf{l}} &= \left(\sum_{\tau(0) = \delta_n} x_1^{\tau_1} x_2^{\tau_2} \dots \right) \sum_{\tau(0) = \mathbf{l}} x_1^{\tau_1} x_2^{\tau_2} \dots \\ &= \sum_{\tau(0) = \delta_n} \sum_{\tau(0) = \mathbf{l}} x_1^{\tau_1 + j_1} x_2^{\tau_2 + j_2} \dots, \end{aligned}$$

where $\mathbf{l} = (l_1, l_2, \dots)$ and $\mathbf{j} = (j_1, j_2, \dots)$. Now the effect of \mathbf{l} is to change a single k th power in \mathbf{j} to a $(k + n)$ th power, for some k , in all possible ways. Each of the resulting

monomials may be obtained in $i_{n+k} + 1$ ways. Thus, if $\mathbf{m} = (m_1, \dots)$,

$$s_n A_{\mathbf{i}} = \sum_{k \geq 1} (1 + i_{n+k}) \sum_{\tau(\mathbf{m}) = i + \delta_{n+k} - \delta_k} x_1^{m_1} x_2^{m_2} \dots + (1 + i_n) \sum_{\tau(\mathbf{m}) = i + \delta_n} x_1^{m_1} x_2^{m_2} \dots$$

and

$$H\left(s_n \sum_{\mathbf{i} \geq 0} c(\mathbf{i}) A_{\mathbf{i}}\right) = C_n\left(\mathbf{y}, \frac{\partial}{\partial \mathbf{y}}\right) H(G),$$

where $C_n(\mathbf{y}, \partial/\partial \mathbf{y}) = y_n + \sum_{i \geq 1} y_{n+i} \partial/\partial y_i$. The result follows immediately. \square

It follows from the H -series theorem that $H(s_1^{i_1} \dots \partial^{i_1}/\partial s_1^{i_1} \dots G) = C_1^{i_1}(\mathbf{y}, \partial/\partial \mathbf{y}) \dots E_1^{i_1}(\partial/\partial \mathbf{y}) \dots H(G)$. Thus any differential equation for $G(\mathbf{s})$ may be translated, by means of the H -series theorem, into a differential equation for $H(G)$. We note that $C_n(\mathbf{y}, \partial/\partial \mathbf{y})$ and E_n are reminiscent of Hammond operators for symmetric functions (MacMahon [3] and Hammond [2]).

3. Preliminary application. In this section and §§ 4 and 5 we consider the enumeration of p -regular labelled graphs and simple graphs. We now set up a system of differential equations for labelled simple graphs and demonstrate the use of the H -series theorem for the 2-regular case.

Let $T(\mathbf{t})$ be the ordinary generating function for simple labelled graphs, where t_j marks the degree of vertex j for $j \geq 1$. The generating function for the pair $\{i, j\}$ of distinct vertices is $1 + t_i t_j$ since if i and j are not joined by an edge there is no contribution to the degrees of i and j , while if they are then there is a contribution of 1 to each of the degrees of i and j . Thus

$$T(\mathbf{t}) = \prod_{1 \leq i < j} (1 + t_i t_j).$$

We next derive $G(\mathbf{s})$, where $G(\mathbf{s}(\mathbf{t})) = T(\mathbf{t})$. Now

$$\begin{aligned} T(\mathbf{t}) &= \exp \log \prod_{1 \leq i < j} (1 + t_i t_j) = \exp \sum_{k \geq 1} \frac{1}{k} (-1)^{k-1} \sum_{1 \leq i < j} (t_i t_j)^k \\ &= \exp \sum_{k \geq 1} \frac{1}{2k} (-1)^{k-1} \{s_k^2(\mathbf{t}) - s_{2k}(\mathbf{t})\}, \end{aligned}$$

whence

$$G(\mathbf{s}) = \exp \sum_{k \geq 1} \frac{1}{2k} (-1)^{k-1} \{s_k^2 - s_{2k}\}.$$

The system of differential equations which $G(\mathbf{s})$ satisfies is

$$\begin{aligned} \frac{\partial G}{\partial s_{2k+1}} &= \frac{1}{2k+1} s_{2k+1} G \quad \text{for } k \geq 0, \\ \frac{\partial G}{\partial s_{2k}} &= \frac{1}{2k} \{(-1)^k - s_{2k}\} G \quad \text{for } k \geq 1. \end{aligned} \tag{3.1}$$

This is the general system of equations for labelled simple graphs. For the moment we confine our attention to 2-regular simple graphs.

Let $r_2(n)$ denote the number of 2-regular simple labelled graphs on n vertices. Then

$$r_2(n) = [t_1^2 \dots t_n^2] T(\mathbf{t}) = \left[\frac{y_2^n}{n!} \right] U(y_1, y_2),$$

where $U(y_1, y_2) = H(G)(y_1, y_2, 0, \dots)$. Thus, applying the H -series theorem to $T(\mathbf{t})$ and setting $y_3 = y_4 = \dots = 0$, we have

$$\frac{\partial U}{\partial y_1} = y_1 U + y_2 \frac{\partial U}{\partial y_1}, \quad 2 \frac{\partial U}{\partial y_2} - \frac{\partial^2 U}{\partial y_1^2} = -U - y_2 U.$$

Eliminating $\partial U/\partial y_1$ and $\partial^2 U/\partial y_1^2$ and then setting $y_1 = 0$, we have $dV/dy_2 = \{(1 - y_2)^{-1} - (1 + y_2)\}V/2$ and $r_2(n) = [y_2^n/n!]V$, where $V(y_2) = U(0, y_2)$ and $V(0) = 1$. Since V is differentiably finite (or D -finite) it follows (Stanley [6, Thm. 1.5]) that $\{r_2(n)|n \geq 0\}$ is P -recursive. Indeed, applying $[y_2^n/n!]$ to both sides of this ordinary differential equation for V we have

$$2r_2(n+1) - 2nr_2(n) - n(n-1)r_2(n-2) = 0,$$

where $r_2(0) = 1$ and $r_2(k) = 0$ for $k < 0$. We note that we may solve the differential equation to obtain

$$V(y_2) = (1 - y_2)^{-1/2} \exp \left\{ -\frac{y_2}{2} - \frac{y_2^2}{4} \right\},$$

the well-known generating function for the number of cycle covers of the complete graph on n vertices. This may, of course, be obtained by a direct argument, but its derivation here has illustrated the use of the H -series.

It is important to note that the ordinary generating function for labelled graphs is, by a similar argument,

$$T'(\mathbf{t}) = \prod_{1 \leq i \leq j} (1 - t_i t_j)^{-1} = G'(\mathbf{s}(\mathbf{t})),$$

where

$$G'(\mathbf{s}) = \exp \sum_{k \geq 1} \frac{1}{2k} \{s_k^2(\mathbf{t}) + s_{2k}(\mathbf{t})\}.$$

The system of differential equations associated with $G'(\mathbf{s})$ is

$$(3.2) \quad \begin{aligned} \frac{\partial G'}{\partial s_{2k+1}} &= \frac{1}{2k+1} s_{2k+1} G' \quad \text{for } k \geq 0, \\ \frac{\partial G'}{\partial s_{2k}} &= \frac{1}{2k} \{1 + s_{2k}\} G' \quad \text{for } k \geq 1. \end{aligned}$$

Systems (3.1) and (3.2) are strongly related to each other. Accordingly, in § 4 we shall give certain details for calculations with (3.1) but totally suppress the corresponding details for calculations with (3.2) since they are similar.

4. The P -recursiveness of the numbers of 3- and 4-regular labelled graphs and simple graphs on n vertices. Let $r_p(n)$ be the number of p -regular simple labelled graphs on n vertices. We apply the method of § 3 to the cases $p = 3$ and 4 to derive differential equations with polynomial coefficients for $\sum_{n \geq 0} r_p(n)(x^n/n!)$. The calculations are of course more prolonged, and we have suppressed their details because they add nothing of conceptual importance to the argument.

We consider first the case $p = 3$. Applying the H -series theorem to system (3.1) for $G(s)$ and putting $y_4 = y_5 = \dots = 0$, we have

$$\begin{aligned}
 (4.1) \quad & \text{a) } \frac{\partial A^{(3)}}{\partial y_1} = y_1 A^{(3)} + y_2 \frac{\partial A^{(3)}}{\partial y_1} + y_3 \frac{\partial A^{(3)}}{\partial y_2}, \\
 & \text{b) } 2 \frac{\partial A^{(3)}}{\partial y_2} - \frac{\partial^2 A^{(3)}}{\partial y_1^2} = -(1 + y_2) A^{(3)} - y_3 \frac{\partial A^{(3)}}{\partial y_1}, \\
 & \text{c) } 3 \frac{\partial A^{(3)}}{\partial y_3} - 3 \frac{\partial^2 A^{(3)}}{\partial y_1 \partial y_2} + \frac{\partial^3 A^{(3)}}{\partial y_1^3} = y_3 A^{(3)},
 \end{aligned}$$

where $A^{(3)}(y_1, y_2, y_3) = H(G)(y_1, y_2, y_3, 0, \dots)$ and $r_3(n) = [y_3^n/n!]A^{(3)}$.

Let $B^{(3)}(y_1, y_3) = A^{(3)}(y_1, 0, y_3)$. By inspection we may express $\partial B^{(3)}/\partial y_3$ and $\partial^2 B^{(3)}/\partial y_3^2$ solely in terms of $\partial B^{(3)}/\partial y_1$. We therefore have a system of two simultaneous linear equations for the unknown $\partial B^{(3)}/\partial y_1$. Eliminating $\partial B^{(3)}/\partial y_1$ between these equations and setting $y_1 = 0$, we obtain a second order linear ordinary differential equation in y_3 for $B^{(3)}(0, y_3)$, with polynomial coefficients, so $\{r_3(n) | n \geq 0\}$ is P -recursive. To simplify this equation we note that $r_3(2n + 1) = 0$ for $n \geq 0$ since the sum of the degrees in a graph is even. Thus $B^{(3)}(0, y_3)$ is a power series $R_3(x)$ in $y_3^2 = x$. Now

$$y_3 \frac{\partial B^{(3)}}{\partial y_3} = 2x \frac{dR_3}{dx} \quad \text{and} \quad \frac{\partial^2 B^{(3)}}{\partial y_3^2} = 2 \frac{\partial R_3}{\partial x} + 4x \frac{\partial^2 R_3}{\partial x^2}.$$

Thus $R_3(x) = \sum_{n \geq 0} r_3(2n)x^n/(2n)!$ satisfies the differential equation given in Table 4.1(i). This agrees with Read [4]. A similar argument applied to system (3.2) gives the ordinary differential equation for $Q_3(x) = \sum_{n \geq 0} q_3(2n)x^n/(2n)!$, where $q_3(2n)$ is the number of labelled 3-regular graphs on $2n$ vertices. This is given in Table 4.1(ii).

TABLE 4.1(i)
The differential equation for the number of 3-regular simple labelled graphs.

i	$\phi_i(x)$
0	$x(-x^2 - 2x + 2)^2$
1	$-6(x^5 + 6x^4 + 6x^3 - 32x + 8)$
2	$36x^2(-x^2 - 2x + 2)$

$$R_3(x) = \sum_{n \geq 0} r_3(2n) \frac{x^n}{(2n)!} : \phi_2(x) \frac{d^2 R_3(x)}{dx^2} + \phi_1(x) \frac{dR_3(x)}{dx} + \phi_0(x) R_3(x) = 0$$

TABLE 4.1(ii)
The differential equation for the number of 3-regular labelled graphs.

i	$\phi_i(x)$
0	$x^5 - 10x^4 + 24x^3 - 4x^2 - 44x - 48$
1	$-6(x^5 - 6x^4 + 6x^3 + 24x^2 + 16x - 8)$
2	$36x^2(x^2 - 2x - 2)$

$$Q_3(x) = \sum_{n \geq 0} q_3(2n) \frac{x^n}{(2n)!} : \phi_2(x) \frac{d^2 Q_3}{dx^2}(x) + \phi_1(x) \frac{dQ_3}{dx}(x) + \phi_0(x) Q_3(x) = 0$$

The case $p = 4$ may be treated in a similar way. Applying the H -series theorem to system (3.1) for $G(s)$ and putting $y_5 = y_6 = \dots = 0$, we have

$$\begin{aligned}
 & \frac{\partial A^{(4)}}{\partial y_1} = y_1 A^{(4)} + y_2 \frac{\partial A^{(4)}}{\partial y_1} + y_3 \frac{\partial A^{(4)}}{\partial y_2} + y_4 \frac{\partial A^{(4)}}{\partial y_3}, \\
 & 2 \frac{\partial A^{(4)}}{\partial y_2} - \frac{\partial^2 A^{(4)}}{\partial y_1^2} = -(1 + y_2)A^{(4)} - y_3 \frac{\partial A^{(4)}}{\partial y_1} - y_4 \frac{\partial A^{(4)}}{\partial y_2}, \\
 & 3 \frac{\partial A^{(4)}}{\partial y_3} - 3 \frac{\partial^2 A^{(4)}}{\partial y_1 \partial y_2} + \frac{\partial^3 A^{(4)}}{\partial y_1^3} = y_3 A^{(4)} + y_4 \frac{\partial A^{(4)}}{\partial y_1}, \\
 & 4 \frac{\partial A^{(4)}}{\partial y_4} - 4 \frac{\partial^2 A^{(4)}}{\partial y_1 \partial y_3} - 2 \frac{\partial^2 A^{(4)}}{\partial y_2^2} + 4 \frac{\partial^3 A^{(4)}}{\partial y_1^2 \partial y_2} - \frac{\partial^4 A^{(4)}}{\partial y_1^4} = (1 - y_4)A^{(4)},
 \end{aligned}
 \tag{4.2}$$

where $A^{(4)}(y_1, y_2, y_3, y_4) = H(G)(y_1, y_2, y_3, y_4, 0, \dots)$ and $r_4(n) = [y_4^n/n!]A^{(4)}$.

Let $B^{(4)}(y_1, y_4) = A^{(4)}(y_1, 0, 0, y_4)$. By inspection, we may express $\partial^m B^{(4)}/\partial y_4^m$ linearly in terms of $B^{(4)}, \partial B^{(4)}/\partial y_1, \partial^2 B^{(4)}/\partial y_1^2$ alone for $m \geq 1$. In fact, when we carry this out for $m = 1, 2$ (using the symbolic algebra system VAXIMA, as described in § 8) and set $y_1 = 0$, we find that the coefficient of $\partial B^{(4)}/\partial y_1$ at $y_1 = 0$ is 0 in both equations. Eliminating $\partial^2 B^{(4)}/\partial y_1^2$ at $y_1 = 0$ between these two equations, we obtain a second order differential equation for $R_4(x) = \sum_{n \geq 0} r_4(n)(x^n/n!)$, where $R_4(x) = B^{(4)}(0, x)$. This differential equation is given in Table 4.2(i) and demonstrates that $R_4(x)$ is D -finite so $\{r_4(n)|n \geq 0\}$ is P -recursive. The corresponding differential equation for $Q_4(x) = \sum_{n \geq 0} q_4(n)(x^n/n!)$, where $q_4(n)$ is the number of 4-regular labelled graphs, is deduced in a similar way from system (3.2) and is given in Table 4.2(ii).

We have therefore established the following result.

COROLLARY. $\{r_p(n)|n \geq 0\}$ and $\{q_p(n)|n \geq 0\}$ are P -recursive for $p = 2, 3, 4$.

TABLE 4.2(i)

The differential equation for the number of 4-regular simple labelled graphs.

i	$\phi_i(x)$
0	$-x^4(x^5 + 2x^4 + 2x^2 + 8x - 4)^2$
1	$-4(x^{13} + 4x^{12} - 16x^{10} - 10x^9 - 36x^8 - 220x^7 - 348x^6 - 48x^5 + 200x^4 - 336x^3 - 240x^2 + 416x - 96)$
2	$16x^2(x - 1)^2(x^5 + 2x^4 + 2x^2 + 8x - 4)(x + 2)^2$

$$R_4(x) = \sum_{n \geq 0} r_4(n) \frac{x^n}{n!} : \phi_2(x) \frac{d^2 R_4(x)}{dx^2} + \phi_1(x) \frac{dR_4(x)}{dx} + \phi_0(x) R_4(x) = 0$$

TABLE 4.2(ii)

The differential equation for the number of 4-regular labelled graphs.

i	$\phi_i(x)$
0	$x^{14} - 4x^{13} - 8x^{12} + 44x^{11} - 8x^{10} - 40x^9 - 244x^8 + 288x^7 + 192x^6 + 1056x^5 - 944x^4 - 2688x^3 + 448x^2 + 1408x + 384$
1	$-4(x^{13} - 4x^{12} + 8x^{10} + 22x^9 - 20x^8 - 92x^7 - 36x^6 + 48x^5 + 760x^4 - 464x^3 - 400x^2 + 160x + 96)$
2	$16x^2(x + 1)^2(x - 2)^2(x^5 - 2x^4 - 2x^2 + 8x + 4)$

$$Q_4(x) = \sum_{n \geq 0} q_4(n) \frac{x^n}{n!} : \phi_2(x) \frac{d^2 Q_4}{dx^2} + \phi_1(x) \frac{dQ_4}{dx} + \phi_0(x) Q_4 = 0$$

Using the recurrence equations implied by the above differential equations for R_3, R_4, Q_3, Q_4 , we have calculated $r_p(n)$ and $q_p(n)$ for $p = 3, 4$ and $n \leq 20$. These numbers are displayed in Tables A and B of the Appendix.

Read [4] has already given the differential equation associated with $\{r_3(n)|n \geq 0\}$. Read and Wormald [5] have given a system of simultaneous recurrence equations for $\{r_4(n)|n \geq 0\}$ and an inspection of these indicates that the P -recursiveness of $\{r_4(n)|n \geq 0\}$ may be deduced quite easily. The differential equations for $\{q_p(n)|n \geq 0\}$ for $p = 3$ and $p = 4$ appear to be new. We draw the reader's attention to the fact that the H -series theorem enables us to write down the system of partial differential equations for the H -series for arbitrary p without difficulty. However, the reduction of this system to a single ordinary differential equation in y_p is a technical task which we are unable to carry out for the general case.

5. A combinatorial construction. The differential equations for the H -series associated with p -regular simple labelled graphs may be given a direct combinatorial interpretation. This is achieved by distinguishing precisely k monovalent vertices for $k = 1, \dots, p$. This clearly involves a difficult case analysis, which is long even for the case $p = 3$. It is noteworthy that in this instance the H -series theorem carries out this case analysis automatically. In this section we give a combinatorial interpretation of system (4.1) for simple labelled 3-regular graphs.

Let \mathcal{A} be the set of simple labelled graphs whose vertices have degree at most 3. Then the power series $A^{(3)}(y_1, y_2, y_3)$ of § 4 is the exponential generating function for the elements of \mathcal{A} with y_i marking vertices of degree i for $i = 1, 2, 3$. Thus if $a(i_1, i_2, i_3)$ is the number of graphs in \mathcal{A} with i_j vertices of degree $j = 1, 2, 3$, then we have

$$A^{(3)}(y_1, y_2, y_3) = \sum_{i_1, i_2, i_3 \geq 0} a(i_1, i_2, i_3) \frac{y_1^{i_1}}{i_1!} \frac{y_2^{i_2}}{i_2!} \frac{y_3^{i_3}}{i_3!}.$$

The combinatorial derivations of (4.1a, b, c) are now given. To obtain these we count the graphs in \mathcal{A} once for each set of i distinct monovalent vertices for $i = 1, 2, 3$. For this purpose the i -set is regarded as being distinguished.

Equation (4.1a). Distinguish exactly one monovalent vertex in each element in \mathcal{A} . The generating function for this is $y_1 \partial A^{(3)} / \partial y_1$. We now derive this in another way.

1) The distinguished monovalent vertex is adjacent to a vertex of degree one, forming a component consisting of a single edge joining two vertices. The generating function for this is $y_1^2 A^{(3)}$.

2) The distinguished monovalent vertex is adjacent to a vertex of degree two. We may construct such graphs by distinguishing a monovalent vertex, v , and then connecting this by an edge to a new monovalent vertex u . Now u is the distinguished monovalent vertex adjacent to a bivalent vertex v . The generating function for this is $y_1 y_2 \partial A^{(3)} / \partial y_1$. We note that the operator $y_2 \partial / \partial y_1$ arises because a monovalent vertex is first distinguished and then connected to another vertex, making the former bivalent.

3) The distinguished monovalent vertex is adjacent to a vertex of degree three. Following 2), the generating function for this is $y_1 y_3 \partial A^{(3)} / \partial y_2$.

It follows that

$$\frac{\partial A^{(3)}}{\partial y_1} = y_1 A^{(3)} + y_2 \frac{\partial A^{(3)}}{\partial y_1} + y_3 \frac{\partial A^{(3)}}{\partial y_2},$$

and we have derived (4.1a) combinatorially.

Equation (4.1b). Distinguish two distinct monovalent vertices in each element in \mathcal{A} . The generating function for this is $(y_1^2/2!) \partial^2 A^{(3)}/\partial y_1^2$. We now derive this in another way.

1) The two distinguished vertices are connected by a path of edge-length 1, forming a component consisting of one edge. The generating function for this is $(y_1^2/2!)A^{(3)}$.

2) The two distinguished vertices are connected by a path of edge-length 2. There are two subcases.

i) The path contains exactly one bivalent vertex. The generating function for this is $(y_1^2/2!)y_2A^{(3)}$, since $(y_1^2/2!)y_2$ is the generating function for a component consisting of a path of edge-length two.

ii) The path contains exactly one trivalent vertex. Such graphs may be obtained by joining a distinguished monovalent vertex in an element of \mathcal{A} to two new monovalent vertices, which are themselves the distinguished monovalent vertices in the resulting graph. The generating function for this is $(y_1^2/2!)y_3 \partial A^{(3)}/\partial y_1$.

The generating function for these two cases is therefore

$$\frac{1}{2!} y_1^2 y_2 A^{(3)} + \frac{1}{2!} y_1^2 y_3 \frac{\partial A^{(3)}}{\partial y_1}.$$

3) We may obtain the remaining such graphs by deleting from a graph in \mathcal{A} a vertex, u , of degree 2 connected to distinct vertices a and b and connecting a distinguished isolated vertex a' to a and a distinguished isolated vertex b' to b . The vertices a' and b' are the distinguished monovalent vertices and are not connected by a path of edge-length 1 or 2. The generating function for this is

$$2 \frac{y_1^2}{2!} \frac{\partial A^{(3)}}{\partial y_2}$$

since a', b' may be labelled in two ways.

It follows that

$$\frac{\partial^2 A^{(3)}}{\partial y_1^2} = 2 \frac{\partial A^{(3)}}{\partial y_2} + A^{(3)} + y_2 A^{(3)} + y_3 \frac{\partial A^{(3)}}{\partial y_1},$$

and we have derived Equation (4.1b) combinatorially.

Equation (4.1c). Distinguish three distinct monovalent vertices in each element in \mathcal{A} . The generating function for this is $(y_1^3/3!) \partial^3 A^{(3)}/\partial y_1^3$. We now derive this in another way.

1) Exactly two of the distinguished vertices are joined by a path of edge-length one. We may construct such graphs by joining two isolated vertices, u and v , by an edge and by distinguishing one monovalent vertex w in a graph in \mathcal{A} . The generating function for this is $(y_1^2/2!)y_1 \partial A^{(3)}/\partial y_1$.

2) At least one pair of distinguished vertices are joined by a path of edge-length two. There are three subcases.

i) All three distinguished vertices are joined by paths of edge-length exactly two. Thus the distinguished vertices are the monovalent vertices of a component whose remaining vertex has degree three. The generating function for this is therefore $(y_1^3/3!)y_3 A^{(3)}$, since the component may be adjoined to any element in \mathcal{A} .

ii) Exactly two of the distinguished vertices are joined by a path of edge-length equal to two. There are two subcases.

a) The path contains a bivalent vertex. Such graphs may be constructed from a path of edge-length two joining two distinguished vertices and a graph in \mathcal{A} with exactly one distinguished monovalent vertex. The generating function for this is $y_2(y_1^2/2!)y_1 \partial A^{(3)}/\partial y_1$.

b) The path contains a vertex of degree three. We may construct such graphs by considering a path uvw , of edge-length two, and a graph in \mathcal{A} with exactly two distinct distinguished monovalent vertices a and b , separated by more than one edge. The vertices v and a are now identified, and u, w and b are the distinct distinguished vertices of the resulting graph. The generating function due to all graphs in \mathcal{A} treated in this way is $(y_1^2/2!)y_1y_3 \partial^2 A^{(3)}/\partial y_1^2$. But this set includes graphs in which two distinguished monovalent vertices are separated by a single edge, and hence form a component, enumerated by y_1^2 , adjoined to an element of \mathcal{A} , enumerated by $A^{(3)}$. When this is treated in the above manner, the generating function is $(y_1^2/2!)y_3y_1A^{(3)}$, so the contribution of this case is

$$\frac{1}{2!} y_1^3 y_3 \left(\frac{\partial^2 A^{(3)}}{\partial y_1^2} - A^{(3)} \right).$$

3) No pairs of the distinguished monovalent vertices are joined by paths of edge-length at least one or two. We may construct such graphs by deleting from a graph in \mathcal{A} a vertex of degree three connected to vertices a, b and c and connecting a to a', b to b' and c to c' , where a', b', c' are isolated vertices. In the resulting graph, a', b', c' are the distinguished vertices. The generating function for this is $y_1^3 \partial A^{(3)}/\partial y_3$ since a', b', c' may be labelled in $3!$ ways.

It follows that

$$\begin{aligned} \frac{\partial^3 A^{(3)}}{\partial y_1^3} &= 3(1 + y_2) \frac{\partial A^{(3)}}{\partial y_1} + 3y_3 \frac{\partial^2 A^{(3)}}{\partial y_1^2} - 2y_3 A^{(3)} + 6 \frac{\partial A^{(3)}}{\partial y_3} \\ &= \frac{\partial}{\partial y_1} \left\{ 3(1 + y_2) A^{(3)} + 3y_3 \frac{\partial A^{(3)}}{\partial y_1} \right\} - 2y_3 A^{(3)} + 6 \frac{\partial A^{(3)}}{\partial y_3} \\ &= -3 \frac{\partial}{\partial y_1} \left\{ 2 \frac{\partial A^{(3)}}{\partial y_2} - \frac{\partial^2 A^{(3)}}{\partial y_1^2} \right\} - 2y_3 A^{(3)} + 6 \frac{\partial A^{(3)}}{\partial y_3} \end{aligned}$$

from (4.1b). Thus

$$\frac{\partial^3 A^{(3)}}{\partial y_1^3} + 3 \frac{\partial A^{(3)}}{\partial y_3} = y_3 A^{(3)} + 3 \frac{\partial^2 A^{(3)}}{\partial y_1 \partial y_2},$$

and we have combinatorially derived (4.1c). This completes the combinatorial treatment of system (4.1).

6. Bisymmetric H -series. In the final part of this paper we consider the extension of the H -series theorem to the bisymmetric case. As an application of this extension we enumerate $n \times n$ matrices over the nonnegative integers with line sum p (each row sum and column sum equals p) for $p = 2, 3$.

Let $\mathbf{r} = (r_1, r_2, \dots)$ and $\mathbf{c} = (c_1, c_2, \dots)$ be sets of indeterminates, and let

$$T(\mathbf{r}, \mathbf{c}) = \sum_{\mathbf{i}, \mathbf{j} \geq \mathbf{0}} c(\mathbf{i}, \mathbf{j}) A_{\mathbf{i}}(\mathbf{r}) A_{\mathbf{j}}(\mathbf{c}),$$

where $A_{\mathbf{i}}$ is the monomial symmetric function defined in § 2. Then the H -series of T is

$$(H(T))(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{i}, \mathbf{j} \geq \mathbf{0}} c(\mathbf{i}, \mathbf{j}) (\tau(\mathbf{i})! \tau(\mathbf{j})!)^{-1} \mathbf{x}^{\tau(\mathbf{i})} \mathbf{y}^{\tau(\mathbf{j})}.$$

Let T be the ordinary generating function for nonnegative integer matrices, with r_i and c_j marking the sums of the elements in row i and column j respectively. Now a k in row i and column j contributes k to the i th row sum and the j th column sum so the generating function for the (i, j) -element is $1 + (r_i c_j) + (r_i c_j)^2 + \dots$ whence

$$T(\mathbf{r}, \mathbf{c}) = \prod_{i,j \geq 1} (1 - r_i c_j)^{-1}.$$

Clearly T is bisymmetric since it is symmetric in \mathbf{r} and in \mathbf{c} . Let $s_k = r_1^k + r_2^k + \dots$ and $t_k = c_1^k + c_2^k + \dots$ for $k \geq 1$, the power sum symmetric functions for \mathbf{r} and \mathbf{c} , respectively. Now

$$T(\mathbf{r}, \mathbf{c}) = \exp \sum_{i,j \geq 1} \log (1 - r_i c_j)^{-1} = \exp \sum_{k \geq 1} \frac{1}{k} \sum_{i,j \geq 1} r_i^k c_j^k,$$

so

$$G(\mathbf{s}, \mathbf{t}) = \exp \sum_{k \geq 1} \frac{1}{k} s_k t_k.$$

The system of differential equations satisfied by $G(\mathbf{s}, \mathbf{t})$ is

$$(6.1) \quad \frac{\partial G}{\partial s_k} = \frac{1}{k} t_k G \quad \text{for } k \geq 1, \quad \frac{\partial G}{\partial t_k} = \frac{1}{k} s_k G \quad \text{for } k \geq 1.$$

We illustrate the use of the H -series theorem in the bisymmetric case by applying it to nonnegative integer matrices with line sum two in this section and line sum three in § 7. Let $l_p(n)$ be the number of $n \times n$ nonnegative integer matrices with line sum p . Then

$$l_2(n) = \left[\frac{x_2^n y_2^n}{n! n!} \right] D^{(2)}(x_1, x_2, y_1, y_2),$$

where $D^{(2)}(x_1, x_2, y_1, y_2) = (H(T))(x_1, x_2, 0, \dots, y_1, y_2, 0, \dots)$. Applying the H -series theorem to system (6.1) and setting $x_3 = x_4 = \dots = 0$ and $y_3 = y_4 = \dots = 0$ we have the following system of equations for $D^{(2)}(x_1, x_2, y_1, y_2)$.

$$(6.2) \quad \begin{aligned} \text{a) } \frac{\partial D^{(2)}}{\partial x_1} &= y_1 D^{(2)} + y_2 \frac{\partial D^{(2)}}{\partial y_1}, & \text{a)' } \frac{\partial D^{(2)}}{\partial y_1} &= x_1 D^{(2)} + x_2 \frac{\partial D^{(2)}}{\partial x_1}, \\ \text{b) } \frac{\partial D^{(2)}}{\partial x_2} - \frac{1}{2} \frac{\partial^2 D^{(2)}}{\partial x_1^2} &= \frac{1}{2} y_2 D^{(2)}, & \text{b)' } \frac{\partial D^{(2)}}{\partial y_2} - \frac{1}{2} \frac{\partial^2 D^{(2)}}{\partial y_1^2} &= \frac{1}{2} x_2 D^{(2)}. \end{aligned}$$

From (6.2a) and (6.2a)' we obtain

$$(6.3) \quad \frac{\partial D^{(2)}}{\partial x_1} = (y_1 + y_2 x_1)(1 - x_2 y_2)^{-1} D^{(2)}.$$

Differentiating (6.2b)' partially with respect to x_2 we have

$$2 \frac{\partial^2 D^{(2)}}{\partial x_2 \partial y_2} = \frac{\partial^3 D^{(2)}}{\partial y_1^2 \partial x_2} + x_2 \frac{\partial D^{(2)}}{\partial x_2} + D^{(2)}.$$

Eliminating $\partial D^{(2)}/\partial x_2$ from the right-hand side of this equation by means of (6.2b) we have

$$(6.4) \quad 4 \frac{\partial^2 D^{(2)}}{\partial x_2 \partial y_2} = \frac{\partial^4 D^{(2)}}{\partial x_1^2 \partial y_1^2} + y_2 \frac{\partial^2 D^{(2)}}{\partial y_1^2} + x_2 \frac{\partial^2 D^{(2)}}{\partial x_1^2} + x_2 y_2 D^{(2)} + 2D^{(2)}.$$

We wish to eliminate x_1 and y_1 . Thus let $E^{(2)}(x_2, y_2) = D^{(2)}(0, x_2, 0, y_2)$ so that, from (6.3),

$$\left. \frac{\partial^2 D^{(2)}}{\partial x_1^2} \right|_{x_1=y_1=0} = y_2(1-x_2y_2)^{-1}E^{(2)}.$$

From (6.2a)' we have

$$\left. \frac{\partial^4 D^{(2)}}{\partial x_1^2 \partial y_1^2} \right|_{x_1=y_1=0} = 2(1+x_2y_2)(1-x_2y_2)^{-2}E^{(2)}$$

and

$$\left. \frac{\partial^2 D^{(2)}}{\partial y_1^2} \right|_{x_1=y_1=0} = x_2(1-x_2y_2)^{-1}E^{(2)}.$$

Substituting these expressions into (6.4) and simplifying we obtain

$$(4-8x_2y_2+4x_2^2y_2^2) \frac{\partial^2 E^{(2)}}{\partial x_2 \partial y_2} = (4-2x_2^2y_2^2+x_2^3y_2^3)E^{(2)}.$$

But a matrix with line sum 2 must be square, so $E^{(2)}(x_2, y_2) = M^{(2)}(x_2y_2)$, where $M^{(2)}(z) = \sum_{n \geq 0} l_2(n) z^n / (n!)^2$.

Thus $M^{(2)}(z)$ satisfies the differential equation

$$4z(1-z)^2 \frac{d^2}{dz^2} M^{(2)}(z) + 4(1-z)^2 \frac{d}{dz} M^{(2)}(z) - (4-2z^2+z^3)M^{(2)}(z) = 0,$$

so $M^{(2)}(z)$ is D -finite and $\{l_2(n) | n \geq 0\}$ is P -recursive. By inspection this equation may be rewritten as

$$\left\{ 2z(1-z) \frac{d}{dz} + 2 + 2z - z^2 \right\} G(z) = 0, \quad \text{where } G(z) = \left\{ 2(1-z) \frac{d}{dz} - (2-z) \right\} M^{(2)}(z).$$

But $G(z)$ is a formal power series with no negative exponents so $G(z) = 0$, yielding the recurrence equation

$$l_2(n+1) = (n+1)^2 l_2(n) - \frac{1}{2} n^2 (n+1) l_2(n-1)$$

for $n \geq 0$, where $l_2(0) = 1, l_2(-1) = 0$. This simplifies the recurrence equation given by Anand, Dumir and Gupta [1]. The differential equation may be solved to give

$$G(z) = (1-z)^{-1/2} \exp\left(\frac{z}{2}\right),$$

which may be obtained immediately by a combinatorial construction involving cycles.

7. Nonnegative integer matrices with line sum 3. Now

$$l_3(n) = \begin{bmatrix} x_3^n & y_3^n \\ n! & n! \end{bmatrix} D^{(3)}(x_1, x_2, x_3, y_1, y_2, y_3),$$

where $D^{(3)}(x_1, x_2, x_3, y_1, y_2, y_3) = (H(T))(x_1, x_2, x_3, 0, \dots, y_1, y_2, y_3, 0, \dots)$, and T is given in § 6. Following § 6 we apply the H -series theorem and set $x_4 = x_5 = \dots = 0$

and $y_4 = y_5 = \dots = 0$ to obtain the following system of equations for $D^{(3)}$.

$$\begin{aligned}
 (7.1) \quad & \text{a)} \quad \frac{\partial D^{(3)}}{\partial x_1} = y_1 D^{(3)} + y_2 \frac{\partial D^{(3)}}{\partial y_1} + y_3 \frac{\partial D^{(3)}}{\partial y_2}, \\
 & \text{a)'} \quad \frac{\partial D^{(3)}}{\partial y_1} = x_1 D^{(3)} + x_2 \frac{\partial D^{(3)}}{\partial x_1} + x_3 \frac{\partial D^{(3)}}{\partial x_2}, \\
 & \text{b)} \quad \frac{\partial D^{(3)}}{\partial x_2} - \frac{1}{2} \frac{\partial^2 D^{(3)}}{\partial x_1^2} = \frac{1}{2} y_2 D^{(3)} + \frac{1}{2} y_3 \frac{\partial D^{(3)}}{\partial y_1}, \\
 & \text{b)'} \quad \frac{\partial D^{(3)}}{\partial y_2} - \frac{1}{2} \frac{\partial^2 D^{(3)}}{\partial y_1^2} = \frac{1}{2} x_2 D^{(3)} + \frac{1}{2} x_3 \frac{\partial D^{(3)}}{\partial x_1}, \\
 & \text{c)} \quad \frac{\partial D^{(3)}}{\partial x_3} - \frac{\partial^2 D^{(3)}}{\partial x_1 \partial x_2} + \frac{1}{3} \frac{\partial^3 D^{(3)}}{\partial x_1^3} = \frac{1}{3} y_3 D^{(3)}, \\
 & \text{c)'} \quad \frac{\partial D^{(3)}}{\partial y_3} - \frac{\partial^2 D^{(3)}}{\partial y_1 \partial y_2} + \frac{1}{3} \frac{\partial^3 D^{(3)}}{\partial y_1^3} = \frac{1}{3} x_3 D^{(3)}.
 \end{aligned}$$

By inspection we may express $\partial^{2i} D^{(3)} / \partial x_3^i \partial y_2^i$ at $x_2 = 0, y_2 = 0$ linearly in terms of $D^{(3)}, \partial D^{(3)} / \partial x_1, \partial D^{(3)} / \partial y_1, \partial^2 D^{(3)} / \partial x_1 \partial y_1$ at $x_2 = y_2 = 0$ for $i \geq 1$. Moreover, when we carry this out for $i = 1, 2$ (again using VAXIMA) and set $x_1 = y_1 = 0$, we discover that the coefficients of $\partial D^{(3)} / \partial x_1$ and $\partial D^{(3)} / \partial y_1$ are 0 in both equations. Eliminating $\partial^2 D^{(3)} / \partial x_1 \partial y_1$ at $x_1 = x_2 = y_1 = y_2 = 0$ between these two equations, we get a linear equation involving $\partial^4 D^{(3)} / \partial x_3^2 \partial y_3^2, \partial^2 D^{(3)} / \partial x_3 \partial y_3$ and $D^{(3)}$, all at $x_1 = x_2 = y_1 = y_2 = 0$. But $D^{(3)}(0, 0, x_3, 0, 0, y_3) = E^{(3)}(x_3 y_3)$, where

$$E^{(3)}(z) = \sum_{n \geq 0} l_3(n) \frac{z^n}{(n!)^2}.$$

Finally, this partial differential equation for $E^{(3)}(x_3 y_3)$ can be transformed to a fourth-order ordinary differential equation for $E^{(3)}(x)$, with polynomial coefficients in x , by making the substitution $x = x_3 y_3$. This differential equation is displayed in Table 7.1.

We therefore have the following result.

COROLLARY.

$\{l_3(n) | n \geq 0\}$ is P-recursive.

TABLE 7.1

The differential equation for the number of nonnegative integer matrices with line sum 3.

i	ϕ_i
0	$x^{11} - 7x^{10} + 30x^9 - 16x^8 - 43x^7 + 51x^6 + 238x^5 + 630x^4 + 36x^3 - 1944x^2 - 1152x + 576$
1	$-9(x^{10} - 4x^9 + 22x^8 - 8x^7 - 4x^6 + 8x^5 + 88x^4 + 252x^3 + 120x^2 - 320x + 64)$
2	$-9(x^{10} - 4x^9 + 22x^8 - 8x^7 - 22x^6 + 8x^5 + 106x^4 + 234x^3 + 48x^2 - 320x + 64)x$
3	$324x^4(x^4 - x^2 + x + 4)$
4	$81x^5(x^4 - x^2 + x + 4)$

$$\begin{aligned}
 E^{(3)}(x) = \sum_{n \geq 0} l_3(n) \frac{x^n}{(n!)^2} : \phi_4(x) \frac{d^4 E^{(3)}}{dx^4} + \phi_3(x) \frac{d^3 E^{(3)}}{dx^3} + \phi_2(x) \frac{d^2 E^{(3)}}{dx^2} \\
 + \phi_1(x) \frac{dE^{(3)}}{dx} + \phi_0(x) E^{(3)} = 0
 \end{aligned}$$

This appears to be a new result (see Stanley [6, p. 186]). The recurrence for $\{l_3(n) | n \geq 0\}$ which follows from the equation in Table 7.1 has been used to compute $l_3(n)$ for $n \leq 15$. These numbers are given in Table C of the Appendix.

8. Concluding comments. Each of the differential equations displayed in tables in this paper was obtained by using the symbolic algebra system called VAXIMA. The elimination procedures for R_4 , Q_4 and $E^{(3)}$ were so substantial that we could not have carried them out by hand. Each of the tables given in the Appendix was computed from the corresponding differential equation by means of VAXIMA. The computer calculations were carried out at the University of Waterloo. VAXIMA is based on the MACSYMA system developed at the Massachusetts Institute of Technology.

Appendix.

TABLE A
Numbers of 3-regular simple labelled graphs (i) and labelled graphs (ii).

n	$r_3(n)$	$q_3(n)$
0	1	1
2	0	2
4	1	47
6	70	4720
8	19355	1256395
10	11180820	699971370
12	11555272575	706862729265
14	19506631814670	1173744972139740
16	50262958713792825	2987338986043236825
18	187747837889699887800	11052457379522093985450
20	976273961160363172131825	5703510582280129537568575
	(i)	(ii)

TABLE B
Numbers of 4-regular simple labelled graphs (i) and labelled graphs (ii).

n	$r_4(n)$	$q_4(n)$
0	1	1
1	0	1
2	0	3
3	0	15
4	0	138
5	1	2021
6	15	43581
7	465	1295493
8	19355	50752145
9	1024380	2533755933
10	66462606	157055247261
11	5188453830	11836611005031
12	480413921130	1066129321651668
13	52113376310985	113117849882149725
14	6551246596501035	13965580274228976213
15	945313907253606891	1985189312618723797371
16	155243722248524067795	321932406123733248625851
17	28797220460586826422720	59079829666712346141491403
18	5993002310427150494060340	12182062872168618012045410805
19	1390759561507559001823665540	2804416350168401031334025488653
20	357920518512934324278467820756	716675823235860386364568072658826
	(i)	(ii)

TABLE C
 Numbers of $n \times n$ nonnegative integer matrices with line sum 3.

n	$l_3(n)$
0	1
1	1
2	4
3	55
4	2008
5	153040
6	20933840
7	4662857360
8	1579060246400
9	772200774683520
10	523853880779443200
11	477360556805016931200
12	569060910292172349004800
13	868071731152923490921728000
14	1663043727673392444887284377600
15	3937477620391471128913917360384000

REFERENCES

- [1] H. ANAND, V. C. DUMIR AND H. GUPTA, *A combinatorial distribution problem*, Duke Math. J., 33 (1966), pp. 757–770.
- [2] J. HAMMOND, *On the use of certain differential operators in the theory of equations*, Proc. Lond. Math. Soc., 14 (1883), pp. 119–129.
- [3] P. A. MACMAHON, *Combinatory Analysis*, Vol. 1 and 2, Chelsea, New York, 1960.
- [4] R. C. READ, *The enumeration of locally restricted graphs* (II), J. Lond. Math. Soc., 35 (1960), pp. 344–351.
- [5] R. C. READ AND N. C. WORMALD, *Number of labelled 4-regular graphs*, J. Graph Theory, 4 (1980), pp. 203–212.
- [6] R. P. STANLEY, *Differentiably finite power series*, European J. Combinatorics, 1 (1980), pp. 175–188.

TRADITIONAL GALLERIES REQUIRE FEWER WATCHMEN*

J. KAHN[†], M. KLAWE[‡] AND D. KLEITMAN[†]

Abstract. Chvátal's watchman theorem shows if the walls of an art gallery form an n -sided polygon then at most $\lceil n/3 \rceil$ watchmen are needed to guard it, and that this number is best possible. In this paper it is shown that if every pair of adjacent sides of the polygon form a right angle then at most $\lceil n/4 \rceil$ guards are needed, and again this result is best possible. Our proof depends on showing that any finite region bounded by a finite number of edges, each of which lies parallel to one of a fixed pair of perpendicular axes, has a partition into convex quadrilaterals.

1. Introduction. The following question is due to Victor Klee. Suppose the walls of an art gallery form an n -sided polygon. How many guards are needed so that every wall is seen by some guard, where it is assumed that the guards are stationary but can see in all directions? In 1975 Vaclav Chvátal [1] showed that $\lceil n/3 \rceil$ are always sufficient, and moreover there are galleries which require at least this number (see Fig. 1). In 1979 Steve Fisk [2] gave a very short proof of this result which goes as follows. Let G be a graph which is obtained from triangulating the polygon. It is well known that any such graph can be three-colored. In any such coloring one of the colors, say green, is used at most $\lceil n/3 \rceil$ times. If a guard is placed at every vertex colored green, then it is easy to see that every wall is seen by some guard since every wall is in some triangle and every triangle has one green vertex.

In this paper we consider art galleries whose walls form an n -sided right-angled polygon, i.e., a polygon in which all angles are right angles. Figure 2 shows such a gallery requiring $\lceil n/4 \rceil$ guards. We will prove that this number is always sufficient. The crux of the argument depends on showing that right-angled polygons can be convexly quadrilateralized. In other words the interior can be partitioned into convex quadrilaterals by adding nonintersecting lines between vertices. An example is shown in Fig. 3(a). If G is the graph obtained from a convex quadrilateralization of any polygon by adding the pair of diagonals to each quadrilateral (see Fig. 3(b)), then it is easy to prove that G can be four-colored by induction on n . Now because the

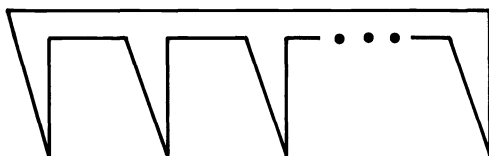


FIG. 1

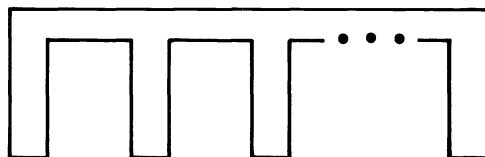


FIG. 2

* Received by the editors November 25, 1980, and in final form March 8, 1982.

[†] Mathematics Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

[‡] Computer Science Department, IBM Research Laboratory, San Jose, California 95193.

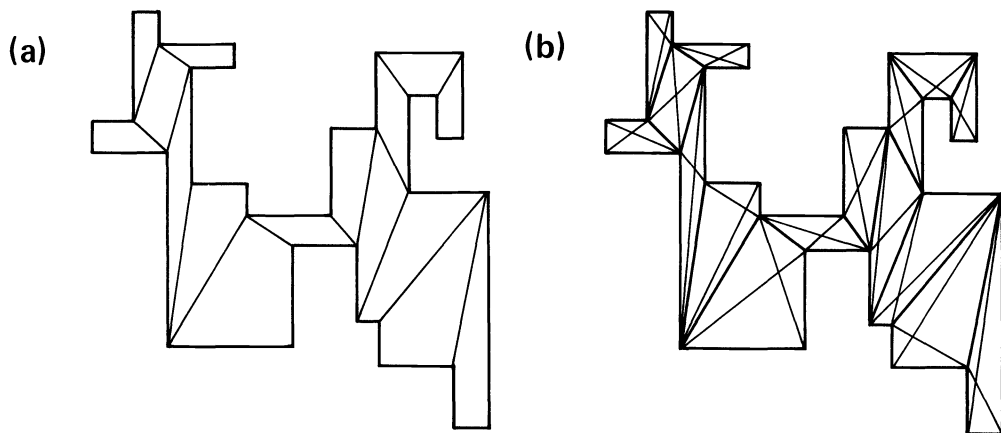


FIG. 3

quadrilaterals are convex, placing guards at each vertex colored by the least frequently used color, completes the solution.

At this point we should point out that in our proof the polygons are assumed to be nonself-intersecting though it is clear that the results also hold for cases where intersections occur (see Fig. 4 for example) since small perturbations can be used to reduce these to the nonintersecting case. The same however is not true when we consider art galleries whose interior is not simply connected. Figures 5(a) and (b) show that $\lfloor n/3 \rfloor$ and $\lfloor n/4 \rfloor$ guards respectively are no longer sufficient in these cases. Notice that although the gallery depicted by Fig. 5(b) is right-angled it is not true that all walls lie parallel to some fixed pair of perpendicular axes. Let us call regions, whose boundaries consist of line segments parallel to some fixed pair of perpendicular axes, rectilinear. We have been able to prove that every rectilinear region with a finite number of edges can be convexly quadrilateralized. Unfortunately it is not true that the graph obtained by “completing” the quadrilateralization is always four-colorable as can be seen by examining the graph in Fig. 5(c). Thus the question of whether $\lfloor n/4 \rfloor$ guards are always sufficient to guard a nonsimply connected rectilinear gallery with n edges remains open.

The remainder of this paper is devoted to proving the existence of convex quadrilateralizations for finite rectilinear regions. It is convenient for the argument to extend the result to the situation in which the region lies not in the plane, but in

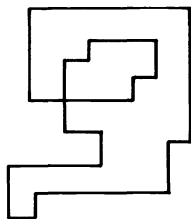


FIG. 4

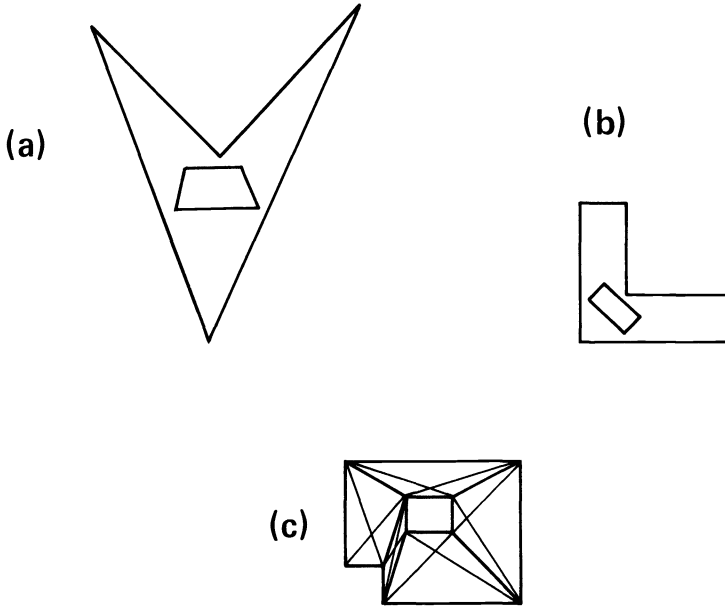


FIG. 5

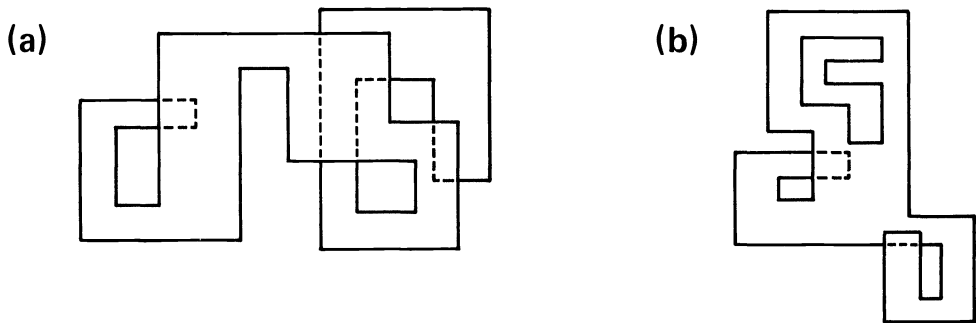


FIG. 6

any Riemann surface corresponding to a function with branch points outside itself. This corresponds in the gallery case to having several levels with ramps leading from one to another. Some examples are shown in Figs. 6(a) and (b).

We therefore prove:

THEOREM. *If R is a closed region bounded by a finite number of straight edges, each parallel to either of two perpendicular axes in a Riemann surface corresponding to a function with singularities outside R , then R has a convex quadrilateralization.*

The proof is inductive in that we will show that if every “smaller” region than R has a convex quadrilateralization then R does too. Our notion of smaller is defined

as follows. A region S is smaller than R if the number of connected components of the complement of S is less than the corresponding number for R or if these two numbers are equal and S has strictly fewer vertices than R . Thus for example any simply connected region is always smaller than a nonsimply connected region, and when restricted to simply connected regions, smaller just means fewer vertices.

We will think of the edges of R as being parallel to either the x or y axis and hence refer to them as horizontal or vertical edges. Similarly we will refer to horizontal and vertical coordinates of points and edges of R .

Remark. We may assume that no two horizontal [vertical] edges in R have the same vertical [horizontal] coordinate.

To see this, notice that for any general finite rectilinear region R , one may take a sequence of finite rectilinear regions which converges to R , such that none of these regions have “offending” edges. Moreover since there are only finitely many different quadrilateralizations, we may assume that each region in the sequence has the same convex quadrilateralization. Finally since convexity is closed under taking limits, this quadrilateralization must also be a convex quadrilateralization of R .

Let us say that a finite rectilinear region is reducible if whenever every smaller finite rectilinear region is convexly quadrilateralizable then so is R .

The inductive argument falls into two parts. First we show that any region R having one of several configurations is reducible. We then conclude the proof by showing that every finite rectilinear region has at least one of these configurations.

2. Reductions. In this section after introducing the necessary terminology, we will give reductions for rectilinear regions which contain any one of three configurations. In many places in this paper our proofs depend on certain configurations having particular properties which appear to follow obviously from the relevant definitions. In fact the first lemma is a case in point. However, as is often true in geometrical problems, despite their “obvious truth” it seems to be both time-consuming and tricky to provide rigorous proofs of these assertions. Rather than distract the reader with the details of the proofs now, here we will simply state these observations, leaving the proofs for the next section where we will develop sufficient machinery to cope with the problem.

We begin by defining notation to describe features of rectilinear regions.

Edges are of four kinds, which we shall call *top*, *bottom*, *left* and *right*, referring to how they bound the region. Thus a top edge means one which bounds the region from above. For a region R we use $\text{int}(R)$ to denote the interior of R , i.e., those points lying inside R but not on any edge of R . We will also refer to the complement of R as the exterior of R . If x and y are points in R we use $[x, y]$ to denote the line segment joining x and y , whereas (x, y) denotes that line segment without its endpoints and $[x, y)$ and $(x, y]$ denote the appropriate half-closed line segments.

We say that two points x and y are *visible* to one another or that x sees y , if (x, y) is contained in $\text{int}(R)$. Similarly two edges S and T see each other if there exist points x and y on S and T respectively, which see each other. For points or horizontal edges we will often use expressions such as x “is higher than” y to indicate that x ’s vertical coordinate is greater than y ’s. Likewise when referring to horizontal coordinates we will say things like x “is to the left of” y .

A top edge S and a bottom edge T which see each other are *neighbors* if S is higher than T and there is no bottom [top] edge visible to S [T] whose vertical coordinate lies between those of S and T . A *tab* is a pair of neighboring edges which are connected to each other by a vertical edge. We say that a tab is a *down-tab* if its top edge extends further, either to the left or right, than the bottom edge. Otherwise

it is an *up-tab*. In Fig. 7, (i) is a pair of neighboring edges which is not a tab, (ii) is a down-tab and (iii) is an up-tab.

If x is either a point or vertical edge and y is either a point or horizontal edge, we will use $x \# y$ to denote the point whose horizontal coordinate is that of x and vertical coordinate is that of y .

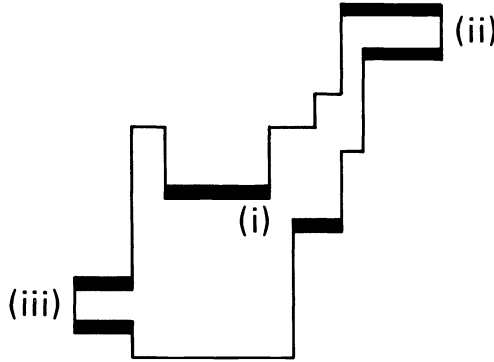


FIG. 7

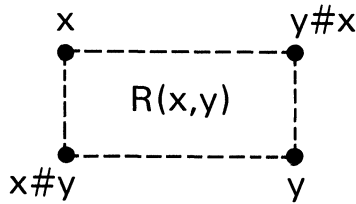


FIG. 8

Given two points x and y , we will denote the rectangle with corners $x, y, x \# y$ and $y \# x$ by $R(x, y)$. Figure 8 illustrates these definitions.

Our first lemma describes properties of a pair of neighboring edges.

LEMMA 2.1. *Let T and S be neighboring edges, where T is the top edge. Then there is a vertical edge M at least as far left as the left endpoints of T and S , whose top [bottom] endpoint is at least as high [low] as T [S]. Similarly there is a vertical edge N on the right of T and S with analogous properties. Moreover the interior of $R(M \# T, N \# S)$, i.e., the rectangle spanned by M, N, T and S , is completely contained in the interior of R .*

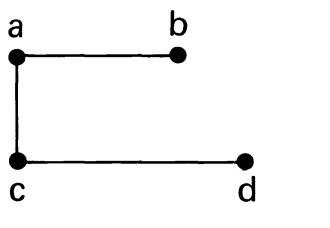


FIG. 9

The proof is given in the next section (see Lemma 3.5). In the case that T and S actually form a tab, then either M or N joins T to S and is called the *side edge* of the tab; the other is called the *facing edge* of the tab.

We will find it useful to note the following remark:

LEMMA 2.2. *If $[a, b]$ and $[c, d]$ form the horizontal edges of a tab then any convex quadrilateralization Q of R must include the quadrilateral with corners a, b, c, d .*

Proof. Let $[a, b]$ be the top edge, $[c, d]$ the bottom edge, and assume that they are joined by the edge $[a, c]$. See Fig. 9 for illustration. Then by Lemma 2.1 and the assumption that no two horizontal edges have the same vertical coordinate, every vertex of R other than d which is visible to a must lie lower than d . This implies that a can be connected in Q by edges only to b and c and possibly other vertices lower than d . If it is connected by an edge $[a, x]$ of this latter kind in Q , then c can only be connected to a and d in Q and cannot possibly be part of a convex quadrilateral. Thus the only edges of Q containing a are $[a, b]$ and $[a, c]$; likewise the only edges containing c are $[a, c]$ and $[c, d]$. Thus one must have the quadrilateral (a, b, c, d) .

We may now note the following reduction.

LEMMA 2.3. *If R possesses a pair of neighboring edges which are not a tab, then R is reducible.*

Proof. Assume every region smaller than R has a convex quadrilateralization. Let the top edge of the neighboring pair be $[a, b]$, with a to the left of b , and the bottom edge be $[c, d]$ with c to the left of d . Suppose for example that c is to the right of b ; all other cases may be handled similarly. Let b' be the point $d \# b$, and let c' be the point $a \# c$.

Consider the multilevel region, R' , obtained by replacing $[a, b]$ and $[c, d]$ by two tabs, one having edges $[a, b']$, $[b', d]$, $[c, d]$, the other having edges $[a, b]$, $[a, c']$, $[c', d]$. Since the insides of these overlap we can imagine one forming a dead-end up-ramp and the other a similar down-ramp. See Fig. 10 for an example.

If R' is disconnected then both connected components are smaller than R since they have fewer vertices. On the other hand if R' is connected then its complement

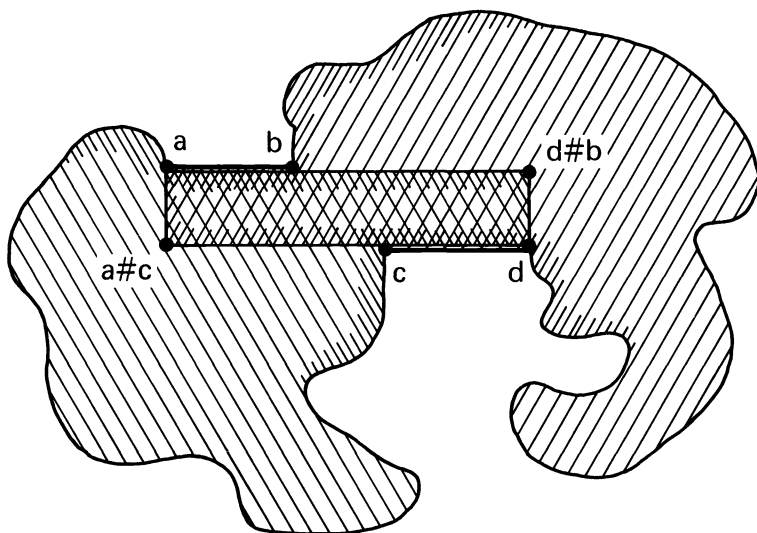


FIG. 10

has one less connected component than the complement of R . To see this consider two points x and y in the exterior of R , where x is just above T and y is just below S . Clearly x and y are in the same component of the exterior of R' , as they are connected by a path going around either one of the new tabs; but since R' is connected x and y must be in different connected components of the exterior of R .

In any case by our assumption R' has a convex quadrilateralization Q' , and by Lemma 2.2, (a, b', c, d) and (a, b, c', d) are quadrilaterals in Q' . Now replacing them by the quadrilateral (a, b, c, d) we obtain a convex quadrilateralization of R . \square

Before describing the next reduction, we must introduce a little more terminology about tabs. We define the *top-tip* and *bottom-tip* of a tab to be the endpoints of its top and bottom edges, respectively, that are closest to its facing edge. If Z is an up-tab [down-tab], then its *step-point* is the upper [lower] endpoint of its facing edge and its *step-edge* is the horizontal edge containing its step-point. We say that an up-tab Z [down-tab] is *bad* if its step-edge is a bottom [top] edge, and if no edges of R intersect the interior of the rectangle $R(s, t)$, where s and t are the step-point and top-tip [bottom-tip] of Z respectively. Naturally a *good tab* is any tab which is not bad.

We are now ready to give a reduction for regions containing a good tab.

LEMMA 2.4. *If R has a good tab Z , then R is reducible.*

Proof. We give the proof for the case that Z is an up-tab; the other case follows by a symmetric argument. Let $[a, b]$, $[a, c]$ and $[c, d]$ form the top, side and bottom edges of Z , respectively, and let s be the step-point of Z . Let $[x, y]$ be the lowest horizontal edge which intersects the interior of $R(b, s)$, and let e be the upper endpoint of the vertical edge containing b . We assume without loss of generality that $[a, c]$ is a left edge and that x lies to the left of y .

We must deal with two cases: whether x is to the left of b or not. The method of argument is of the same form as in Lemma 2.3: in either of the cases we replace R by a new region R' obtained by cutting R and splicing tabs on the wounds. Again by analogous arguments each of the connected components R' will be smaller than R yielding by inductive assumption a convex quadrilateralization Q' of R' , and as before we will be able to replace quadrilaterals of Q' by some convex quadrilateral to obtain a convex quadrilateralization of R .

It will be important to note that no edges intersect the interior of the rectangle $R(b \# c, s \# y)$. By Lemma 2.1, no edge intersects $\text{int}(R(b \# c, s \# b))$, and by the definition of $[x, y]$, no horizontal edge intersects $\text{int}(R(b, s \# y))$. Combining this with the fact that no other horizontal edge has the same horizontal coordinate as $[a, b]$ we see that we have shown that no horizontal edge intersects $\text{int}(R(b \# c, s \# y))$. However it is easy to see that no vertical edge may intersect either, since by 2.1 one of its endpoints (and hence a horizontal edge) would lie in $\text{int}(R(b, s \# y))$. It is this fact that allows us to conclude that the edges we add to obtain R' in the next paragraph form tabs.

When x is to the left of b , we replace the edges $[x, y]$, $[e, b]$, $[a, b]$, $[a, c]$ and $[c, d]$ by the tab $[b \# y, y]$, $[b \# y, b \# c]$, $[b \# c, d]$ on say a down-ramp and the "sideways" tab $[e, b]$, $[b, y \# b]$, $[y \# b, y]$ on an up-ramp along with $[x, y]$. Now any convex quadrilateralization of R' must contain the quadrilaterals $(b \# y, y, d, b \# c)$ and $(e, b, y \# b, y)$, and replacing them by (e, y, b, d) and adding (a, b, c, d) yields a convex quadrilateralization of R . This case is shown in Fig. 11.

When x is not to the left of b , we make the identical replacements. Let Q' be a convex quadrilateralization of R' . Again for the same reason as before, the first of these is a tab and we may conclude that $(b \# y, y, d, b \# c)$ is a convex quadrilateral in Q' . Now however it is not necessarily true that $(e, b, y \# b, y)$ be in Q' , since $[e, b]$

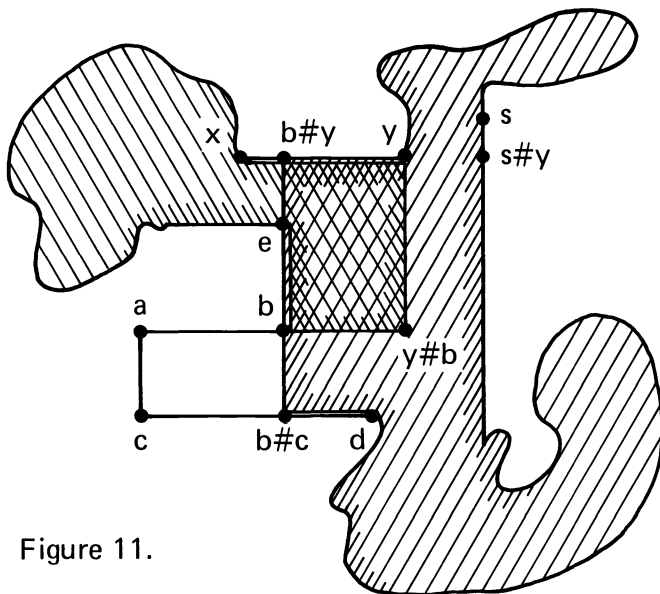


Figure 11.

FIG. 11

and $[y \# b, y]$ are no longer a “sideways” tab. However it is necessary that $y \# b$ lie in only one quadrilateral in Q' . Otherwise the line segment separating two of them would necessarily go either to the left of $[e, b]$, preventing any convex quadrilateral containing b , or above $[x, y]$, preventing any such quadrilateral from containing y , because of the fact that there are no edges (and hence no corners) of R in the interior of $R(b, y)$.

One can similarly show that Q' contains either $(e, b, y \# b, y)$ or $(b, x, y, y \# b)$. We may therefore replace this quadrilateral and $(b \# y, y, d, b \# c)$ by (e, b, d, y) or (b, x, y, d) respectively, then adding (a, b, c, d) to obtain a convex quadrilateralization of R . See Fig. 12 for a diagram of this case. \square

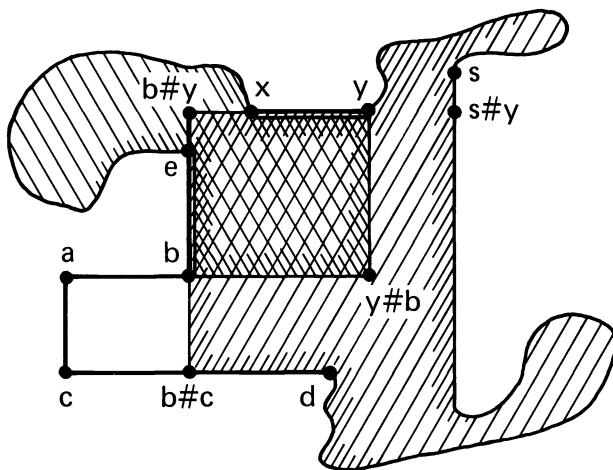


FIG. 12

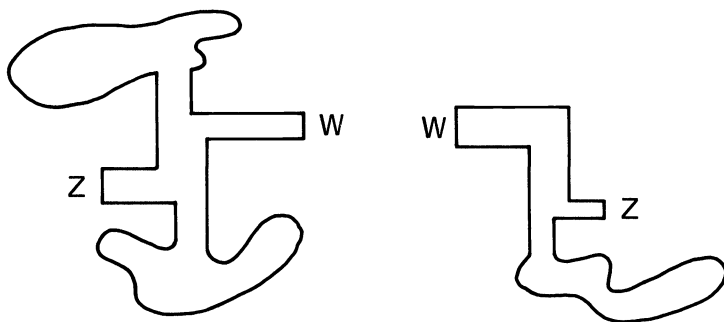


FIG. 13

We are now ready to present the final reduction after one more definition. We say that an up-tab Z and a down-tab W form a *tab-pair* if the step-edge of Z is the bottom edge of W and the step-edge of W is the top edge of Z . Some examples are shown in Fig. 13.

LEMMA 2.5. *If R contains a tab-pair, then R is reducible.*

Proof. Let $[a, b]$, $[c, d]$ and $[a, c]$ form the top, bottom and side edges of the up-tab, while $[h, i]$, $[f, g]$ and $[i, g]$ form the top, bottom and side edges of the down-tab. If we replace $[a, b]$, $[a, c]$, $[f, g]$ and $[i, g]$ by $[a \# f, f]$, $[a \# f, c]$, $[b, g \# b]$ and $[g \# b, i]$, we obtain a region R' whose connected components are smaller than R as before. By inductive assumption these have convex quadrilateralizations with quadrilaterals $(c, d, f, a \# f)$ and $(h, i, g \# b, b)$ which may be replaced by (a, b, c, d) , (g, h, i, f) and (b, d, f, h) to give a convex quadrilateralization of R . This is shown in Fig. 14. \square

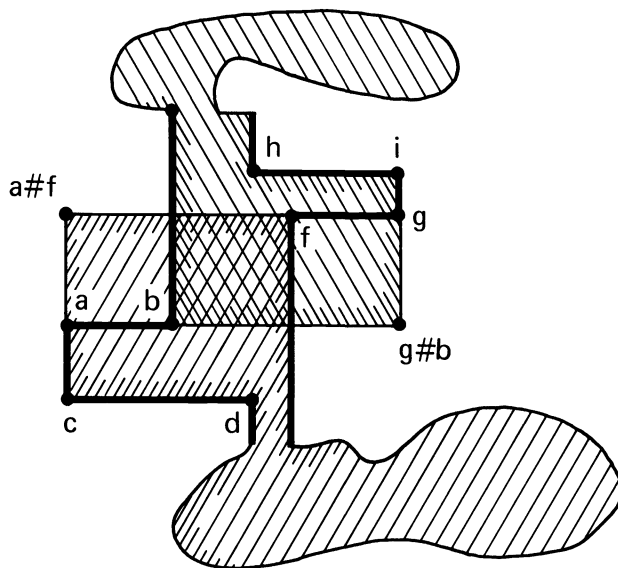


FIG. 14

3. Properties of rectilinear regions. In this section we will show that every finite rectilinear region must have one of the following three configurations: a pair of neighboring edges that do not form a tab, a good tab or a tab-pair. Let us say that a rectilinear region R is irreducible if it contains none of the above configurations. We will show that every irreducible rectilinear region has an infinite sequence of distinct horizontal edges. To begin with however, we will concentrate on proving some elementary facts about rectilinear regions, which will yield the results we needed in the preceding section.

First we must introduce some more notation. For any interior point w of R consider the vertical line L passing through w . It is easy to see that the first edge that L intersects above w must be a top edge (or possibly the corner of a top edge and a vertical edge). We call this top edge the *top-bounding* edge of w . Similarly we define *bottom-, left- and right-bounding* edges of w .

We will use expressions such as x lies *vertically above* an edge S to mean that some point of S has the same horizontal coordinate as x and is lower than x . Formally we should add here that the line joining x to S lies on the Riemann surface containing R ; “obvious” restrictions of this type will hereafter be omitted. Similarly x lies *horizontally between* a and b means that its horizontal coordinate lies between those of a and b .

One of the standard remarks that we often need to make is that no edges of R intersect some rectangle or more general region. Our first aim is to establish some simple results of this form that we can apply whenever necessary. The following lemma is the basic tool that we will use over and over again for many different purposes.

LEMMA 3.1. *Let S be a horizontal edge, z an interior point of R lying vertically above S and w a point of R at least as high as z such that w sees z . Then there is some bottom edge T , whose vertical coordinate is between those of S and w , which is visible to w .*

Proof. Choose T to be a horizontal edge such that some point v of $(w, z]$ lies vertically above some point u of T , and such that the distance from u to v is minimal. First note that T is a bottom edge since by its minimum distance property, T is the bottom-bounding edge of v . Thus since the vertical coordinate of T is clearly between those of S and w , it suffices to show that u is visible to w . Obviously no horizontal edge intersects (u, w) since taking u' to be the point of intersection and v' to be the point on $(w, z]$ vertically above u' , it is clear that u' would be closer to v' than u to v . A similar argument shows that no vertical edge intersects (u, w) since its upper endpoint u'' and the point v'' on $(w, z]$ vertically above u'' would be closer than u and v . \square

COROLLARY 3.2. *Suppose x is a point outside R , z is a point in the interior of R vertically above x and w is a point at least as high as z which sees z . Then w sees a bottom edge higher than x .*

Proof. Use Lemma 3.1, where S is the bottom-bounding edge of z . \square

COROLLARY 3.3. *Suppose x and y are points of neighboring edges T and S respectively, such that x is visible to y . Then no edge intersects $\text{int}(R(x, y))$.*

Proof. Without loss of generality let T be the top edge. Since one of the diagonals of $\text{int}(R(x, y))$ lies in the interior of R , it is easy to see that it suffices to show that no horizontal edge intersects $\text{int}(R(x, y))$. Suppose B is a horizontal edge intersecting $\text{int}(R(x, y))$. If B lies below (x, y) , then by Lemma 3.1, T sees a higher bottom edge than S . On the other hand if B lies above (x, y) , then by a symmetric version of Lemma 3.1, S sees a lower top edge than T . However both of these are impossible by the definition of neighboring edges. \square

The next lemma notes another useful fact.

LEMMA 3.4. *If T is a top edge which sees a lower bottom edge S , then there are interior points of T and S which are visible to each other.*

Proof. It is straightforward to show that if any point z in T sees a lower bottom edge then it can see an interior point of that edge. Applying this followed by the symmetric fact for higher top edges completes the proof. \square

We are now ready to give a proof of the first lemma of the last section.

LEMMA 3.5. *Let T and S be neighboring edges, where T is the top edge. Then there is a vertical edge M at least as far left as the left endpoints of T and S , whose top [bottom] endpoint is at least as high [low] as T [S]. Similarly there is a vertical edge N on the right of T and S with analogous properties. Moreover the interior of $R(M \# T, N \# S)$, i.e., the rectangle spanned by M, N, T and S , is completely contained in the interior of R .*

Proof. By Lemma 3.4 we can find x and y , interior points of T and S , respectively, which see each other. Choose M to be the rightmost vertical edge which is a left-bounding edge of some point of (x, y) . We first show that we may assume that M lies strictly to the left of both x and y . This is obviously true if x and y have the same horizontal coordinate, so assume without loss of generality that x is to the left of y . Now M is at least as far left as x since by Corollary 3.3, M cannot intersect the interior of $R(x, y)$. If M is not strictly to the left of x replace x by any point x' of T whose horizontal coordinate lies strictly between those of x and y . It is clear that x' and y are visible to each other, M is strictly to the left of both x' and y and M is still the rightmost left-bounding edge of points on (x', y) .

We next show that if w is a point of M which is lower than T and higher than S , then w is visible to both x and y . Suppose w is not visible to x . Then there is some point v of (w, x) which is on an edge of R . Let u be the point on (x, y) with the same vertical coordinate as v , and let v' be the rightmost point of $[v, u]$ which is on an edge. Then by definition v' is on the left-bounding edge of u , and hence is at least as far left as M . This shows that v , and hence x , is at least as far left as w , a contradiction since M is to the left of x . A similar argument shows that w is visible to y .

The edge N is defined analogously to M as the leftmost vertical edge which is a right-bounding edge of some point of (x, y) , and by analogous reasoning every point of N lying lower than T and higher than S is visible to both x and y . A direct consequence of this is that both M and N must span the vertical width of T and S , since otherwise one of their endpoints, and hence a horizontal edge, would be visible to both x and y , and moreover this horizontal edge would be lower than T and higher than S , contradicting T and S being neighboring edges.

All that remains is to show that no edges intersect the interior of the rectangle $R(M \# T, N \# S)$. In fact it suffices to show that no horizontal edge intersects the interior of $R(M \# T, N \# S)$, since it is clear that because there are line segments from x to points on M and N which lie entirely in the interior of R , any vertical edge intersecting $\text{int}(R(M \# T, N \# S))$ would have to have at least one endpoint in $\text{int}(R(M \# T, N \# S))$. Thus suppose some bottom edge P intersects $\text{int}(R(M \# T, N \# S))$. Choose a point w on either M or N , visible to x , such that some point z of (w, x) lies vertically above P . Then by Lemma 1 there is a bottom edge which is higher than S and visible to T , a contradiction. The symmetric argument handles the case that P is a top edge. \square

If S is a top [bottom] edge we define $n(S)$ to be the highest [lowest] bottom [top] edge which is visible to S and lower [higher] than S . If R is a finite irreducible rectilinear region, then for any horizontal edge S , we must have $n^k(S) = n^{k+2}(S)$ for

some k . Moreover since $n^k(S)$ and $n^{k+1}(S)$ are neighboring edges, by the irreducibility of R they must form a tab, which we call $g(S)$.

We will need the following lemma concerning the horizontal location of S relative to $g(S)$.

LEMMA 3.6. *Let S be a horizontal edge which does not form part of $g(S)$, and let b , d and N be the top-tip, bottom-tip and facing edge of $g(S)$ respectively. If S is a top [bottom] edge, then the horizontal coordinate of every point of S lies between those of b [d] and N .*

Proof. Let $j(S) = k$, where k is the smallest integer such that $n^k(S)$ is an edge of $g(S)$. The proof is by induction on $j(S)$. For simplicity's sake suppose $g(S)$ is a down-tab opening to the right. Thus d is to the left of b , which is to the left of N . First suppose S is a top edge and $n(S)$ is the bottom edge of $g(S)$. Let x and y be points of S and $n(S)$ respectively which see each other. Notice that just above b , the exterior of R lies immediately to the left of b and to the right of N . Thus if the horizontal coordinate of x is not between those of b and N , then some point of (x, y) lies vertically above a point of the exterior which is higher than b . But now by Lemma 3.2, x sees a lower bottom edge which is higher than $n(S)$, which contradicts the definition of $n(S)$. Thus we have shown that x lies horizontally between b and N . A similar argument shows that every point of S lies horizontally between b and N since for any point w of S we can find a point z which is vertically below w and visible to x because S is a top edge. If w is not between b and N , then some point of (x, z) lies above a point in the exterior of R which is impossible by Corollary 3.2 again.

A symmetric argument handles the case that S is a bottom edge with $n(S)$ the top edge of $g(S)$.

Now suppose the lemma holds for $k - 1$ and that S is a top edge with $n^k(S)$ a horizontal edge of $g(S)$. Then since by assumption every point of $n(S)$ lies horizontally between d and N , it is easy to see that the preceding argument can be applied to obtain the desired result. The case that S is a bottom edge is entirely analogous. \square

We give one final lemma before completing the proof of the main theorem. This is the lemma that will enable us to find an infinite sequence of edges if R is irreducible.

LEMMA 3.7. *Suppose R is irreducible and that S is a bottom edge such that $g(S)$ is a down-tab not containing S as an edge. Then there is a bottom edge $h(S)$ which is higher than S and is not part of a down-tab.*

Proof. Let P be the step-edge of $g(S)$. As R is irreducible P is a top edge, which is higher than S by Lemma 3.6 because $g(S)$ is a bad tab. Also P is not visible to S because it is a top edge which is lower than the top edge of $g(S)$. Let d and N be as in Lemma 3.6, and let w be the step point of $g(S)$. Let x be an interior point of S , and let y be the point closest to x on (x, w) which is on an edge of R . Now since y is visible to x and higher than x , y cannot be an interior point of a bottom edge; also y is not on a top edge because y is lower than the top edge of $g(S)$. Thus y is on some vertical edge M . Let $h(S)$ be the horizontal edge meeting the upper endpoint of M . Clearly $h(S)$ is higher than S . Also $h(S)$ is no higher than P by the definition of a bad tab, since like S it has points with horizontal coordinates strictly between those of d and N . We will now show that $h(S)$ is a bottom edge. First notice that $h(S)$ is a top edge if and only if $h(S)$ lies on the same side of M as the line $[x, y]$. But then if $h(S)$ is a top edge it is vertically above a point on (x, y) , and by a symmetric version of Lemma 3.1, x can see a top edge at least as low as $h(S)$. This is a contradiction since $h(S)$ is lower than $g(S)$.

Finally $h(S)$ is not part of a down-tab since if it were the top edge T of the down-tab would have to lie vertically above some point of (x, y) , and again by Lemma

3.1, x would be able to see a top edge at least as low as T . But now we reach a contradiction as before since by the same argument as for S and $h(S)$ we have that T is at least as low as P , which is lower than $g(S)$. \square

THEOREM 3.8. *If R is irreducible, then R has infinitely many edges.*

Proof. Suppose R has only finitely many edges. Then clearly R has some neighboring pair of edges, and since R is irreducible these form a tab. Without loss of generality let us assume it to be an up-tab. Let Z be the up-tab of R which has the highest top edge $[a, b]$, and let P be the step-edge of Z . Note that P is a bottom edge since Z must be a bad tab as R is irreducible. In order to obtain a contradiction it will suffice to show that R has a bottom edge S which is higher than $[a, b]$ and not part of $g(S)$. To see this note that $g(S)$ must be a down-tab because it is higher than $[a, b]$, and we may apply Lemma 3.7 to obtain another bottom edge $h(S)$ higher than $[a, b]$ which is not part of a down-tab and hence not part of $g(h(S))$. Repeating this argument we obtain an infinite sequence $S, h(S), h(h(S)), h(h(h(S))), \dots$ of distinct edges in R .

We may assume that P is part of $g(P)$, since otherwise we may take S to be P . Let $[e, b]$ be the vertical edge meeting $[a, b]$, and let S be the horizontal edge meeting $[e, b]$ at e . Let d and f be the step-points of Z and $g(P)$ respectively. Since $g(P)$ and Z do not form a tab-pair, f is higher than $[a, b]$. Moreover the horizontal coordinate of e lies between those of f and d since no edge intersects the interior of $R(b, d)$ as Z is a bad tab. But now since likewise no edge intersects the interior of $R(f, d)$, clearly e must be lower than f . Also notice that since S cannot intersect $R(b, d)$ either, S must be a bottom edge.

Finally suppose S is an edge of $g(S)$, and let T be the top edge of $g(S)$. Since $g(S)$ must be a down-tab we can find a point x of T whose horizontal coordinate lies strictly between those of b and d . Also x must be both higher than e , lower than the top edge of $g(P)$ and cannot have the same vertical coordinate as P . However this is impossible since this area is covered by the interiors of the rectangles $R(b, d)$ and $R(f \# V, d)$, where V is the top edge of $g(P)$, and thus no edges intersect this area at all. Thus S is not part of $g(S)$, and we are done. \square

Acknowledgments. The authors would like to thank Jim Shearer and Dean Sturtevant for helpful discussions.

REFERENCES

- [1] V. CHVÁTAL, *A combinatorial theorem in plane geometry*, J. Combin. Theory, Ser. B, 18 (1975), pp. 39–41.
- [2] S. FISK, *A short proof of Chvátal's watchman theorem*, J. Combin. Theory, Ser. B, 24 (1978), p. 374.

FINDING LEAST-DISTANCES LINES*

NIMROD MEGIDDO† AND ARIE TAMIR‡

Abstract. We consider the following problem related to both location theory and statistical linear regression. Given n points in the plane find a straight line L so as to minimize the weighted sum of the distances of the points to L relative to either the Euclidean metric or the l_1 -metric. We present $O(n^2 \log n)$ and $O(n \log^2 n)$ time algorithms for the Euclidean and rectilinear cases, respectively.

1. Introduction. We consider the following problem which is related to both location theory and statistics: Given n points in the plane $(x_1, y_1), \dots, (x_n, y_n)$ together with positive weights w_1, \dots, w_n , find a straight line L so as to minimize $\sum_{i=1}^n w_i d(x_i, y_i; L)$, where d is the distance function from L relative to either the Euclidean metric or the l_1 -metric.

The location theory aspects of the problem are obvious. One may think of locating a portion of a new railroad so as to minimize the average cost to the users who have to reach the tracks from different small towns. The problem is also closely related to linear regression, with the difference that here we minimize the sum of distances instead of the squared distances. The latter case is computationally much easier since there are easy formulas available for the least-squares line. This is true both in the case where the distance is measured parallel to one of the axes and also when the distance is measured vertically to the line.

We note that the problem is related to the classic Weber problem [5], [13]. The Weber problem is to find a single point so as to minimize the average distance from it to n given points. When the problem is posed with respect to the Euclidean metric no polynomial time algorithms are known even when all the weights are equal. Relative to the l_1 -metric the Weber problem is separable into two one-dimensional problems and hence is solvable in linear time by a weighted-median-finding algorithm [1].

Following the terminology of location theory we call our problem the 1-line median problem. We present in this paper an $O(n^2 \log n)$ algorithm for the Euclidean problem and an $O(n \log^2 n)$ algorithm for the rectilinear problem.

2. The Euclidean problem. In this section we focus on the Euclidean case. It is easy to see that a 1-line median can always be chosen so as to contain one of the n given points. This is because a parallel translation of the line which contains none of the points results in a linear change in the objective function as long as none of the points is reached. We, however, claim that a 1-line median can be chosen so as to contain at least two of our n points. This will enable us to consider only a set of $O(n^2)$ candidate lines for the 1-line median.

LEMMA 1. *Relative to the Euclidean metric there exists a 1-line median which contains at least two points from the set $\{(x_1, y_1), \dots, (x_n, y_n)\}$.*

Proof. We have already argued that at least one point lies on the line. Thus, we assume without loss of generality that the point (x_1, y_1) lies on the line. Moreover, we may translate the coordinate system so that $x_1 = y_1 = 0$. In other words, we may pose our problem as of finding a straight line of the form $ax + by = 0$ which minimizes the sum of weighted distances from the points (x_i, y_i) ($i = 2, \dots, n$) to the line. The

* Received by the editors November 30, 1981, and in revised form July 14, 1982.

† Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel. This author's work was supported in part by the National Science Foundation under grant ECS8121741.

‡ Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel.

distance between a point (x_i, y_i) and a line $ax + by = 0$ ($a^2 + b^2 \neq 0$) is equal to $(a^2 + b^2)^{-1/2}|ax_i + by_i|$ so, formally, we now wish to minimize the function $f(a, b) = \sum_{i=2}^n w_i |ax_i + by_i|$ subject to the constraint $a^2 + b^2 = 1$.

Suppose (a^*, b^*) is an optimal solution for the optimization problem we have posed. Let $S^+ = \{i: 2 \leq i \leq n, a^*x_i + b^*y_i \geq 0\}$ and $S^- = \{i: 2 \leq i \leq n, a^*x_i + b^*y_i \leq 0\}$. It follows that

$$\begin{aligned} f(a^*, b^*) &= \sum_{i \in S^+} w_i(a^*x_i + b^*y_i) - \sum_{i \in S^-} w_i(a^*x_i + b^*y_i) \\ &= \left(\sum_{i \in S^+} w_i x_i - \sum_{i \in S^-} w_i x_i \right) a^* + \left(\sum_{i \in S^+} w_i y_i - \sum_{i \in S^-} w_i y_i \right) b^*. \end{aligned}$$

Let α and β denote the coefficients of a^* and b^* , respectively, in the latter equality, i.e., $f(a^*, b^*) = \alpha a^* + \beta b^*$. A necessary condition for (a^*, b^*) to minimize $f(a, b)$ (subject to $a^2 + b^2 = 1$) is that it is also an optimal solution for the following optimization problem:

$$\begin{aligned} &\text{minimize } \alpha a + \beta b, \\ &\text{s.t. } ax_i + by_i \geq 0 \quad (i \in S^+), \\ &\quad \quad \quad ax_i + by_i \leq 0 \quad (i \in S^-), \\ &\quad \quad \quad a^2 + b^2 = 1. \end{aligned}$$

If $a^*x_i + b^*y_i = 0$ for some i ($2 \leq i \leq n$), then the lemma holds since the line $a^*x + b^*y = 0$ passes through (x_1, y_1) and (x_i, y_i) . Thus, assume $a^*x_i + b^*y_i \neq 0$ for all i ($i = 2, \dots, n$). We now observe that the constraints $ax_i + by_i \geq 0$ ($i \in S^+$) and $ax_i + by_i \leq 0$ ($i \in S^-$) are not binding at the point (a^*, b^*) . This implies that (a^*, b^*) is in fact an optimal solution for the problem of minimizing $\alpha a + \beta b$ subject only to $a^2 + b^2 = 1$. We note that under the present assumptions $\alpha^2 + \beta^2 \neq 0$, since otherwise all the points are colinear, which in turn implies $a^*x_i + b^*y_i = 0$ for all i . The latter optimization problem has a unique local minimum (a', b') , where $a' = -\alpha(\alpha^2 + \beta^2)^{-1/2}$ and $b' = -\beta(\alpha^2 + \beta^2)^{-1/2}$ and the corresponding objective-function value is $-(\alpha^2 + \beta^2)^{1/2}$. Thus $(a^*, b^*) = (a', b')$ and hence $\alpha a^* + \beta b^* = -(\alpha^2 + \beta^2)^{1/2} < 0$. This however is a contradiction since $\alpha a^* + \beta b^* = \sum_{i=2}^n w_i |a^*x_i + b^*y_i| \geq 0$. It follows that at least for one i ($2 \leq i \leq n$) $a^*x_i + b^*y_i = 0$ and that completes the proof.

An obvious consequence of Lemma 1 is that a 1-line median can be found in $O(n^3)$ time: Enumerate all the $O(n^2)$ candidates and compute the weighted sum of distances in each case.

We now develop an $O(n^2 \log n)$ algorithm for the 1-line median problem. The idea is to sort the candidate lines according to their slopes and then enumerate them in that order so that it takes only constant time to evaluate the sum of distances in each case. Let $-\infty < s_1 \leq s_2 \leq \dots \leq s_j \leq \dots \leq s_m \leq \infty$ denote these slopes and assume that together with each slope we have an associated pair of points.

A necessary condition for a line $ax + by + c = 0$ to be a 1-line median is that it separates the set of points into two sets of approximately the same weight; more precisely, if $W = \sum_{i=1}^n w_i$, $T^+ = \{i: ax_i + by_i > -c\}$ and $T^- = \{i: ax_i + by_i < -c\}$, then the necessary condition is that $\sum_{i \in T^+} w_i, \sum_{i \in T^-} w_i \leq \frac{1}{2}W$. In other words, the number $-c$ has to be a weighted-median of the set $H = H(a, b) = \{ax_i + by_i\}$ of the ‘‘heights’’ of the different points above the line $ax + by = 0$.

Obviously, for every pair (a, b) there is such a number c . Imagine that we increase the slope of our line continuously from $-\infty$ to $+\infty$, always selecting the number c so

as to satisfy the necessary condition. Consider the linear order induced on the set of points by heights relative to the line. This order changes only when the slope of the line coincides with one of the s_j 's, in which case the ranks of the two points associated with the critical slope are interchanged. This observation enables us to keep track of the sets T^+ , T^- as we continuously change the slope of the line. Specifically, the sets T^+ , T^- change only when the pair of points involved in a critical slope consists of no more than one member from either set. We note that some of the critical slopes may coincide (if three or more points are colinear), however this does not affect the complexity of the algorithm since we traverse all the pairs of points in any case. Given a, b and the sets T^+ , T^- , c may be redefined as $-\max\{ax_i + by_i; i \notin T^+\}$ and then the weighted sum of distances becomes

$$(a^2 + b^2)^{-1/2} \left[\left(\sum_{i \in T^+} w_i x_i - \sum_{i \in T^-} w_i x_i \right) a + \left(\sum_{i \in T^+} w_i y_i - \sum_{i \in T^-} w_i y_i \right) b + \left(\sum_{i \in T^+} w_i - \sum_{i \in T^-} w_i \right) c \right].$$

Suppose that we keep track of the sets T^+ and T^- as well as the quantities

$$\sum_{i \in T^+} w_i x_i, \quad \sum_{i \in T^-} w_i x_i, \quad \sum_{i \in T^+} w_i y_i, \quad \sum_{i \in T^-} w_i y_i, \quad \sum_{i \in T^+} w_i, \quad \sum_{i \in T^-} w_i, \quad \max\{ax_i + by_i; i \notin T^+\}$$

when we sweep the slopes in a nondecreasing order. Then it takes only $O(n^2)$ time to evaluate the objective function at all $O(n^2)$ critical slopes and choose the optimal slope. (To avoid the square-root operation we may instead maximize our objective function squared.)

3. The rectilinear problem. In the present section we consider the 1-line-median problem in the case where the distances are measured rectilinearly, i.e.,

$$d(x_i, y_i; x_j, y_j) = |x_i - x_j| + |y_i - y_j|.$$

It turns out that the distance between a line $\{ax + by + c = 0\}$ and a point (x_i, y_i) is given simply by $|ax_i + by_i + c| / \max(|a|, |b|)$. In other words, if the slope of the line is between -1 and 1 , then the distance is measured from the point to the line in parallel to the y -axis; otherwise it is measured in parallel to the x -axis. Thus, we can solve two problems: one in which all distances are measured in parallel to the y -axis and another one in which they are measured in parallel to the x -axis; we then select one of the two accordingly.

We shall now describe an algorithm for finding a straight line $y = ax + b$ so as to minimize $\sum_{i=1}^k w_i |y_i - ax_i - b|$. This problem resembles the problem of linear regression where we seek best fit in least squares. However, we do not have available a nice formula for this least-distances line like the one for the regression line. Nevertheless, the present case is more favorable than the Euclidean one due to convexity properties which are discussed below.

Let $f(a, b) = \sum_{i=1}^n w_i |y_i - ax_i - b|$ and $g(a) = \min_b f(a, b)$. Obviously, $f(a, b)$ is convex and this implies that $g(a)$ is convex.

We will find the minimum of $g(a)$. We note that the function $g(a)$ is linear on intervals between consecutive slopes of lines determined by two of the given points. Thus, $g(a)$ is piecewise linear with breakpoints only at these values. The latter can be proved along the lines of Lemma 1. It is easy to devise an $O(n^2 \log n)$ algorithm like the one in § 2. We will, however, develop a more efficient algorithm, exploiting the convexity of g .

It is easy to verify that, given a , the number $b = b(a)$ which minimizes $f(a, b)$ is a weighted-median of the set $\{y_i - ax_i\}$. Thus, $g(a)$ can be evaluated in $O(n)$ time [1].

Furthermore, even if a is a breakpoint of g , we can evaluate the one-sided derivatives $g'_+(a)$, $g'_-(a)$ of g at a . This is carried out as follows. Let $S^- = \{i: y_i - ax_i < b\}$, $S^0 = \{i: y_i - ax_i = b\}$ and $S^+ = \{i: y_i - ax_i > b\}$. We know that $w(S^-)$, $w(S^+) \leq \frac{1}{2}W$. (For any $X \subseteq \{1, 2, \dots, n\}$, $w(X) = \sum_{i \in X} w_i$.) Consider the set S^0 with the order induced by the x_i 's. According to our choice of b , $S^0 \neq \emptyset$. Thus, there exists an $i \in S^0$ such that the sets $S^{0-} = \{j \in S^0: x_j > x_i\}$ and $S^{0+} = \{j \in S^0: x_j < x_i\}$ satisfy $w(S^-) + w(S^{0-}) \leq \frac{1}{2}W$ and $w(S^+) + w(S^{0+}) \leq \frac{1}{2}W$. If $\epsilon > 0$ is sufficiently small, then $b(a + \epsilon) = y_i - (a + \epsilon)x_i$. This implies that $g'_+(a) = \sum_{j \in S^- \cup S^{0-}} w_j x_j - \sum_{j \in S^+ \cup S^{0+}} w_j x_j$. S^{0+} and S^{0-} can be obtained in $O(n)$ time, [1], which is, therefore, also the time to compute $g'_+(a)$. The evaluation of the left-hand side derivative is analogous. Thus we conclude that for a given a , it takes $O(n)$ time to compute $g(a)$, $g'_+(a)$ and $g'_-(a)$.

Let a^* denote the slope of the 1-line-median. For any a if $g'_+(a) < 0$, then $a \leq a^*$ and if $g'_-(a) > 0$, then $a \geq a^*$; otherwise, $g'_-(a) \leq 0 \leq g'_+(a)$ and that implies $a = a^*$. This enables us to search for a^* efficiently.

We will search for a' by applying a general method for solving parametric combinatorial problems first introduced in [6]. Efficient implementations are achieved with the aid of parallel computation algorithms as explained in [7]. The application in the present case is as follows. We utilize a parallel sorting algorithm by Preparata [8] which employs $n \log n$ "processors" and runs in $O(\log n)$ time. We will sort the set $\{1, \dots, n\}$ by the numbers $\{y_i - a^* x_i\}$ without actually knowing the value of a^* . Instead, throughout the process an interval $[\alpha, \beta]$ such that $\alpha \leq a^* \leq \beta$ will be maintained. At any stage, the interval will have the property that the outcomes of all the comparisons executed so far will be independent of a provided $a \in [\alpha, \beta]$. Finally, the entire order will be constant over the current interval.

Suppose that we sort the set $\{y_i - ax_i\}$, where a is restricted to some interval $[\alpha, \beta]$, but unspecified yet. Then, when we need to compare some $y_i - ax_i$ with $y_j - ax_j$, the ratio $a' = (y_i - y_j)/(x_i - x_j)$ becomes critical for that comparison. However, we can test in $O(n)$ time whether $a' \geq a^*$ or $a' \leq a^*$ and update the interval accordingly. Corresponding to each step in Preparata's sorting scheme, there will be $n \log n$ such critical values produced, one by each processor. We can search the set of critical values for a^* , namely, we will perform a binary search until our interval is narrowed down so that it does not contain any critical value in its interior. This binary search requires $O(\log n)$ tests, where each test decides whether a critical point is to the left or to the right of a^* . Thus a single stage requires $O(n \log n)$ time. However, the entire sort runs in $O(\log n)$ stages, so that our algorithm finds in $O(n \log^2 n)$ time an interval $[\alpha_0, \beta_0]$ such that $a^* \in [\alpha_0, \beta_0]$ and $g(a)$ is linear over $[\alpha_0, \beta_0]$. Finding a^* is now straightforward.

To conclude this section we contrast our $O(n \log^2 n)$ algorithm with the different solution approaches to the problem which have appeared in the statistics literature. The first approach was to apply infinite iterative processes to find the least weighted absolute deviation line. References [4], [10] represent this approach. It should be noted that some of these iterative procedures do not even guarantee convergence (e.g., [10]). The second approach (e.g., [2], [3], [12]) was to formulate and solve the problem as a linear programming problem. These methods (which are also applicable to the multidimensional case) are finite, but it is not at all clear whether their bounds are polynomial in the number of points. To our knowledge, the method in [9], [11] is the only one which has a polynomial bound. Using our notation, the method amounts to the evaluation of all the breakpoints of the piecewise linear function $g(a)$, which are between some arbitrary value \bar{a} and a^* (the minimum of $g(a)$). In the worst-case all the breakpoints of $g(a)$ may have to be looked at. Since no method is known to

perform this task in $o(n^2)$ time, our $O(n \log^2 n)$ algorithm improves considerably over all existing methods.

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 (1972), pp. 448–461.
- [2] A. CHARNES, W. W. COOPER AND R. O. FERGUSON, *Optimal estimation of executive compensation by linear regression*, Management Sci., 1 (1955), pp. 138–151.
- [3] W. D. FISHER, *A note on curve fitting with minimum deviations by linear programming*, J. Amer. Statist. Assoc., 56 (1961), pp. 359–362.
- [4] O. J. KARST, *Linear curve fitting using least deviations*, J. Amer. Statist. Assoc., 53 (1958), pp. 118–132.
- [5] H. W. KUHN AND R. E. KUENNE, *An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics*, J. Regional Sci., 4 (1962), pp. 21–33.
- [6] N. MEGIDDO, *Combinatorial optimization with rational objective functions*, Math. Oper. Res., 4 (1979), pp. 414–424.
- [7] N. MEGIDDO, *Applying parallel computation algorithms in the design of serial algorithms*, Proc. 22nd Annual IEEE Symposium on Foundations of Computer Science, 1981, pp. 399–408.
- [8] F. P. PREPARATA, *New parallel-sorting schemes*, IEEE Trans. Comp., C-27 (1978), pp. 669–673.
- [9] M. R. RAO AND V. SRINIVASAN, *A note on Sharpe's algorithm for minimizing the sum of absolute deviations in a simple regression problem*, Management Sci., 19 (1972), pp. 222–225.
- [10] E. J. SCHLOSSMACHER, *An iterative technique for absolute deviations curve fitting*, J. Amer. Statist. Assoc., 68 (1973), pp. 857–859.
- [11] W. G. SHARPE, *Mean-absolute deviation characteristic lines for securities and portfolios*, Management Sci., 18 (1971), pp. B1–B13.
- [12] H. M. WAGNER, *Linear programming techniques for regression analysis*, J. Amer. Statist. Assoc., 54 (1959), pp. 206–212.
- [13] A. WEBER, *Über Den Standort der Industrien*, Tubingen, 1909.

A MINIMAL TOTALLY DUAL INTEGRAL DEFINING SYSTEM FOR THE b -MATCHING POLYHEDRON*

WILLIAM COOK†

Abstract. Totally dual integral linear systems are intimately related to polyhedra that have the property that every nonempty face contains an integer point. A minimal totally dual integral defining system for a certain polyhedron related to b -matchings is given.

1. Introduction. The study of dual integrality is the study of integral optimal solutions to dual linear programs.

Dual integrality is studied in complexity combinatorics for several reasons. One is that often a combinatorial problem is better described as the dual of another problem. Another is to obtain combinatorial min-max theorems via the duality theorem of linear programming.

Alan Hoffman [5] introduced the concept of total dual integrality, which was latter studied and used by Edmonds–Giles [3].

A finite linear system $Ax \leq b$, with A and b rational, is called totally dual integral (TDI) when the dual linear program of the linear program

$$\max \{cx : Ax \leq b\}$$

has an integral optimal solution for integral c such that it has an optimal solution.

TDI linear systems are intimately related to integer polyhedra (those polyhedra that have the property that every nonempty face contains an integer point).

This paper investigates the relation of TDI linear systems to a combinatorial problem known as the b -matching problem. A minimal TDI defining system for a certain integer polyhedron related to b -matchings is given. Pulleyblank [9] independently obtained this result in a different way, using the results contained in [8].

2. TDI linear systems and integer polyhedra. The relation of TDI linear systems to integer polyhedra is made clear by the following two theorems.

THEOREM 1 (Edmonds–Giles [3]). *If $Ax \leq b$ is a TDI linear system with b integral, then*

$$P = \{x \in \mathbb{R}^n : Ax \leq b\}$$

is an integer polyhedron.

THEOREM 2 (Giles and Pulleyblank [4]). *Let*

$$P = \{x \in \mathbb{R}^n : Ax \leq b\},$$

where A and b are rational. If P is an integer polyhedron, then there exists a TDI linear system $A'x \leq b'$ with b' integral such that

$$P = \{x \in \mathbb{R}^n : A'x \leq b'\}.$$

Theorem 1 is a nice generalization of a theorem of Hoffman [5].

The above theorems can be combined to produce an interesting and useful technique for proving that a particular linear system is a defining system for an integer

* Received by the editors October 13, 1981, and in revised form August 19, 1982.

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

polyhedron. Suppose the goal is to prove that

$$P = \{x \in \mathbb{R}^n : Ax \leq b\}$$

is an integer polyhedron (often the goal is to prove that P is the convex hull of a certain set of integral points). By multiplying the inequalities by a positive constant if necessary, it can be assumed that b is integral. By Theorem 2, a set of inequalities that are implied by $Ax \leq b$ can be added to $Ax \leq b$ to form a TDI linear system with integral right-hand side. Theorem 1 implies that the polyhedron defined by the new system is an integer polyhedron. But the polyhedron defined by the new system is P .

The importance of the condition that b be integral in the above development is shown by the following theorem.

THEOREM 3 (Giles and Pulleyblank [4]). *For any finite, rational, linear system $Ax \leq b$, there is a positive rational number d such that $dAx \leq db$ is a TDI linear system.*

3. b -matchings. Before b -matchings are described, some notation will be given. For a real vector $x = (x_i : i \in I)$ and $S \subseteq I$, where I is a finite set, let

$$x(S) = \sum \{x_i : i \in S\}.$$

Let G be a graph with node set VG and edge set EG . For i in VG , let $N(i)$ denote the set of nodes adjacent to i (i is not adjacent to itself). For $S \subseteq VG$, let $\delta(S)$ denote the subset of edges of G that are incident to exactly one node of S (for i in VG , $\delta(i)$ will denote $\delta(\{i\})$) and let $\gamma(S)$ denote the subset of edges having both ends in S . Let $G[S]$ denote the graph with node set S and edge set $\gamma(S)$. For j in EG , let $\psi(j)$ denote the subset of VG that makes up the two ends of j (each edge is assumed to have two distinct ends). If \mathcal{S} is a collection of subsets of VG and j is an edge of G , let

$$\mathcal{S}(j) = \{R \in \mathcal{S} : j \in \gamma(R)\}.$$

Let $b = (b_i : i \in VG)$, where b_i is a positive integer for each i in VG . A b -matching in G is an integral solution to the linear system

$$(3.1) \quad x_j \geq 0 \quad \text{for every } j \text{ in } EG,$$

$$(3.2) \quad x(\delta(i)) \leq b_i \quad \text{for every } i \text{ in } VG.$$

Let $P(G, b)$ denote the convex hull of the b -matchings of G .

Edmonds has proved the following theorem by means of a good algorithm known as the blossom algorithm.

THEOREM 4 (Edmonds [2]). *A defining system for $P(G, b)$ is (3.1) and (3.2) together with*

$$(3.3) \quad \begin{aligned} x(\gamma(S)) &\leq \lfloor b(S)/2 \rfloor \text{ for all } S \subseteq VG \text{ such that } |S| \geq 3 \text{ and} \\ b(S) &\geq 3 \text{ is an odd integer.} \end{aligned}$$

A perfect b -matching of a graph G is a b -matching of G such that $x(\delta(i)) = b_i$ for all i in VG . A near perfect b -matching of G deficient at node i is a b -matching of G such that

$$x(\delta(i)) = b_i - 1$$

and

$$x(\delta(v)) = b_v \quad \text{for all } v \text{ in } VG - \{i\}.$$

A graph G is called b -critical if for every i in VG there exists a near perfect b -matching of G deficient at node i and $|VG| \geq 3$.

A balanced edge of G is a pair of nodes i, j that are joined by one or more edges and satisfy $b_i = b_j$.

For a graph G and positive, integral b , let

$$\mathcal{F} = \{S \subseteq VG : G[S] \text{ is } b\text{-critical and } G[S] \text{ contains no cutnode } i \text{ for which } b_i = 1\}$$

and

$$\mathcal{V} = \{i \in VG : i \text{ belongs to a component of } G \text{ that is a balanced edge; or } b(N(i)) > b_i \text{ and if } b(N(i)) = b_i + 1, \text{ then } \gamma(N(i)) = \emptyset\}.$$

A theorem of Pulleyblank can now be stated.

THEOREM 5 (Pulleyblank [7]). *A minimal set of inequalities that define $P(G, b)$ is (3.1) together with*

$$(3.4) \quad x(\delta(i)) \leq b_i \quad \text{for every } i \text{ in } \mathcal{V}$$

and

$$(3.5) \quad x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor \quad \text{for every } S \text{ in } \mathcal{F}.$$

That (3.1), (3.4), and (3.5) is a defining system for $P(G, b)$ follows from a result of the next section, but the minimality seems more difficult to demonstrate.

4. TDI linear systems and b -matchings. The defining system for $P(G, b)$ given in Theorem 4 is not in general a TDI linear system. This can be seen by considering a triangle with $b_i = 2$ for each node i and $c_j = 1$ for each edge j in the objective function

$$\max \sum \{c_j x_j : j \in EG\}.$$

By Theorem 2, there does exist a TDI defining system for $P(G, b)$ which has integral right-hand side. Such a TDI defining system is given in the following theorem, which can be proven easily using Edmonds' blossom algorithm (see Pulleyblank [8]).

THEOREM 6. *A TDI defining system for $P(G, b)$ is (3.1), (3.2), and*

$$(4.1) \quad x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor \quad \text{for every } S \subseteq VG.$$

Theorem 6 has been proven without making use of the blossom algorithm by Hoffman and Oppenheim [6] and Schrijver and Seymour [11] (it should be noted that although [11] deals with the special case of 1-matchings, its proof generalizes easily to b -matchings).

The system given by Theorem 6 is much larger than necessary. The result will now be improved to get a smaller TDI defining system for $P(G, b)$.

Pulleyblank has introduced the idea of b -bicritical graphs in his study of dual integrality in b -matching problems. A graph G is b -bicritical if G is connected, $|VG| \geq 3$, $b_i \geq 2$ for all i in VG , and for every i in VG there exists a b -matching of G such that

$$x(\delta(i)) = b_i - 2$$

and

$$x(\delta(v)) = b_v \quad \text{for all } v \text{ in } VG - \{i\}.$$

Some results on the structure of b -critical and b -bicritical graphs are needed to proceed further.

For a graph G , positive integral b , and $S \subseteq VG$, let

$$\begin{aligned} \mathcal{C}^0(S) &= \{i \in V - S: G[\{i\}] \text{ is a component of } G[V - S]\}, \\ \mathcal{C}^1(S) &= \{\mathcal{R} \subseteq V - S: |\mathcal{R}| \geq 2, b(\mathcal{R}) \text{ is odd, and } G[\mathcal{R}] \text{ is a} \\ &\quad \text{component of } G[V - S]\}, \\ \mathcal{C}^2(S) &= \{\mathcal{R} \subseteq V - S: |\mathcal{R}| \geq 2, b(\mathcal{R}) \text{ is even, and } G[\mathcal{R}] \text{ is a} \\ &\quad \text{component of } G[V - S]\}. \end{aligned}$$

The following theorem of Tutte characterizes those graphs which have a perfect b -matching.

THEOREM 7 (Tutte [12]). *A graph, G , has a perfect b -matching if and only if for every $S \subseteq VG$*

$$b(S) \geq b(U\mathcal{C}^0(S)) + |\mathcal{C}^1(S)|.$$

Using Theorem 7, the following two lemmas of Pulleyblank can be proven.

LEMMA 1 (Pulleyblank [7]). *A connected graph, G , is b -critical if and only if $b(VG)$ is odd, $|VG| \neq 1$, and for every nonempty $S \subseteq VG$*

$$b(S) \geq b(U\mathcal{C}^0(S)) + |\mathcal{C}^1(S)| + 1.$$

LEMMA 2 (Pulleyblank [8]). *A connected graph, G , is b -bicritical if and only if $b(VG)$ is even, $|VG| \neq 1$, and for every nonempty $S \subseteq VG$*

$$b(S) \geq b(U\mathcal{C}^0(S)) + |\mathcal{C}^1(S)| + 2.$$

It is useful to combine the above lemmas to get the following lemma, which can be proved by noting that if G is a b -bicritical graph and S is a subset of VG , then

$$b(S) + b(U\mathcal{C}^0(S)) + |\mathcal{C}^1(S)|$$

is an even number.

LEMMA 3. *A connected graph, G , is one of b -critical or b -bicritical if and only if $|VG| \neq 1$, and for every nonempty $S \subseteq VG$*

$$b(S) \geq b(U\mathcal{C}^0(S)) + |\mathcal{C}^1(S)| + 1.$$

Using Lemma 3 and the TDI-ness of the system given by Theorem 6, a theorem which gives a smaller TDI defining system for $P(G, b)$ can be obtained. The result can also be obtained by using the results of Pulleyblank [8], but it is simpler to prove it directly. The proof uses an idea of Paul Seymour for proving the same result in the special case of 1-matchings.

For a graph G and positive integral b , let

$$\mathcal{D} = \{S \subseteq VG: G[S] \text{ is } b\text{-critical or } G[S] \text{ is } b\text{-bicritical}\}.$$

THEOREM 8. *A TDI defining system for $P(G, b)$ is (3.1), (3.2), and*

$$(4.2) \quad x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor \quad \text{for every } S \text{ in } \mathcal{D}.$$

Proof. It will be shown that (3.1), (3.2), and (4.2) is a TDI linear system. That it is a defining system for $P(G, b)$ will then follow from the Edmonds–Giles theorem (Theorem 1) by noting that every b -matching of G satisfies (3.1), (3.2), and (4.2).

Let c be an integral vector and consider the linear program

$$(4.3) \quad \max \{ \sum (c_j x_j: j \in EG): (3.1), (3.2), (4.2) \}.$$

The dual linear program of (4.3) is

$$\begin{aligned} & \min \sum \{b_i y_i : i \in VG\} + \sum \{[b(S)/2] Y_S : S \in \mathcal{D}\} \\ & \text{subject to} \\ & y(\psi(j)) + Y(\mathcal{D}(j)) \geq c_j \quad \text{for every } j \text{ in } EG, \\ (4.4) \quad & y_i \geq 0 \quad \text{for every } i \text{ in } VG, \\ & Y_S \geq 0 \quad \text{for every } S \text{ in } \mathcal{D}. \end{aligned}$$

By Theorem 6, there exists an integral optimal solution, (y, Y) , to the dual linear program of

$$(4.5) \quad \max \{\sum (c_j x_j : j \in EG) : (3.1), (3.2), (4.1)\}.$$

Suppose there exists $S \subseteq VG$ such that $Y_S > 0$ and $G[S]$ is not connected. Let S_1, \dots, S_k be the subsets of S such that $G[S_1], \dots, G[S_k]$ are the components of $G[S]$. Let

$$\begin{aligned} Y'_S &= 0, \\ Y'_{S_i} &= Y_{S_i} + Y_S \quad \text{for } i = 1, \dots, k, \\ Y'_R &= Y_R \quad \text{for all other } R \subseteq VG. \end{aligned}$$

Now (y, Y') is an integral optimal solution to the dual linear program of (4.5).

This procedure allows the assumption to be made that (y, Y) is such that if $Y_S > 0$, then $G[S]$ is connected.

Suppose there exists $S \subseteq VG$ such that $Y_S > 0$ and S is not in \mathcal{D} . Since $Y_S > 0$, $G[S]$ is connected. It can be assumed that $|S|$ is not equal to 1. By Lemma 3, there exists a nonempty $X \subseteq S$ such that in $G[S]$ (notation is relative to $G[S]$)

$$(4.6) \quad b(X) < b(U\mathcal{C}^0(X)) + |\mathcal{C}^1(X)| + 1.$$

Let

$$\begin{aligned} y'_v &= y_v + Y_S \quad \text{for every } v \text{ in } X, \\ y'_v &= y_v \quad \text{for all other } v \text{ in } VG, \\ Y'_S &= 0, \\ Y'_R &= Y_R + Y_S \quad \text{for every } R \text{ in } \mathcal{C}^1(X) \cup \mathcal{C}^2(X), \\ Y'_R &= Y_R \quad \text{for all other } R \subseteq VG. \end{aligned}$$

It is easy to check that (y', Y') is a feasible solution to the dual linear program of (4.5). To show that (y', Y') is an optimal solution, it must be shown that

$$(4.7) \quad [b(S)/2] \geq b(X) + b(U\mathcal{C}^1(X))/2 + b(U\mathcal{C}^2(X))/2 - |\mathcal{C}^1(X)|/2.$$

Since the right-hand side of (4.7) is integral, it suffices to show that

$$(4.8) \quad \frac{b(S)}{2} \geq b(X) + \frac{b(U\mathcal{C}^1(X))}{2} + \frac{b(U\mathcal{C}^2(X))}{2} - \frac{|\mathcal{C}^1(X)|}{2}.$$

Since

$$b(X) = b(S) - b(U\mathcal{C}^0(X)) - b(U\mathcal{C}^1(X)) - b(U\mathcal{C}^2(X)),$$

(4.8) is equivalent to

$$(4.9) \quad b(U\mathcal{C}^0(X)) + |\mathcal{C}^1(X)| \cong b(X).$$

Now (4.9) follows from (4.6). So (y', Y') is an integral optimal solution to the dual linear program of (4.5).

The above procedure makes it possible to assume that (y, Y) is such that if $Y_S > 0$, then S is in \mathcal{D} .

Let

$$\bar{Y} = (Y_S : S \in \mathcal{D}).$$

Now (y, \bar{Y}) is an integral optimal solution to (4.4). \square

Using the techniques of Pulleyblank [8, see § 7], Theorem 8 can be sharpened.

For a graph G and positive integral b , let

$$\mathcal{D}' = \mathcal{F} \cup \{S \subseteq VG : G[S] \text{ is } b\text{-bicritical and there does not exist a node } u \in S \text{ that is adjacent to } v \in VG - S \text{ with } b_v = 1\}.$$

where \mathcal{F} is as defined in § 3. Let \mathcal{V} be defined as in § 3.

THEOREM 9. *A TDI defining system for $P(G, b)$ is (3.1),*

$$(4.10) \quad x(\delta(i)) \leq b_i \quad \text{for every } i \text{ in } \mathcal{V},$$

and

$$(4.11) \quad x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor \quad \text{for every } S \text{ in } \mathcal{D}'.$$

Proof. Again, it suffices to show that (3.1), (4.10), and (4.11) is a TDI linear system.

Let c be an integral vector. It will be shown that there exists an integral optimal solution, (y, Y) , to (4.4) such that if $Y_S > 0$ for some S in \mathcal{D} , then S is in \mathcal{D}' and if $y_i > 0$ for some i in VG , then i is in \mathcal{V} . This will prove the theorem.

By Theorem 8, there exists an integral solution, (y, Y) , to (4.4). Suppose there is an S in \mathcal{D} such that $Y_S > 0$ and S is not in \mathcal{D}' . If $G[S]$ is not b -critical, then letting

$$\begin{aligned} Y'_S &= 0, \\ Y'_{S \cup \{v\}} &= Y_{S \cup \{v\}} + Y_S, \\ Y'_R &= Y_R \quad \text{for all other } R \text{ in } \mathcal{D}, \end{aligned}$$

where $v \in VG - S$ is adjacent to a node in S and $b_v = 1$, gives an optimal solution (y, Y') to (4.4) ($G[S \cup \{v\}]$ is b -critical). So it can be assumed that $G[S]$ is b -critical. Since S is not in \mathcal{D}' , $G[S]$ is a b -critical subgraph with a cutnode v such that $b_v = 1$. Let S_1, \dots, S_k be the subsets of $S - \{v\}$ such that $G[S_1], \dots, G[S_k]$ are the components of $G[S - \{v\}]$. Let

$$S'_i = S_i \cup \{v\} \quad \text{for } i = 1, 2, \dots, k.$$

Using the definition of a b -critical graph, it is easy to check that $G[S'_i]$ is a b -critical graph for $i = 1, 2, \dots, k$. Let

$$\begin{aligned} Y'_S &= 0, \\ Y'_{S'_i} &= Y_{S'_i} + Y_S \quad \text{for } i = 1, 2, \dots, k, \\ Y'_R &= Y_R \quad \text{for all other } R \text{ in } \mathcal{D}. \end{aligned}$$

The solution (y, Y') is an optimal solution to (4.4). This procedure allows the assumption to be made that (y, Y) has the property that if $Y_S > 0$, then S is in \mathcal{D}' .

Suppose there exists an i in VG such that $y_i > 0$ and i is not in \mathcal{V} . If $b(N(i)) \leq b_i$, let

$$\begin{aligned} y'_i &= 0, \\ y'_v &= y_v + y_i \quad \text{for } v \text{ in } N(i), \\ y'_v &= y_v \quad \text{for all other } v \text{ in } VG. \end{aligned}$$

The solution (y', Y) is an optimal solution to (4.4). If $b(N(i)) = b_i + 1$, let

$$S = N(i) \cup \{i\}.$$

Since $\gamma(N(i)) \neq \emptyset$, $G[S]$ is a b -critical subgraph with no cutnode v such that $b_v = 1$. Let

$$\begin{aligned} y'_i &= 0, \\ y'_v &= y_v \quad \text{for all other } v \text{ in } VG, \\ Y'_S &= Y_S + y_i, \\ Y'_R &= Y_R \quad \text{for all other } R \text{ in } \mathcal{D}. \end{aligned}$$

Again, (y', Y') is an optimal solution to (4.4). These two operations allow the assumption to be made that (y, Y) is such that if $y_i > 0$, then i is in \mathcal{V} . \square

As was mentioned earlier, a result of this theorem is that (3.1), (3.4), and (3.5) is a defining system for $P(G, b)$. The stronger result of Pulleyblank (Theorem 5) will be needed to prove the minimality of the TDI defining system given in Theorem 9.

THEOREM 10. *A minimal TDI defining system for $P(G, b)$ is (3.1), (4.10), and (4.11).*

Proof. By Theorem 9, (3.1), (4.10), and (4.11) is a TDI defining system for $P(G, b)$. Write the system as (3.1), (4.10),

$$(4.12) \quad x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor \quad \text{for every } S \text{ in } \mathcal{F},$$

and

$$(4.13) \quad x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor \quad \text{for every } S \text{ in } \mathcal{D}' - \mathcal{F}.$$

By Pulleyblank's theorem (Theorem 5), (3.1), (4.10), and (4.12) is a minimal defining system for $P(G, b)$. Since $P(G, b)$ is a full dimensional polytope, any defining system for $P(G, b)$ must include some multiple of each inequality in (3.1), (4.10), and (4.12). So each inequality in (3.1), (4.10), and (4.12) is necessary for (3.1), (4.10), and (4.11) to be a defining system for $P(G, b)$.

To prove the theorem, all that remains to be shown is that if any inequality in (4.13) is removed, the resulting linear system is not TDI.

Let G be a graph and let $S \subseteq VG$ be such that S is in \mathcal{D}' but not in \mathcal{F} . Now let (4.13') be the set of inequalities (4.13) with

$$x(\gamma(S)) \leq \lfloor b(S)/2 \rfloor$$

removed.

It must be shown that for some integral c , the dual linear program of

$$(4.14) \quad \max \{ \sum (c_j x_j : j \in EG) : (3.1), (4.10), (4.12), (4.13') \}$$

has no integral optimal solution. This is equivalent to showing that the dual linear program of

$$(4.15) \quad \max \{ \sum (c_j x_j : j \in EG) : (3.1), (4.10), (4.11) \}$$

has no integral optimal solution, (y, Y) , such that $Y_S = 0$.

Let

$$c_j = \begin{cases} 1 & \text{for every } j \text{ in } \gamma(S), \\ 0 & \text{for every other } j \text{ in } EG. \end{cases}$$

An optimal solution to (4.15) has objective value $b(S)/2$, since any b -bicritical graph contains a perfect b -matching. It will be shown that the dual linear program of (4.15) has no optimal solution, (y, Y) , such that $Y_S = 0$.

Since c is 0, 1-valued, only 0, 1-valued solutions to the dual linear program of (4.15) need be considered. A 0, 1-valued solution to the dual linear program of (4.15) corresponds to a subset Q of \mathcal{V} and a subset T of \mathcal{D}' such that for every edge j in $\gamma(S)$, either j has an end in Q or j is contained in $\gamma(R)$ for some R in T . Such a pair (Q, T) is called a cover of $\gamma(S)$. The weight of a cover (Q, T) of $\gamma(S)$ is

$$w(Q, T) = \sum (b_i : i \in Q) + \sum (\lfloor b(R)/2 \rfloor : R \in T).$$

It must be shown that there does not exist a cover (Q, T) of $\gamma(S)$ such that S is not in T and $w(Q, T) \leq b(S)/2$.

It is straightforward to check that

(4.16) if $S' \subseteq S$ is such that $G[S']$ is connected, then there does not exist a covering (Q, T) of $\gamma(S')$ such that $Q = \emptyset$ and $w(Q, T) < \lfloor (S')/2 \rfloor$

and

(4.17) there does not exist a covering (Q, T) of $\gamma(S)$ such that $w(Q, T) \leq b(S)/2$ with $Q = \emptyset$ and S not in T .

Now (4.16) and (4.17) will be used to finish the proof. Let (Q, T) be a cover of $\gamma(S)$ such that S is not in T . If $Q = \emptyset$, then $w(Q, T) > b(S)/2$. Suppose that $Q \neq \emptyset$. It can be assumed that $Q \subseteq S$. Since $G[S]$ is b -bicritical, by Lemma 3

(4.18)
$$b(Q) \geq b(U\mathcal{C}^0(Q)) + |\mathcal{C}^1(Q)| + 1,$$

where all notation is with respect to $G[S]$.

Now (4.16) implies that

(4.19)
$$w(Q, T) \geq b(Q) + \frac{b(U\mathcal{C}^1(Q))}{2} + \frac{b(U\mathcal{C}^2(Q))}{2} - \frac{|\mathcal{C}^1(Q)|}{2}.$$

By (4.18) and (4.19),

(4.20)
$$w(Q, T) > b(S)/2. \quad \square$$

In the special case of 1-matchings, Theorem 9 implies that

(4.21) the minimal defining system for $P(G, 1)$ given by Theorem 5 is TDI.

This result on 1-matchings has been proven by Cunningham and Marsh [1].

Schrijver [10] has shown that every full dimensional, rational polyhedron is defined by a unique minimal TDI system with integral left-hand sides. This implies that the system given in Theorem 10 is the unique minimal integral TDI defining system for $P(G, b)$.

Acknowledgments. I thank the Complexity Combinatorics Fraternity of the University of Waterloo for their helpful discussions concerning this work. I also thank

Bill Pulleyblank for pointing out the necessity of the condition on b -bicritical graphs given in Theorem 9.

REFERENCES

- [1] W. CUNNINGHAM AND A. MARSH, *A primal algorithm for optimum matching*, Math. Programming Study, 8 (1978), 50–72.
- [2] J. EDMONDS, *Maximum matching and a polyhedron with $(0, 1)$ -vertices*, J. Res. Nat. Bur. Standards, 69B (1965), pp. 125–130.
- [3] J. EDMONDS AND R. GILES, *A min-max relation for submodular functions on graphs*, Ann. Discr. Math., 1 (1977), pp. 185–204.
- [4] R. GILES AND W. PULLEYBLANK, *Total dual integrality and integer polyhedra*, Linear Algebra Appl. 25 (1979), pp. 191–196.
- [5] A. J. HOFFMAN, *A generalization of max flow-min cut*, Math. Programming, 6 (1974), 352–359.
- [6] A. J. HOFFMAN AND R. OPPENHEIM, *Local unimodularity in the matching polytope*, Ann. Discr. Math., 2 (1978), pp. 201–209.
- [7] W. PULLEYBLANK, *Faces of matching polyhedra*, Doctoral thesis, Univ of Waterloo, Waterloo, Ontario, 1973.
- [8] ———, *Dual integrality in b -matching problems*, Math. Programming Study, 12 (1980), pp. 176–196.
- [9] W. PULLEYBLANK, *Total dual integrality and b -matchings*, Oper. Res. Letters, 1 (October 1981), pp. 28–30.
- [10] A. SCHRIJVER, *On total dual integrality*, Linear Algebra Appl., 38 (1981), pp. 27–32.
- [11] A. SCHRIJVER AND P. SEYMOUR, *A proof of the total dual integrality of matching polyhedra*, Report ZN79177, Mathematisch Centrum, Amsterdam.
- [12] W. TUTTE, *The factors of graphs*, Canad. J. Math., 4 (1952), pp. 314–328.

ON THE FRACTIONAL SOLUTION TO THE SET COVERING PROBLEM*

DORIT S. HOCHBAUM†

Abstract. We study the gap between the value of the integer solution to the set covering problem and the value of the fractional solution that solves the linear programming relaxation of this problem.

Key words. set covering, fractional solution, linear programming relaxation, worst case analysis

The set covering problem asks for a collection of sets the union of which covers a given set of elements at minimum cost. It is formulated as an optimization problem:

$$\begin{aligned} \min \quad & C^T \cdot x \\ \text{subject to} \quad & Ax \geq 1, \quad x \text{ a 0-1 vector,} \end{aligned}$$

where x is a vector such that x_j is equal to 1 if the j th set is selected and 0 otherwise. The vector C is the vector of costs associated with the sets and A a 0-1 matrix where $A = (a_{ij})$ and $a_{ij} = 1$ if set j covers element i and 0 otherwise. Without loss of generality we may assume that $C > 0$ ($C_j \leq 0$ implies implicitly $x_j = 1$). By replacing the requirement that x is a 0-1 vector by the nonnegativity requirement ($x \geq 0$), we obtain a linear programming problem. The purpose of this note is to study the gap between the optimal solution value to the linear program Z^f (f stands for "fractional" solution) and the integral solution to the set covering problem Z^* .

Chvátal's analysis of the greedy heuristic for set covering problems [2] uses the following ideas: he shows that the greedy solution value is equal to $y^T 1$ where y is certain feasible solution to the constraint set $y^T A \leq H(D) \cdot C^T$ and where $H(d) = \sum_{l=1}^d 1/l$ and d is the maximum column sum of the matrix.

Therefore,

$$H(d) \cdot Z^f = \min_{\substack{Ax \geq 1 \\ x \geq 0}} H(d) \cdot C^T \cdot x \geq y^T \cdot 1 \geq Z^*.$$

The above inequalities imply that $Z^*/Z^f \leq H(d)$.

Consider now the "LP heuristic" (a detailed worst case analysis of which is given in [3]): Let x^* be the fractional optimal solution of the linear program

$$\begin{aligned} \min \quad & C^T \cdot x \\ \text{subject to} \quad & Ax \geq 1, \quad x \geq 0. \end{aligned}$$

Then the set of columns $\bar{J} = \{j | x_j^* \geq 1/f\}$ is a cover where f is the maximum row sum. That is, the vector \bar{x} with $\bar{x}_j = 1$ only if $j \in \bar{J}$ and 0 otherwise, satisfies $A\bar{x} \geq 1$. Now

$$Z^f \geq \sum_{j \in \bar{J}} C_j x_j^* \geq \frac{1}{f} C \cdot \bar{x} \geq \frac{1}{f} Z^*,$$

which implies that $Z^*/Z^f \leq f$. It follows that $Z^*/Z^f \leq \min\{H(d), f\}$.

In fact, the bound f is tight in the sense that there exist problems for which the ratio Z^*/Z^f is arbitrarily close to f . We are going to show that next.

Consider the following family of problems: Let $C = 1$ (the vector of all entries equal 1) and the 0-1 matrix A contain n columns and $\binom{n}{f}$ rows each a characteristic

* Received by the editors December 15, 1981. This research was supported in part by the National Science Foundation under grant ESC-8204695.

† School of Business Administration, University of California, Berkeley, California 94720.

vector of a different subset of size f of the n columns. (Another way of interpreting A is the matrix representing a clique in a hypergraph with all edges of size f). Now, the optimal solution is equal to $n - (f - 1)$ (in fact, any solution with $n - (f - 1)$ variables set to 1 is optimal). To see that the optimal value is at least $n - (f - 1)$ we consider any solution of size $n - f$, so there are f columns that are not included. Since each subset of the columns set is represented by one of the rows there is a row that has f 1's in precisely those columns. This row is not covered, hence there is no solution of size $n - f$. The fact that any optimal solution contains at most $n - (f - 1)$ columns follows easily from the existence of f 1's in each row.

The fractional solution to such problems assigns the value $1/f$ to all variables, i.e., its value is n/f . This solution is optimal since there is a dual solution with value n/f ; the solution with all dual variables equal to $1/\binom{n-1}{f-1}$. The ratio of the optimum to the fractional solution is

$$\frac{Z^*}{Z^f} = \frac{n - (f - 1)}{n/f} = f - \frac{f(f - 1)}{n}.$$

Therefore for all f and $\epsilon > 0$ there exists an n_0 such that for $n > n_0$ the bound is at least $f - \epsilon$. Hence, the bound f is tight.

It is interesting to compare the bound to the one derived from the greedy heuristic for this family of problems. Each column sum is $\binom{n-1}{f-1}$, so

$$H(d) \geq \ln d = \sum_{i=f}^{n-1} \ln i - \sum_i \ln i \geq [\ln(n-1) - \ln(f-1)] \cdot (f-1) \geq f.$$

Therefore $H(d)$ is actually greater or equal to f for any nontrivial problem of this type.

Since this family of problems is unweighted (i.e., $C = 1$), we can compare to the bound derived in [1]. This bound applies when $C = 1$ and is equal to $n/4 + 1/2 + 1/4n$ if n is odd). Hence, for *unweighted problems*

$$\frac{Z^*}{Z^f} \leq \min \left\{ H(d), f, \frac{n}{4} + \frac{1}{2} \left(+ \frac{1}{4n} \text{ if } n \text{ is odd} \right) \right\}.$$

Note, however, that for the family of problems discussed above and ϵ sufficiently small (≤ 0.05) the value of f is smaller than $n/4$, and in that sense this bound is better than the one in [1] for these problems. Generally, however, none of these bounds dominate the others for all problem instances.

Acknowledgment: I wish to thank Tom Magnanti and an anonymous referee for their useful and insightful comments.

REFERENCES

[1] E. BALAS, *Optimal integer and fractional covers: a sharp bound on their ratio*, Rep. 81-13, Mathematisches Institut, Universitat Zu Koln, May 1981.
 [2] V. CHVÁTAL, *A greedy heuristic for the set-covering problem*, Math. Oper. Res., 4 (1979), pp. 233-235.
 [3] D. S. HOCHBAUM, *Approximation algorithms for the weighted set covering and node cover problems*, W. P. # 64-79-80, GSIA Carnegie-Mellon University, Pittsburgh, PA; SIAM J. Comput., 11 (1982), pp. 555-556.

MATRIX PRODUCTS THAT CAN BE EVALUATED IN CLOSED FORM*

RAY REDHEFFER† AND ALEXANDER VOIGT‡

Abstract. A number of elementary but novel approximations are developed for certain continued products of 2×2 matrices. In many cases the error is so small that, if the matrices have integral elements, an exact formula is obtained by taking the nearest integer to the approximation.

1. Introduction. Continued products of matrices $M_0 M_1 M_2 \cdots M_{n-1}$ are encountered in the theory of layered dielectric media, in the problem of cascaded networks and in the theory of Markov processes, to name three examples. They are also involved in the problem of explicit evaluation of product integrals. When all the matrices M_k are the same, or when they have the form $S^{-1} \Lambda_k S$ with Λ_k diagonal, the product can be evaluated with ease; but the general case is more difficult and few explicit evaluations are known.

In a study of certain definite integrals we came upon a two-parameter family of matrices for which the associated products can be evaluated in closed form and can be approximated by simple expressions of astonishing accuracy. The method which leads most naturally to these results is based on a dual interpretation of certain recurrence formulas and is presented in §§ 2, 4 and 7. Discussion from another point of view is given in §§ 5 and 8. The latter leads to efficient proofs, but gives no clue as to where the results come from. In this paper, we make a clear distinction between derivation and verification.

2. An example. Let us denote by P_n the continued product

$$P_n = \begin{pmatrix} 1 & 1 \cdot 2 \\ 1 & 2 \cdot 2 \end{pmatrix} \begin{pmatrix} 3 & 3 \cdot 4 \\ 3 & 4 \cdot 4 \end{pmatrix} \begin{pmatrix} 5 & 5 \cdot 6 \\ 5 & 6 \cdot 6 \end{pmatrix} \cdots \begin{pmatrix} 2n-1 & (2n-1)(2n) \\ 2n-1 & (2n)(2n) \end{pmatrix}.$$

Then, as will be shown, P_n is given by the exact formula

$$(1) \quad P_n = (\text{nearest integer}) \begin{pmatrix} (2n)!/e & (2n+1)!/e \\ (2n)!(1-1/e) & (2n+1)!(1-1/e) \end{pmatrix},$$

where $e = 2.718 \cdots$ is the base of natural logarithms. The phrase “nearest integer” means that each of the four entries in the following matrix is to be replaced by the nearest integer thereto. Naturally, without this modification, the formula cannot be exact, since e is irrational.

Since P_n has integral entries, (1) follows if it can be shown that the error in each of the four elements is less than $1/2$. In fact the error is less than $1/(2n)$, as will be seen shortly. For example, when $n = 5$ the approximation is

$$\begin{pmatrix} 1334960.92 & 14684570.08 \\ 2293839.08 & 25232229.92 \end{pmatrix}$$

and P_n is obtained by taking the nearest integer. If $n = 5,000$, the entries in P_n are integers involving thousands of digits and yet the error in the approximation is at

* Received by the editors April 6, 1981, and in revised form July 19, 1982.

† University of California, Los Angeles, California 90024. The research of this author was supported in part by the Mathematisches Institut I, University of Karlsruhe, under auspices of the Deutsche Forschungsgemeinschaft.

‡ Mathematisches Institut I, University of Karlsruhe, West Germany.

most 0.0001. The formula is one of those rare formulas that are both asymptotic and exact.

To obtain (1) without a priori knowledge of the value of P_n , let a_n and b_n satisfy

$$(2) \quad \int_0^1 t^{2n} e^t dt = a_n e - (2n)!, \quad n \geq 0,$$

$$\int_0^1 t^{2n+1} e^t dt = (2n+1)! - b_n e, \quad n \geq 0.$$

Then partial integration gives the recurrence formulas

$$(3) \quad a_n = 2nb_{n-1} + 1, \quad b_n = (2n+1)a_n - 1$$

valid for $n \geq 1$ and for $n \geq 0$, respectively. Now comes an important point. By the second equation (3)

$$1 = (2n+1)a_n - b_n = (2n-1)a_{n-1} - b_{n-1}, \quad n \geq 1.$$

If this expression is used instead of 1 in both equations (3) they reduce to

$$(a_n, b_n) = (a_{n-1}, b_{n-1}) \begin{pmatrix} 2n-1 & 2n(2n-1) \\ 2n-1 & (2n)^2 \end{pmatrix}.$$

The evaluation of (a_n, b_n) in terms of (a_0, b_0) leads to $(a_n, b_n) = (a_0, b_0)P_n$, P_n being the matrix product introduced above.

On the other hand the choice $a_n = (2n)!u_n$, $b_n = (2n+1)!v_n$ in (3) gives

$$u_n = v_{n-1} + 1/(2n)!, \quad v_n = u_n - 1/(2n+1)!$$

Eliminating v_{n-1} from the first equation by use of the second we get an expression for $u_n - u_{n-1}$ which implies

$$(4) \quad u_n = u_0 + \sum_{k=1}^{2n} \frac{(-1)^k}{k!}, \quad v_n = v_0 + \sum_{k=2}^{2n+1} \frac{(-1)^k}{k!}.$$

The choices $(a_0, b_0) = (1, 0)$ and $(a_0, b_0) = (0, -1)$ give respectively the first and, apart from a change of sign, the second row of the matrix P_n . The approximate formula with an accurate estimate of error is obtained when the finite sums in (4) are replaced by the corresponding infinite series.

The formulas (2) were introduced only to show how we were led to (3). We mention, however, that (2) gives

$$0 < a_n - \frac{(2n)!}{e} < \frac{1}{2n+1}, \quad 0 < \frac{(2n+1)!}{e} - b_n < \frac{1}{2n+2}$$

for the special case $(a_0, b_0) = (1, 0)$ leading to the top row of P_n .

3. A generalization. For $k \geq 0$ let $M_k(x, y)$ be defined by

$$(5) \quad M_k = \begin{pmatrix} -y(x+2k) & -y(x+2k)(x+y+2k+2) \\ x+y+2k+1 & (x+2k+1)^2 + (y+1)(x+y+2k+1) \end{pmatrix}.$$

Thus $M_k(1, -1)$ leads to the expression considered in § 2. We define further $(x)_0 = 1$ and

$$(6) \quad (x)_n = x(x+1)(x+2) \cdots (x+n-1), \quad n \geq 1,$$

$$(7) \quad F(x, y) = \frac{1}{x} + \frac{y}{x(x+1)} + \frac{y^2}{x(x+1)(x+2)} + \cdots$$

If x is 0 or a negative integer $F(x, y)$ is undefined, but we interpret products such as $(x)_n F(x, y)$ by continuity whenever possible.

Together with its proof, the following theorem provides motivation for more refined results given later.

THEOREM 1. *With M_k as in (5) let $|y| \leq 1$ and $n > (e + 1/2)|x|$. Suppose further that x and y are integers of an imaginary quadratic field. Then*

$$M_0 M_1 \cdots M_{n-1} = \begin{pmatrix} \text{nearest} \\ \text{integer} \end{pmatrix} \begin{pmatrix} 1 - xF(x, y) & 1 - xF(x, y) \\ F(x, y) & F(x, y) \end{pmatrix} \begin{pmatrix} (x)_{2n} & 0 \\ 0 & (x)_{2n+1} \end{pmatrix}.$$

The series $F(x, y)$ satisfies the difference equation

$$xF(x, y) - yF(x + 1, y) = 1$$

and belongs to a well-known class of functions which can be expressed by means of the confluent hypergeometric function [2]. In terms of the incomplete gamma function we have

$$F(x, y) = y^{-x} e^y \int_0^y e^{-t} t^{x-1} dt, \quad \text{Re } x > 0.$$

It is necessary to choose branches of y^{-x} and of the integral in such a way that the product behaves like $1/x$ near the branch-point $y = 0$, in agreement with the analytical continuation given by (7). This well-known formula [1] follows from the fact that $w(y) = y^x F(x, y)$ satisfies a first-order differential equation. It is also known [2] that

$$\lim_{x \rightarrow -m} \frac{x F(x, y)}{\Gamma(x + 1)} = y^m e^y, \quad m = 0, 1, 2, \dots$$

If $x = p$ is a positive integer we have

$$y^p F(p, y) = (p - 1)! \sum_{k=p}^{\infty} \frac{y^k}{k!}$$

which is readily expressed in terms of e^y . The case $x = 1, y = -1$ leads to the result of § 2 while $x = 2, y = -1$ gives

$$\begin{aligned} & \begin{pmatrix} 2 & 2 \cdot 3 \\ 2 & 3 \cdot 3 \end{pmatrix} \begin{pmatrix} 4 & 4 \cdot 5 \\ 4 & 5 \cdot 5 \end{pmatrix} \cdots \begin{pmatrix} 2n & 2n(2n + 1) \\ 2n & (2n + 1)^2 \end{pmatrix} \\ &= (\text{nearest integer}) \begin{pmatrix} (2n + 1)!(1 - 2/e) & (2n + 2)!(1 - 2/e) \\ (2n + 1)!/e & (2n + 2)!/e \end{pmatrix}. \end{aligned}$$

Further examples are given below, in connection with an improved formulation due to Prof. David Cantor.

4. Derivation and proof. In this section we derive the basic formulas underlying Theorem 1 without assuming a priori knowledge of them. The discussion is presented briefly, because an independent proof by mathematical induction is given in § 5 and also in § 8.

For $n \geq 0$ and $\text{Re } x > -2n$ let a_n and b_n satisfy

$$(8) \quad \begin{aligned} y^{2n} \int_0^1 t^{x+2n-1} e^{-yt} dt &= e^{-y} a_n - (x)_{2n}, \\ y^{2n+1} \int_0^1 t^{x+2n} e^{-yt} dt &= e^{-y} b_n - (x)_{2n+1}. \end{aligned}$$

By partial integration

$$(9) \quad a_n = (x + 2n - 1)b_{n-1} - y^{2n-1}, \quad b_n = (x + 2n)a_n - y^{2n}$$

for $n \geq 1$ and $n \geq 0$, respectively. Elimination of the inhomogeneous terms gives

$$(a_n, b_n) = (a_{n-1}, b_{n-1})M_{n-1}, \quad n \geq 1,$$

and hence $(a_n, b_n) = (a_0, b_0)P_n$, where $P_n = P_n(x, y)$ is the product $M_0M_1 \cdots M_{n-1}$ considered in Theorem 1. On the other hand (9) is readily solved by setting

$$a_n = (x)_{2n}u_n, \quad b_n = (x)_{2n+1}v_n.$$

If we denote by

$$F_m(x, y) = \sum_{k=0}^m \frac{y^k}{(x)_{k+1}}$$

the m th partial sum corresponding to the infinite series $F(x, y)$, the final result is the exact formula

$$(10) \quad P_n(x, y) = \begin{pmatrix} 1 - xF_{2n-1}(x, y) & 1 - xF_{2n}(x, y) \\ F_{2n-1}(x, y) & F_{2n}(x, y) \end{pmatrix} \begin{pmatrix} (x)_{2n} & 0 \\ 0 & (x)_{2n+1} \end{pmatrix}.$$

The singularities for $x = 0, -1, -2, \dots$ are removable since the factor $(x)_m$ cancels the zeros in the denominators of F_{m-1} .

To complete the proof of Theorem 1 we must estimate the remainder

$$F(x, y) - F_m(x, y) = \sum_{k=m+1}^{\infty} \frac{y^k}{(x)_{k+1}}$$

with $m = 2n - 1$ or $2n$ as the case may be. Let us suppose $x \neq 0, -1, -2, \dots$ and let us also assume $\text{Re } x \geq -R$, where R is fixed. Then by a short calculation

$$|F(x, y) - F_m(x, y)| \leq \frac{|y|^{m+1}}{|(x)_{m+2}|} e^{|y|}, \quad m \geq R - 1,$$

where we have used the fact that $|x + m + k| \geq \text{Re}(x + m + k) \geq k - 1$. Upon incorporating the factors $(x)_{2n}, (x)_{2n+1}$, we get a result which is slightly stronger than

$$(11) \quad \|P_n(x, y) - \tilde{P}_n(x, y)\| \leq \frac{|y|^{2n}}{|x + 2n|} e^{|y|} \begin{pmatrix} |x| & |xy| \\ 1 & |y| \end{pmatrix}.$$

Here $\tilde{P}_n(x, y)$ is the matrix on the right in the approximate formula of Theorem 1 and the notation $\|A\| \leq B$ for matrices $A = (a_{ij}), B = (b_{ij})$ means $|a_{ij}| \leq b_{ij}$ for all relevant values of i and j . Inequality (11) holds under the sole hypothesis that $\text{Re } x > -2n$.

Theorem 1 follows if the error is less than $1/2$, and by (11) this holds if $|y| \leq 1$ and x is in the circle of Apollonius defined by $|x + 2n| = 2e|x|$. The latter circle contains the circle $n = (e + \frac{1}{2})|x|$ and is contained in the circle $n = (e - \frac{1}{2})|x|$. Hence, if the result is based on (11), the optimal constant c in the hypothesis $n > c|x|$ lies between $e + \frac{1}{2}$ and $e - \frac{1}{2}$. Equation (11) can be sharpened in special cases, e.g., if the series for $F(x, y)$ is alternating, but these refinements will not be discussed here.

5. Cantor's factorization. It was observed by Prof. David Cantor of UCLA that $M_k = N_{2k}N_{2k+1}$, where

$$(12) \quad N_k(x, y) = \begin{pmatrix} 0 & -y(x+k) \\ 1 & x+y+k+1 \end{pmatrix}.$$

Thus, $M_0M_1 \cdots M_{n-1} = N_0N_1 \cdots N_{2n-1}$ and (10) would follow by taking $m = 2n - 1$ in the formula

$$(13) \quad N_0N_1 \cdots N_m = \begin{pmatrix} 1 - xF_m(x, y) & 1 - xF_{m+1}(x, y) \\ F_m(x, y) & F_{m+1}(x, y) \end{pmatrix} \begin{pmatrix} (x)_{m+1} & 0 \\ 0 & (x)_{m+2} \end{pmatrix}.$$

It is not difficult to show that

$$(x + y + m + 2)F_{m+1}(x, y) = yF_m(x, y) + (x + m + 2)F_{m+2}(x, y)$$

and, using this, to establish (13) by mathematical induction. This gives a new proof of (10) and shows that (10) is the special case of (13) in which there is an even number of factors on the left. Grouping factors in pairs produces the matrices M_k considered in Theorem 1. The same calculation shows that the error introduced with F instead of F_m and F_{m+1} in (13) satisfies (11) with $2n$ replaced by $m + 1$.

As pointed out by Prof. Cantor, there is a one-to-one correspondence between a matrix product of the form

$$\prod_{k=0}^m \begin{pmatrix} 0 & a_k \\ 1 & b_k \end{pmatrix}$$

and a continued fraction. If the latter admits a simple closed-form expression, or approximation, the same is true of the former. This remark obviously applies to products of matrices each of which has the form

$$\begin{pmatrix} 0 & a_k \\ 1 & b_k \end{pmatrix} \begin{pmatrix} 0 & a_{k+1} \\ 1 & b_{k+1} \end{pmatrix} = \begin{pmatrix} a_k & a_k b_{k+1} \\ b_k & a_{k+1} + b_k b_{k+1} \end{pmatrix}$$

and is the first generalization suggested by Theorem 1. A second generalization is discussed in § 7.

6. Two additional examples. The factorization introduced in § 5 allows us to replace the matrices M_k by simpler matrices N_k in forming the product. As an illustration, it is readily checked that

$$F\left(\frac{1}{2}, -1\right) = \frac{2}{e} \int_0^1 e^{t^2} dt = \alpha = 1.0761590138255 \cdots,$$

where α is defined by this equation. The choice $x = 1/2, y = -1$ in Cantor's formulation of (10), (11) gives

$$(14) \quad \begin{pmatrix} 0 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 0 & 5 \\ 2 & 5 \end{pmatrix} \cdots \begin{pmatrix} 0 & 2m-1 \\ 2 & 2m-1 \end{pmatrix} \\ = \begin{pmatrix} 1 - \alpha/2 & 1 - \alpha/2 \\ \alpha & \alpha \end{pmatrix} \begin{pmatrix} u_m & 0 \\ 0 & v_m \end{pmatrix} + O\left(\frac{2m}{m}\right),$$

where $u_m = 1 \cdot 3 \cdot 5 \cdots (2m - 1)$ and $v_m = u_{m+1}/2$. Formulation in terms of M_k with $m = 2n$ involves a more complicated product with no increase in mathematical content.

Although the error does not tend to 0, and hence does not lead to an exact formula of "nearest integer" type, the error is extremely small compared to the principal term. For example if $m = 10$ the true value of the product is

$$\begin{pmatrix} 302432822 & 3175544119 \\ 704592506 & 7398222337 \end{pmatrix}$$

and the error in the nearest-integer approximation (14) is

$$(\text{true}) - (\text{approx.}) = \begin{pmatrix} 45 & -42 \\ -90 & 79 \end{pmatrix}.$$

(This is comparable to $2^m/m = 102.4$.) In general it can be shown that the number of correct significant digits given by the approximation is about $m \log_{10}(m/e)$. Thus one can expect more than 150 correct digits when $m = 100$ and nearly 6,000 when $m = 2,000$.

As a second example, we point out that if y is complex, the formulas reveal a periodicity with respect to $\text{Im } y$ which is far from obvious at first glance. Let $(x, y) = (1, it)$, with t real, and let

$$A_n = \begin{pmatrix} 0 & 0 \\ 1 & n+1 \end{pmatrix} + it \begin{pmatrix} 0 & -n \\ 0 & 1 \end{pmatrix}.$$

The structure of A_n gives no clue that the product of A_k should involve a 2π periodicity with respect to t . Yet by (13) with $A_n = N_{n-1}$

$$\lim_{n \rightarrow \infty} \frac{A_1 A_2 \cdots A_n (nt \ 0)}{n!} \begin{pmatrix} 0 \\ t \end{pmatrix} = \begin{pmatrix} t - \sin t & t - \sin t \\ \sin t & \sin t \end{pmatrix} + i \begin{pmatrix} \cos t - 1 & \cos t - 1 \\ 1 - \cos t & 1 - \cos t \end{pmatrix}.$$

An approximate equality of the same sort holds for a finite product of n factors provided $|t|$ is small compared to n ; in fact, $|t| \leq n/e$ gives an error $o(1)$ as $n \rightarrow \infty$.

7. Another generalization. It was observed by the referee that the argument presented in § 4 can be significantly generalized. The generalization sheds light on the topic of this paper and is presented now. Consider the system

$$(15) \quad a_n = \frac{\phi_n}{\psi_{n-1}} b_{n-1} + \alpha_n, \quad b_n = \frac{\psi_n}{\phi_n} a_n + \beta_n$$

for $n \geq 1$ and $n \geq 0$, respectively, where the coefficients are complex numbers with ϕ_n, ψ_n and β_n nonzero. Just as in §§ 2 and 4 it is found that the equations can be written in homogeneous form, on the one hand, and can be solved explicitly on the other. The homogeneous form is

$$(16) \quad (a_n, b_n) = (a_{n-1}, b_{n-1}) M_{n-1}, \quad n \geq 1,$$

where

$$(17) \quad M_{n-1} = \begin{pmatrix} -\frac{\alpha_n}{\beta_{n-1}} \frac{\psi_{n-1}}{\phi_{n-1}} & -\frac{\beta_n}{\beta_{n-1}} \frac{\psi_{n-1}}{\phi_{n-1}} - \frac{\alpha_n}{\beta_{n-1}} \frac{\psi_{n-1}}{\phi_{n-1}} \frac{\psi_n}{\phi_n} \\ \frac{\alpha_n}{\beta_{n-1}} + \frac{\phi_n}{\psi_{n-1}} & \frac{\beta_n}{\beta_{n-1}} + \frac{\psi_n}{\psi_{n-1}} + \frac{\alpha_n}{\beta_{n-1}} \frac{\psi_n}{\phi_n} \end{pmatrix}$$

and the solution is obtained by setting $a_n = \phi_n u_n, b_n = \psi_n v_n$. If we let

$$(18) \quad S_n = \sum_{k=1}^n \frac{\alpha_k}{\phi_k} + \sum_{k=0}^n \frac{\beta_k}{\psi_k},$$

the final result gives the product $P_n = M_0 M_1 \cdots M_{n-1}$ in the explicit form

$$(19) \quad P_n = \frac{1}{\beta_0 \phi_0} \begin{pmatrix} \beta_0 - \psi_0 S_n + \frac{\psi_0 \beta_n}{\psi_n} \beta_0 - \psi_0 S_n \\ \phi_0 S_n - \frac{\phi_0 \beta_n}{\psi_n} \phi_0 S_n \end{pmatrix} \begin{pmatrix} \phi_n & 0 \\ 0 & \psi_n \end{pmatrix}.$$

This leads to a simple approximation when the series

$$(20) \quad S = \sum_{k=1}^{\infty} \frac{\alpha_k}{\phi_k} + \sum_{k=0}^{\infty} \frac{\beta_k}{\psi_k}$$

converges. The special case $\phi_n = (x)_{2n}$, $\psi_n = (x)_{2n+1}$, $\alpha_n = -y^{2n-1}$, $\beta_n = -y^{2n}$ in (19) gives (10). Equations (18) and (19), together with the approximation given by (20), constitute the second generalization suggested by Theorem 1.

8. A direct proof. It will be seen next that (19) admits a short direct proof, and in fact follows from the special case $\phi_n = \psi_n = 1$. This does not quite reduce (19) to a triviality, however, because the central problem is not the proof of (19) but the formulation of it in the first place.

Instead of the above substitution $a_n = \phi_n u_n$, $b_n = \psi_n v_n$ let us set $\alpha_n = p_n \phi_n$ and $\beta_n = q_n \psi_n$. Then, by a short calculation, the matrix M_{n-1} in (17) satisfies

$$M_{n-1} = (\Lambda_{n-1})^{-1} H_{n-1} \Lambda_n,$$

where

$$\Lambda_n = \begin{pmatrix} \phi_n & 0 \\ 0 & \psi_n \end{pmatrix}, \quad H_{n-1} = \frac{1}{q_{n-1}} \begin{pmatrix} -p_n & -p_n - q_n \\ p_n + q_{n-1} & p_n + q_n + q_{n-1} \end{pmatrix}.$$

Since $M_0 M_1 \cdots M_{n-1} = (\Lambda_0)^{-1} H_0 H_1 \cdots H_{n-1} \Lambda_n$, formula (19) is equivalent to

$$(21) \quad H_0 H_1 \cdots H_{n-1} = \frac{1}{q_0} \begin{pmatrix} q_0 + q_n - S_n & q_0 - S_n \\ S_n - q_n & S_n \end{pmatrix},$$

where

$$(22) \quad S_n = \sum_{k=1}^n p_k + \sum_{k=0}^n q_k.$$

That (21) holds is readily proved by mathematical induction, thus giving a new proof of (19). If $\phi_j = \psi_j = 1$ for all j , clearly $M_j = H_j$.

9. Discussion. There are so many parameters in the matrix (17) that one might well expect that an arbitrary 2×2 matrix L_{n-1} could be represented in this form. By a somewhat laborious calculation it is found that this is indeed possible, but that the conditions for ϕ_n and ψ_n are

$$(\phi_n, \psi_n) = (\phi_{n-1}, \psi_{n-1}) L_{n-1}.$$

Thus, determination of ϕ_n and ψ_n involves the very product $L_0 L_1 \cdots L_{n-1}$ that one would like to find. Something of the sort is to be expected, since there is no formula for a product of arbitrary 2×2 matrices comparable in simplicity to those of the foregoing discussion.

The generalization in § 7 gives us an opportunity to point out that the high degree of accuracy in the approximate formulas of §§ 2, 3 and 6 is the exception rather than the rule. For example, the choice $\alpha_k = 0$, $\beta_k = 1/(k+1)^2$, $\phi_k = \psi_k = 1$ in (19) gives the

approximate formula

$$(23) \quad \prod_{k=1}^{n-1} \begin{pmatrix} 0 & -k^2 \\ (k+1)^2 & (k+1)^2 + k^2 \end{pmatrix} = (n!)^2 \begin{pmatrix} 1 - \pi^2/6 & 1 - \pi^2/6 \\ \pi^2/6 & \pi^2/6 \end{pmatrix}$$

which has a superficial resemblance to those with which our discussion began. But the error is of the order of magnitude of $(n!)^2/n$ and hence is hopelessly far from giving a result of the "nearest-integer" type. If $n = 7$ the exact and nearest-integer approximate results are respectively

$$\begin{pmatrix} -12482064 & -13000464 \\ 37883664 & 38402064 \end{pmatrix}, \quad \begin{pmatrix} -16382357 & -16382357 \\ 41783957 & 41783957 \end{pmatrix}.$$

It is readily checked that the correct matrix following $(n!)^2$ in (23) is

$$\begin{pmatrix} 1 - \pi^2/6 & 1 - \pi^2/6 \\ \pi^2/6 & \pi^2/6 \end{pmatrix} + \frac{1}{n} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} + \frac{1}{2n^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

aside from terms of order $1/n^3$. The resulting approximation

$$\begin{pmatrix} -12494357 & -13012757 \\ 37895957 & 38414357 \end{pmatrix}$$

is much better, but is still wide of the mark.

Finally, we mention that the representation

$$\begin{pmatrix} 0 & -k^2 \\ (k+1)^2 & (k+1)^2 + k^2 \end{pmatrix} = \begin{pmatrix} a_k & a_k b_{k+1} \\ b_k & a_{k+1} + b_k b_{k+1} \end{pmatrix}$$

is obviously impossible, so that the product (23) does not admit the factorization discussed in § 5. This indicates that the generalizations of Theorem 1 given in §§ 3 and 7 are truly different.

REFERENCES

- [1] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series and Products*, Academic Press, New York, 1980, pp. 940-942.
- [2] W. MAGNUS, F. OBERHETTINGER AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966, pp. 262-263, p. 285.

SUBBLOCK OCCURRENCES IN THE q -ARY REPRESENTATION OF n^*

PETER KIRSCHENHOFER†

Abstract. Let $B_q(w, n)$ denote the number of subblocks w in the q -ary representation of $n \in \mathbb{N}$ (overlapping allowed). The paper deals with the mean value $m^{-1} \cdot \sum_{n=0}^{m-1} B_q(w, n)$ and an application of this result on the summing function of the generalized "sum of digits" function introduced by H. Prodinger [SIAM J. Alg. Discr. Meth., 3 (1982), pp. 35-42].

1. Introduction. In a recent paper [4] H. Prodinger has proved the following result on the number $B_2(1^s, n)$ of subblocks of s consecuting ones in the binary representation of $n \in \mathbb{N}$ (where overlapping is allowed):

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

where H_s is continuous, periodic with period 1 and satisfies $H_s(0) = 0$ and the error term $E = E_s(m)$ is bounded by $0 \leq E < 1$. The Fourier series of H_s is derived, too.

In the present paper we shall study the following generalized question:

Let $B_q(w, n)$ be the number of subblocks w in the q -ary representation of n (overlapping allowed). Is it possible to establish a result like (1) for the average numbers

$$\frac{1}{m} \sum_{n=0}^{m-1} B_q(w, n) \quad ?$$

By Theorem 1 we answer this question for subblocks w , the first digit of which differs from zero: the reason to exclude subblocks w that begin with a zero is that we do not want to allow the subblock w to overhang to the left of the most significant digit of the representation of n . If we did count occurrences that overhang in this way, then the result of our Theorem 1 would hold for all strings w that do not consist entirely of zeros.

In order to prove our desired result, we will first deal with a generalized version of the problem: let

$$(2) \quad r = \sum_{j \in \mathbb{Z}} \left(\left\lfloor \frac{r}{q^j} \right\rfloor - q \left\lfloor \frac{r}{q^{j+1}} \right\rfloor \right) q^j$$

be the q -ary representation of the real number $r \geq 0$ and $A_q(w, r)$ denote the number of occurrences of the subblock w that start to the left of the radix point (overhanging to the right allowed). In Proposition 1 we will show for the average of $A_q(w, r)$ that

$$(3) \quad \frac{1}{m} \int_0^m A_q(w, r) dr = \frac{\log_q m}{q^{|w|}} + H_w(\log_q m),$$

with H_w continuous, periodic with period 1 and $H_w(0) = 0$.

In the second step (Proposition 2) we consider the difference between $A_q(w, r)$ and the number $B_q(w, r)$ of subblocks w in the q -ary representation of r that are entirely to the left of the radix point (i.e., we consider the occurrences of w that

* Received by the editors August 19, 1981, and in revised form August 10, 1982.

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

straddle the radix point) and prove

$$(4) \quad \frac{1}{m} \int_0^m (A_q(w, r) - B_q(w, r)) dr = \frac{|w|-1}{q^{|w|}} - \frac{E_w(m)}{m},$$

where $E_w(m)$ is bounded (estimates are given in Proposition 2).

Combining (3) and (4) yields the average of $B_q(w, r)$, but since we have only counted occurrences of the subblock w that are entirely to the left of the radix point by $B_q(w, r)$, we have

$$\frac{1}{m} \int_0^m B_q(w, r) dr = \frac{1}{m} \sum_{n=0}^{m-1} B_q(w, n)$$

and the desired result is established (Theorem 1).

The proofs of Propositions 1 and 2 will be based on a “counting” lemma, which gives the connection between the numbers $A_q(w, r)$ and $B_q(w, r)$ and terms of the type $\sum_k \lfloor (n/q^k) + \beta \rfloor$ with $\beta \in [0, 1 - 1/q]$. ([4, Thm. 1] is in fact a special case.)

As a consequence of the main theorem we get a result on the summing function of the following generalized “sum of digits” function (compare [4], where the summing function is investigated in the case $q = 2$):

$$S_{q,q^{-s}}(n) = n - \sum_{k \geq 1} \sum_{1 \leq j \leq q} \left\lfloor \frac{n}{q^k} + \frac{j}{q^s} \right\rfloor.$$

We use the following abbreviations:

For any $u, w \in \{0, 1, \dots, q-1\}^*$, $v \in \{0, 1, \dots, q-1\}^{\aleph_0}$, $(u \cdot v)_q$ denotes the real number with q -ary representation $u.v$; $|w|$ the length of w and \tilde{w} the $(q-1)$ -complement of w (that means $\tilde{w} = \tilde{w}_1 \dots \tilde{w}_k$ with $\tilde{w}_i = q-1-w_i$ if $w = w_1 \dots w_k$).

2. The average value of $B_q(w, n)$. Following the plan as mentioned in the introduction we start by establishing explicit “counting” formulas for $A_q(w, r)$ and $B_q(w, r)$.

LEMMA 1. *Let w be a string of $\{0, 1, \dots, q-1\}^*$ with first digit different from 0 and let $A_q(w, r)$ denote the number of occurrences of w as a subblock in the q -ary representation (2) of the real number $r \geq 0$, where all those occurrences are counted that start to the left of the radix point (straddling allowed). Then*

$$A_q(w, r) = \sum_{k \geq 1} \left(\left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q + \frac{1}{q^{|w|}} \right\rfloor - \left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q \right\rfloor \right).$$

The number $B_q(w, r)$ of occurrences of w that are entirely to the left of the radix point is given by

$$B_q(w, r) = \sum_{k \geq |w|} \left(\left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q + \frac{1}{q^{|w|}} \right\rfloor - \left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q \right\rfloor \right).$$

Proof. The terms

$$\left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q + \frac{1}{q^{|w|}} \right\rfloor - \left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q \right\rfloor$$

can only take the values 0 or 1 and take the value 1 if and only if

$$\frac{r}{q^k} + (0 \cdot \tilde{w})_q = (u \cdot (q-1)^{|w|}v)_q$$

with strings $u \in \{0, 1, \dots, q-1\}^*$, $v \in \{0, 1, \dots, q-1\}^{\aleph_0}$.

This is equivalent to

$$\frac{r}{q^k} = (u \cdot (q-1)^{|w|v})_q - (0 \cdot \tilde{w})_q = (u \cdot wv)_q,$$

that is to an occurrence of w as a subblock in the q -ary representation of r , as w starts with a digit different from zero. The cases $1 \leq k \leq |w| - 1$ obviously correspond to the occurrences of w that straddle the radix point, while $k \geq |w|$ yields the occurrences entirely to the left of the radix point. \square

In the next step we show that the average of $A_q(w, r)$ is built up by a main term of logarithmic order and a periodic error term.

PROPOSITION 1. *Let w and $A_q(w, r)$ be as in Lemma 1. Then the mean value of $A_q(w, r)$ fulfills*

$$\frac{1}{m} \int_0^m A_q(w, r) dr = \frac{\log_q m}{q^{|w|}} + H_w(\log_q m),$$

where H_w is a continuous, periodic function with period 1 and $H_w(0) = 0$.

Proof. Using the formula for $A_q(w, r)$ established in Lemma 1 gives

$$\int_0^m A_q(w, r) dr = \int_0^m \sum_{k \geq 1} \left(\left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q + \frac{1}{q^{|w|}} \right\rfloor - \left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q \right\rfloor \right) dr,$$

which equals

$$\sum_{k=1}^{l+1} \int_0^m \left(\left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q + \frac{1}{q^{|w|}} \right\rfloor - \left\lfloor \frac{r}{q^k} + (0 \cdot \tilde{w})_q \right\rfloor \right) dr,$$

with $l = \lfloor \log_q m \rfloor$ by locating the leftmost significant digit of m .

Let $\beta = (0 \cdot \tilde{w})_q$, $s = |w|$ and

$$(5) \quad g_{\beta,s}(x) = \int_0^x \left(\left\lfloor u + \beta + \frac{1}{q^s} \right\rfloor - \left\lfloor u + \beta \right\rfloor - \frac{1}{q^s} \right) du.$$

Then $g_{\beta,s}$ is continuous, periodic with period 1, $g_{\beta,s}(0) = 0$ and by a simple substitution the sum from above turns to be

$$\frac{1}{q^s} \cdot m(l+1) + \sum_{r=0}^l q^{r+1} g_{\beta,s} \left(\frac{m}{q^{r+1}} \right)$$

which can be rewritten as

$$= \frac{1}{q^s} \cdot m(l+1) + \sum_{k \geq 0} q^{1+l-k} g_{\beta,s}(m \cdot q^{k-l-1})$$

since $g_{\beta,s}(j) = 0$ for integers j .

Now $l = \lfloor \log_q m \rfloor$ and $\{x\} = x - \lfloor x \rfloor$, so that the last expression from above equals

$$\frac{1}{q^s} \cdot m \cdot \log_q m + \frac{m}{q^s} (1 - \{\log_q m\}) + m \cdot q^{1-\{\log_q m\}} \cdot h_{\beta,s}(q^{-1+\{\log_q m\}})$$

with

$$(6) \quad h_{\beta,s}(x) = \sum_{k \geq 0} q^{-k} \cdot g_{\beta,s}(x \cdot q^k).$$

Defining

$$(7) \quad H_w(x) = \frac{1 - \{x\}}{q^s} + q^{1-\{x\}} h_{\beta,s}(q^{\{x\}-1})$$

(again with $\beta = (0 \cdot \tilde{w})_q$, $s = |w|$), H_w is continuous, periodic with period 1 and fulfills $H_w(0) = 0$ and

$$\int_0^m A_q(w, r) \, dr = \log_q m + m \cdot H_w(\log_q m),$$

so that the proof of the proposition is complete. \square

A Fourier analysis of the functions H_w can be done just in the same way as in [4] with the functions H_s from relation (1) from above and yields the following:

COROLLARY 1. $H_w(x) = \sum_{k \in \mathbb{Z}} h_k \cdot e^{2\pi i k x}$ with

$$h_0 = \log_q \frac{\Gamma((0 \cdot w)_q)}{\Gamma((0 \cdot w)_q + q^{-|w|})} - \frac{1}{q^{|w|}} \left(\frac{1}{2} + \frac{1}{\log q} \right),$$

$$h_k = \frac{\zeta\left(\frac{2k\pi i}{\log q}, (0 \cdot w)_q\right) - \zeta\left(\frac{2k\pi i}{\log q}, (0 \cdot w)_q + q^{-|w|}\right)}{2k\pi i \left(1 + \frac{2k\pi i}{\log q}\right)}, \quad k \neq 0,$$

where $\zeta(z, a)$ is the ζ -function of Hurwitz (see, e.g., [5]).

We continue our approach to Theorem 1 by studying the contribution of the occurrences of w that straddle the radix point.

PROPOSITION 2. Let $w, A_q(w, r)$ and $B_q(w, r)$ be as in Lemma 1. Then

$$\frac{1}{m} \int_0^m (A_q(w, r) - B_q(w, r)) \, dr = \frac{|w| - 1}{q^{|w|}} - \frac{E_w(m)}{m},$$

where the error term $E_w(m)$ is bounded by

$$-(1 - q^{1-|w|}) \frac{(0 \cdot \tilde{w})_q}{q - 1} \leq E_w(m) \leq (1 - q^{1-|w|}) \frac{(0 \cdot w)_q}{q - 1}.$$

Proof. Starting again with Lemma 1 and setting $\beta = (0 \cdot \tilde{w})_q$, $s = |w|$, we have

$$\int_0^m (A_q(w, r) - B_q(w, r)) \, dr = \sum_{k=1}^{s-1} \int_0^m \left(\left\lfloor \frac{r}{q^k} + \beta + \frac{1}{q^s} \right\rfloor - \left\lfloor \frac{r}{q^k} + \beta \right\rfloor \right) \, dr.$$

Using $g_{\beta,s}$ from (5) in a similar way as in the proof of Proposition 1, the last expression can be rewritten as

$$m \frac{s-1}{q^s} + \sum_{k=1}^{s-1} q^k g_{\beta,s} \left(\frac{m}{q^k} \right) = m \frac{s-1}{q^s} - E_w(m).$$

As an immediate consequence of definition (5) we have

$$-\frac{1}{q^s} \left(1 - \beta - \frac{1}{q^s} \right) \leq g_{\beta,s}(x) \leq -\frac{1}{q^s} \left(1 - \beta - \frac{1}{q^s} \right) + \left(1 - \frac{1}{q^s} \right) \frac{1}{q^s} = \beta \frac{1}{q^s}$$

so that we get the following bounds for the error term $E_w(m)$:

$$-\beta \cdot \frac{1 - q^{1-s}}{q - 1} \leq E_w(m) \leq \left(1 - \beta - \frac{1}{q^s} \right) \cdot \frac{1 - q^{1-s}}{q - 1}.$$

Observing $1 - (0 \cdot \tilde{w})_q - q^{-|w|} = (0 \cdot w)_q$, the proof of Proposition 2 is complete. \square

$$(8) \quad \frac{1}{m} \int_0^m B_q(w, r) dr = \frac{\log_q m - (|w| - 1)}{q^{|w|}} + H_w(\log_q m) + \frac{E_w(m)}{m}.$$

Since $B_q(w, r)$ counts the number of occurrences of w as a subblock in the q -ary representation that are entirely to the left of the radix point, we have

$$B_q(w, r) = B_q(w, \lfloor r \rfloor) \quad (r \geq 0)$$

and therefore

$$(9) \quad \frac{1}{m} \int_0^m B_q(w, r) dr = \frac{1}{m} \sum_{n=0}^{m-1} B_q(w, n).$$

So we have proved our desired main result:

THEOREM 1. *Let $w \in \{0, 1, \dots, q-1\}^*$ be a string with first digit different from zero. Then the following relation holds for the mean value of the number of occurrences of w as a subblock in the q -ary representation of n :*

$$\frac{1}{m} \sum_{n=0}^{m-1} B_q(w, n) = \frac{\log_q m - (|w| - 1)}{q^{|w|}} + H_w(\log_q m) + \frac{E_w(m)}{m},$$

where H_w is continuous, periodic with period 1 and $H_w(0) = 0$ and the error term E_w is bounded by

$$-(1 - q^{1-|w|}) \frac{(0 \cdot \tilde{w})_q}{q - 1} \leq E_w(m) \leq (1 - q^{1-|w|}) \frac{(0 \cdot w)_q}{q - 1}.$$

As mentioned in the introduction Theorem 1 will not remain valid for subblocks w that begin with a zero if we do not allow the subblock w to overhang to the left of the most significant digit. Nevertheless the case of strings w starting with 0 can be treated as follows: we have

$$(10) \quad B_q(0v, n) = B_q(v, n) - \sum_{j=1}^{q-1} B_q(jv, n) - \delta_1(v, n),$$

where $\delta_1(v, n)$ is 1 if the q -ary representation of n starts with the string v and 0 if not. By (10) $B_q(0^i v, n)$ can be transformed stepwise into a sum of terms as treated in Theorem 1 and “correcting” sums coming up from the δ_1 ’s. Since the complete formulas derived in this way are somewhat complicated and do not give new insight, we omit them here.

3. The summing function of the generalized “sum of digits” function. In [4] H. Prodinger studies the following generalized “sum of digits” function:

$$(11) \quad S_{q,\alpha}(n) = n - \sum_{k \geq 1} \sum_{1 \leq j < q} \left\lfloor \frac{n}{q^k} + j\alpha \right\rfloor, \quad \alpha \in [0, 1/q],$$

and evaluates the mean value

$$\frac{1}{m} \cdot \sum_{n=0}^{m-1} S_{2,2^{-s}}(n).$$

Theorem 1 from above allows to treat the case of $S_{q,q^{-s}}$ with a general $q \geq 2$:

If $S_q(n)$ denotes the usual sum of digits, then

$$(12) \quad S_{q,q^{-s}}(n) = S_q(n) - \sum_{1 \leq i < q} i \cdot B_q((q-1)^{s-1}i, n)$$

by Theorem 1 of [4] or Lemma 1 from above.

The mean value of $S_q(n)$ has been established by H. Delange [1]:

$$(13) \quad \frac{1}{m} \cdot \sum_{n=0}^{m-1} S_q(n) = \frac{q-1}{2} \cdot \log_q m + F(\log_q m)$$

with F continuous, periodic with period 1 and $F(0) = 0$.

Identity (12) and Theorem 1 yield with the abbreviation

$$(14) \quad K_{q,s}(x) = \sum_{1 \leq j < q} j \cdot H_{q-1-j/q^s, s}(x).$$

COROLLARY 2. *The generalized “sum of digits” function defined in (11) has the following mean value for $\alpha = q^{-s}$:*

$$\begin{aligned} \frac{1}{m} \cdot \sum_{n=0}^{m-1} S_{q,q^{-s}}(n) &= \left(\frac{q-1}{2} - \binom{q}{2} q^{-s} \right) \cdot \log_q m + \frac{s-1}{q^s} \binom{q}{2} \\ &\quad + F(\log_q m) - K_{q,s}(\log_q m) - \frac{D}{m}, \end{aligned}$$

where both F and $K_{q,s}$ are continuous, periodic with period 1, $F(0) = K_{q,s}(0) = 0$ and $|D| \leq q/2$.

The Fourier series of F has been developed by Delange [1], the Fourier coefficients of $K_{q,s}$ are established immediately by the defining relation (14) and Corollary 1.

COROLLARY 3. $K_{q,s}(x) = \sum_{k \in \mathbb{Z}} c_k e^{2\pi i k x}$ with

$$\begin{aligned} c_0 &= \sum_{1 \leq j < q} \log_q \Gamma\left(1 - \frac{j}{q^s}\right) - \binom{q}{2} \frac{1}{q^s} \left(\frac{1}{2} + \frac{1}{\log q}\right), \\ c_k &= \frac{\sum_{1 \leq j < q} \left(\zeta\left(\frac{2k\pi i}{\log q}, 1 - \frac{j}{q^s}\right) - \zeta\left(\frac{2k\pi i}{\log q}\right) \right)}{2k\pi i \left(1 + \frac{2k\pi i}{\log q}\right)}, \quad k \neq 0. \end{aligned}$$

Acknowledgment. The author would like to express his deep gratitude to Lyle Ramshaw for numerous suggestions which helped to increase the clarity of the ideas of the paper.

REFERENCES

[1] H. DELANGE, *Sur la fonction sommatoire de la fonction somme des chiffres*, l'Enseignement Mathématique, 21 (1975), pp. 31–47.
 [2] P. FLAJOLET AND L. RAMSHAW, *A note on Gray code and odd-even merge*, SIAM J. Comput., 9 (1980), pp. 142–158.
 [3] D. E. KNUTH, *The Art of Computer Programming, Vol. 1*, Addison-Wesley, Reading MA, 1972.
 [4] H. PRODINGER, *Generalizing the “sum of digits” function*, this Journal, 3 (1982) pp. 35–42.
 [5] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge Univ. Press, Cambridge, 1927.

INDEX TWO LINEAR TIME-VARYING SINGULAR SYSTEMS OF DIFFERENTIAL EQUATIONS*

STEPHEN L. CAMPBELL†

Abstract. An analytic method of solution is given for systems of differential equations of the form $A(t)x'(t) + B(t)x = f(t)$, where $A(t)$ may be singular and the system has index at most two.

AMS(MOS) subject classifications. 34A08, 15A09.

1. Introduction. In the last ten years the singular system of differential equations

$$(1.1) \quad Ax' + Bx = f,$$

where A, B are $n \times n$ constant matrices with A and possibly B singular has received a great deal of study. Applications have been given in economics (the Leontief model) [3], cheap control problems [1] and singular perturbations, where (1.1) is the reduced order model [11]. The early results of [9], [10] on (1.1) have also proved useful in studying nonlinear circuits with implicit models [4], [5], [7] and higher order singular arcs in control problems [8]. While some questions remain, the theory for (1.1) has reached a fairly mature level. For a detailed development and reasonably complete bibliography see [6], [7], [9] which contain discussions of most of the previously mentioned work.

The situation for the time varying case is quite different. While there do exist procedures for solving

$$(1.2) \quad A(t)x'(t) + B(t)x(t) = f(t)$$

in some cases [6], [7], [12] no analogue of the explicit formula given in [10], (or [9], [6]) has been derived except for special cases. This has complicated the numerical and analytical analysis of a variety of problems which contain (1.2) as a subsystem. This paper provides a solution of (1.2) under assumptions that include prior results and important new special cases. Examples will be given that show the difficulty in extending these results.

2. Preliminaries. All matrices are taken to be complex matrices. For a square matrix E , there exists a nonsingular matrix R so that

$$(2.1) \quad E = R \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix} R^{-1},$$

where C is invertible, N is nilpotent of index k , that is, $N^k = 0$, $N^{k-1} \neq 0$ and either C or N may be absent. The index of E , $\text{Index}(E)$, is k . If E is invertible, then $\text{Index}(E) = 0$. The rank of C is called the core-rank of E [9]. The Drazin inverse of E , denoted E^D , is given by

$$(2.2) \quad E^D = R \begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix} R^{-1}$$

* Received by the editors June 11, 1982, and in revised form August 30, 1982. This research was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, under grant AFOSR-81-0052A.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

if C is present and by $E^D = 0$ if E is nilpotent. The Drazin inverse has the properties:

$$E^D E E^D = E^D, \quad E E^D = E^D E, \quad E^{k+1} E^D = E^k.$$

Further properties of the Drazin inverse may be found in [9]. The range and nullspace of E are denoted $\mathcal{R}(E), \mathcal{N}(E)$.

The system (1.2) is said to be *solvable* at t if there exists a scalar $\lambda(t)$ so that $\lambda(t)A(t) + B(t)$ is invertible. If the system (1.2) is solvable for all t of interest, it is said to be a solvable system. For (1.1), solvability is directly related to the uniqueness of solutions. This is not true for (1.2), [6], [7]. Most attempts to numerically solve (1.2) or (1.1) have involved backward difference schemes [7]. Solvability is needed to insure that the resulting linear systems will be consistent and have a unique solution for all but a finite number of possible time step sizes. Since in most cases (1.2) will in fact be solved numerically, solvability, where possible, is a natural assumption to make.

If (1.2) is solvable at t , then the index of (1.2) at t is $\text{Index}((\lambda(t)A(t) + B(t))^{-1}A(t))$. The index depends only on $A(t), B(t)$ and not $\lambda(t)$ [10], [6], [9]. For (1.2), the property that the index is identically zero, identically one, or greater than or equal to two on an interval $[0, T]$ is invariant under “most” coordinate changes of the form $x = P(t)y$ [7].

By a *solution* of (1.2) on $[0, T]$ we mean a differentiable function of t on $[0, T]$. An initial condition x_0 is called *consistent* at t_0 if there is a solution $x(t)$ so that $x(t_0) = x_0$. Even for (1.1), the consistent initial conditions form a $\text{Core-rank}[(\lambda A + B)^{-1}A]$ -dimensional linear manifold.

There are, in general, other kinds of solutions to (1.2) or (1.1) which are impulsive or distributional. However, the derivation of functional solutions and the consistent initial conditions is of interest. Knowing the consistent initial conditions tells when impulses may be present. If a singular perturbation approach is being used, the functional solution is all that is used from the reduced order problem. In many singular nonlinear systems, such as those involving relaxation oscillations, the functional solutions or parts of them represent the “observable” or physically realizable states [7]. Finally, even if (1.2) can be rewritten as a nonsingular subsystem, the form (1.2) may be used to preserve system structure. In such “descriptor systems” the functional solutions are often of most interest.

The system (1.2) often appears in the form

$$(2.3a) \quad x'(t) = A_{11}(t)x(t) + A_{12}(t)y(t) + f_1(t),$$

$$(2.3b) \quad 0 = A_{21}(t)x(t) + A_{22}(t)y(t) + f_2(t)$$

especially in singular perturbation and nonoptimal control problems [13], [14], [15]. System (2.3) has index one if and only if $A_{22}(t)$ is invertible [7]. This is the case that has been most studied to date [7], [12]. In this paper we shall solve (1.2) when the index is two, so that A_{22} will be singular. Such systems arise in circuits with operational amplifiers and in singular control problems [7]. Examples will be given to show that our analysis does not extend to the index 3 case.

3. Main results. Our starting point will be

$$(3.1) \quad \hat{A}(t)x' + x = f(t), \quad 0 \leq t \leq T,$$

where \hat{A}, f are assumed to be differentiable. If the original system is in the form (1.2) and is solvable, the change of variables $x = \exp(\int_0^t \lambda(s) ds)z$ may be used to rewrite (1.2) in the form (3.1), where $\hat{A}(t) = (\lambda(t)A(t) + B(t))^{-1}A(t)$ and the new $f(t)$ is

$(\lambda(t)A(t) + B(t))^{-1}f(t)$. Alternatively, one may just multiply (3.1) by $(\lambda(t)E(t) + F(t))^{-1}$ and slightly modify the arguments that follow.

There are also two ways to proceed with the derivation, directly or through "canonical" forms. We shall proceed directly and then mention the canonical form version.

Our main result is the following.

THEOREM 1. *In the system (3.1) assume that f and $\hat{A}(t)$ are continuously differentiable on $[0 T]$, $\text{Index}(\hat{A}(t)) \leq 2$ on $[0 T]$ and $\text{Rank}(\hat{A}^2(t)) = \text{Core-rank}(\hat{A}(t))$ is constant on $[0 T]$. Define the projections P, Q by $P = \hat{A}^D \hat{A}$, $Q = I - P$. Let $N = \hat{A}(I - \hat{A}^D \hat{A})$. Assume that*

$$(3.2) \quad I - N'(t) \text{ is invertible on } [0 T].$$

Then $x(t)$ is a smooth functional solution of (3.1) if and only if $x = [Px] + [Qx]$, where $[Px]$ and $[Qx]$ are given by

$$(3.3a) \quad [Px]' = (P' - \hat{A}^D)[Px] + P'[Qx] + \hat{A}^D f,$$

$$(3.3b) \quad [Qx] = [Qf] - [I - N']^{-1} N[Qf]' - [I - N']^{-1} NP'[Px].$$

Thus the dimension of the manifold of consistent initial conditions is the Core-rank of $\hat{A}(0)$ and the manifold is; $P(0)x(0)$ is arbitrary and $Q(0)x(0)$ is (3.3b) evaluated at $t = 0$.

Proof. Since $\text{Rank}(\hat{A}^2(t))$ is constant, P, Q and \hat{A}^D are as smooth as \hat{A} on $[0 T]$. Thus N is differentiable and $I - N'$ is well defined.

Differentiating the expressions $N^2 = 0, P^2 = P, Q^2 = Q$ and using $P + Q = I, PQ = QP = 0$ the following identities are easily derived:

$$(I - N')N = N(N' + I), \quad PP' = P'Q, \quad PP'P = 0,$$

$$QQ' = Q'P, \quad P' = -Q', \quad QQ'Q = 0.$$

These facts will be used repeatedly in the following derivation. Note that $x = [Px] + [Qx]$. Substituting this into (3.1) gives the equivalent system:

$$(3.4) \quad \hat{A}[Px]' + \hat{A}[Qx]' + [Px] + [Qx] = f.$$

Multiplying (3.4) by $\hat{A}^D = \hat{A}^D P$ and Q yields

$$(3.5a) \quad P[Px]' + P[Qx]' + \hat{A}^D [Px] = \hat{A}^D f,$$

$$(3.5b) \quad N[Px]' + N[Qx]' + [Qx] = Qf.$$

Now

$$(3.6) \quad \begin{aligned} P[Qx]' &= PQ'x + PQx' = PQ'x = PQ'([Px] + [Qx]) \\ &= -PP'Px + PQ'[Qx] = PQ'[Qx] = -PP'Qx = -P'[Qx]. \end{aligned}$$

Similarly,

$$(3.7) \quad Q[Px]' = -Q'[Px] = P'[Px].$$

Thus (3.5) is

$$(3.8a) \quad P[Px]' - P'[Qx] + \hat{A}^D [Px] = \hat{A}^D f,$$

$$(3.8b) \quad NP'[Px] + N[Qx]' + [Qx] = Qf.$$

Now multiply (3.8b) by N so that $N[Qx] = NQf$ or, upon differentiating, $N'[Qx] + N[Qx]' = [NQf]'$. Thus

$$(3.9) \quad NQ[x]' = [NQf]' - N'[Qx].$$

Substitute (3.9) into (3.8b) to get $[I - N'] [Qx] = Qf - [NQf]' - NP'[Px] = [I - N'] [Qf] - N[Qf]' - NP'[Px]$. Since $\mathcal{R}(N) = \mathcal{R}([I - N']N) \subseteq \mathcal{R}([I - N']Q)$, this equation can be solved for $[Qx]$ to yield (3.8b). Substituting (3.7) and (3.8a) into $[Px]' = P[Px]' + Q[Px]'$ gives (3.3a). \square

The system (3.3) completely determines $[Px]$, $[Qx]$ since if (3.3b) is substituted into (3.3a) a nonsingular system just in $[Px]$ results. Its solution can then be used in (3.3b).

If $\hat{A}^2(t)$ has constant range and nullspace and N is constant, then P is constant and (3.3) becomes $[Px]' = \hat{A}^D [Px] + \hat{A}^D f$, $[Qx] = [Qf] - N[Qf]'$ as in [6]. Thus Theorem 1 includes the constant coefficient index two case. If \hat{A} has index one and constant rank on $[0 T]$, then (3.2) holds (since $N \equiv 0$) so that the index one case is also completely included.

Note that Theorem 1 does not require that any ranges or nullspaces are constant nor does it require that either Rank $(\hat{A}(t))$ or Index $(\hat{A}(t))$ be constant.

If (3.2) does not hold, there are several possibilities. If $\mathcal{R}(N) \subseteq \mathcal{R}([I - N']Q)$, then (3.1) is consistent for all f but uniqueness depends on whether $\mathcal{N}([I - N']Q) \neq \{0\}$. If $\mathcal{R}(N) \not\subseteq \mathcal{R}([I - N']Q)$, then (3.1) may not be consistent for all f . Alternatively, it is possible that (3.1) could be transformed to a system with index greater than two and the solution involves higher than first derivatives of f [7, Example 5.2.1, p. 117]. See also [6, Example 6.4.1, p. 147] and [7, Examples 5.4.1, 5.4.2, pp. 124–125].

The usefulness, either conceptually or in practice, of Theorem 1 remains to be determined and is under investigation. In any attempt to utilize (3.3) on test problems, for example, to provide “true” solutions to compare implicit numerical methods with, several observations need to be made. First, even if $A(t)$ is invertible in (1.2), then a linear system $E(t)z(t) = u(t)$ will have to be solved for some multiple of the number of time steps and the more rapidly E changes, the more solutions will be needed. Thus frequent “inversions” are intrinsic to problems in the form (1.2) or (3.1).

Second, there are several ways to compute $\hat{A}^D(t)$ or $\hat{A}^D(t)h(t)$ for a known vector $h(t)$, [9], [16] for a given value of t . One of them, [9, Algorithm 7.55, p. 134] which would not ordinarily be used, consists of a procedure which terminates in at most n steps, where \hat{A} is $n \times n$, and involves only matrix products, taking traces and dividing by a scalar function. In principle it is possible to obtain $(\hat{A}^D)'(t)$ and hence $P'(t)$ for small matrices with simple entries by this approach in a program language that does symbolic manipulation.

Another approach is motivated by the observation that in practice it is usually easier or possibly safer to obtain $(\hat{A}')^D$ rather than $(\hat{A}^D)'$ since explicit computation of \hat{A}^D as a function of t is difficult and taking the Drazin inverse numerically and then numerically differentiating is likely to create greater error. The key fact relating $(\hat{A}')^D$ and $(\hat{A}^D)'$ is the following unpublished result of Carl Meyer.

PROPOSITION 1 (Meyer). *Suppose that $A(t)$ is an $n \times n$ differentiable matrix valued function on $[0 T]$ with constant core-rank. Then*

$$(3.10) \quad \begin{bmatrix} A & A' \\ 0 & A \end{bmatrix}^D = \begin{bmatrix} A^D & (A^D)' \\ 0 & A^D \end{bmatrix}.$$

Proof. From [9, Thm. 7.7.1, p. 139],

$$\begin{bmatrix} A & A' \\ 0 & A \end{bmatrix}^D = \begin{bmatrix} A^D & H \\ 0 & A^D \end{bmatrix},$$

where

$$\begin{aligned} (3.11) \quad H &= (A^D)^2 \left[\sum_{i=0}^n (A^D)^i A' A^i \right] (I - AA^D) \\ &+ (I - AA^D) \left[\sum_{i=0}^n A^i A' (A^D)^i \right] (A^D)^2 - A^D A' A^D. \end{aligned}$$

But from [2, Thm. 2], $H = (A^D)'$. \square

There is a trade-off here, of course, in that

$$\begin{bmatrix} A & A' \\ 0 & A \end{bmatrix}^D$$

is $2n \times 2n$. On the other hand, its computation gives not only $(A^D)'$ but also A^D . Proposition 1 has several nonobvious consequences such as the facts that

$$\begin{bmatrix} A & A' \\ 0 & A \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A^D & (A^D)' \\ 0 & A^D \end{bmatrix}$$

have the same core-rank and that the right-hand side of (3.10) has index one. Even for the index two case where $n = 1$ in (3.11), the computation of A^D , A' and using (3.11) is probably more effort than computing A' and the left-hand side of (3.10) by a procedure such as Wilkinson's [16].

It is also possible to derive Theorem 1 from a coordinate change point of view. We shall omit the details. Under the assumptions of Theorem 1, there exists a differentiable matrix $P(t)$ on $[0, T]$ such that

$$(3.12) \quad P^{-1}(t)\hat{A}(t)P(t) = \begin{bmatrix} C(t) & 0 \\ 0 & N(t) \end{bmatrix},$$

where C is invertible and $N^2 = 0$ on $[0, T]$. Let $x = Py$ and $H = P^{-1}P'$, $\tilde{f} = P^{-1}f$. Then, decomposing H, y as in (3.12), (3.1) becomes

$$(3.13a) \quad Cy'_1 + (I + CH_{11})y_1 + CH_{12}y_2 = \tilde{f}_1,$$

$$(3.13b) \quad Ny'_2 + NH_{12}y_1 + (I + NH_{22})y_2 = \tilde{f}_2.$$

The analogue of (3.3) is

$$(3.14a) \quad y'_1 = -(C^{-1} + Q_{11})y_1 - Q_{12}y_2 + C^{-1}\tilde{f}_1$$

$$(3.14b) \quad y_2 = [N' - I - NQ_{22}]^{-1} \{-\tilde{f}_2 + (Nf_2)' + NQ_{12}y_1\}.$$

4. An example. A slight modification of an example in [6] can be used to point out difficulties in the extension of Theorem 1 to the index three case.

Let

$$J = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 & 0 \\ -t & 1 & 0 \\ 0 & -t & 1 \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ t & 1 & 0 \\ t^2 & t & 1 \end{bmatrix},$$

$$A(t) = RJR^{-1} = \begin{bmatrix} 2t & 2 & 0 \\ 0 & 0 & 2 \\ -2t^3 & -2t^2 & -2t \end{bmatrix}.$$

Then

$$(4.1) \quad A(t)x' + x = 0$$

has constant core-rank (namely zero), constant index 3 and $I - N' = I - A'$ is invertible. Let $x = Ry$. Then (4.1) becomes

$$(4.2) \quad \begin{bmatrix} 0 & -2 & 0 \\ 0 & 4t^2 & -2 \\ 0 & 0 & 0 \end{bmatrix} \dot{y} + y = 0.$$

System (4.2) has index two on $(0, T]$, index three at zero, core-rank one on $(0, T]$, core-rank zero at zero. The solution of (4.2) is

$$(4.3) \quad y = \begin{bmatrix} t^2 e^{-4/t} c \\ -2t^2 c \\ 0 \end{bmatrix}, \quad c \text{ an arbitrary constant.}$$

Thus (4.1) has a nonzero one-dimensional manifold of solutions even though A is nilpotent, which never occurs in the constant coefficient or index two case as shown by Theorem 1. Also consistent initial conditions at $t = 0$ do not uniquely determine solutions.

Example 1 also shows that a system with index greater than two can sometimes be changed into an index two system.

5. Conclusion. A fairly general solution of the index two singular linear system has been derived and discussed. An example has been given to show that similar results for higher index systems will probably be less general and involve additional technical assumptions.

REFERENCES

- [1] S. L. CAMPBELL, *Optimal control of autonomous linear processes with singular matrices in the quadratic cost functional*, SIAM J. Control, 1 (1976), pp. 1092–1106.
- [2] ———, *Differentiation of the Drazin inverse*, SIAM J. Appl. Math., 30 (1976), pp. 703–707.
- [3] ———, *Nonregular singular dynamic Leontief systems*, Econometrica, 47 (1979), pp. 1565–1568.
- [4] ———, *A procedure for analyzing a class of non-linear equations that arise in circuit and control problems*, IEEE Trans. Circuits and Systems, 28 (1981), pp. 256–261.
- [5] ———, *Consistent initial conditions for singular nonlinear systems*, Circuits, Systems and Signal Processing, 1 (1982), to appear.
- [6] ———, *Singular Systems of Differential Equations*, Pitman, London, 1980.
- [7] ———, *Singular Systems of Differential Equations II*, Pitman, London, 1982.

- [8] S. L. CAMPBELL AND K. CLARK, *Order and the index of singular time invariant linear systems*, Systems and Control Letters, 1 (1981), pp. 119–122.
- [9] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.
- [10] S. L. CAMPBELL, C. D. MEYER, JR. AND N. J. ROSE, *Applications of the Drazin inverse to linear systems of differential equations*, SIAM J. Appl. Math., 31 (1976), pp. 411–425.
- [11] S. L. CAMPBELL AND N. J. ROSE, *A second order singular linear system arising in electric power systems analysis*, Int. J. Systems Sci., 13 (1981), pp. 101–108.
- [12] V. DOLEZAL, *Some properties of non-canonic systems of linear integro-differential equations*, Cas. pest matem., 89 (1964), pp. 470–490.
- [13] R. E. O'MALLEY, JR., *On singular singularly-perturbed initial value problems*, Applicable Analysis, 8 (1978), pp. 71–81.
- [14] R. E. O'MALLEY, JR. AND J. E. FLAHERTY, *Singular singular-perturbation problems*, Lecture Notes in Mathematics 594, Springer-Verlag, New York, 422–436.
- [15] ———, *Analytical and numerical methods for nonlinear singular singularly perturbed initial value problems*, SIAM J. Appl. Math., 38 (1980), pp. 225–248.
- [16] J. H. WILKINSON, *Note on the practical significance of the Drazin inverse*, in Recent Applications of Generalized Inverses, S. L. Campbell, ed., Pitman, London, 1982, pp. 82–99.

THE CONCEPT OF TWO-CHORD TIESETS AND ITS APPLICATION TO AN ALGEBRAIC CHARACTERIZATION OF NON-SERIES-PARALLEL GRAPHS*

SHOJI SHINODA,[†] WAI-KAI CHEN[‡] AND SHU-PARK CHAN[§]

Abstract. The concept of two-chord tiesets is introduced and a necessary and sufficient condition for a graph to be a non-series-parallel graph is given by use of the rank condition of a two-chord tieset matrix.

1. Introduction. Throughout this paper, G is used to denote a nonseparable directed graph with edge set E and node set V . The edges contained in a tree t of G are called the *branches* of t and the edges not contained in t are called the *chords* of t where the set of all chords of t is called a *cotree* of t , denoted by \bar{t} . For a tree t of G , an elementary tieset of G containing exactly k of its chords, with an arbitrarily assigned tieset direction, is called a *k-chord tieset*¹ with respect to t of G , and a tieset matrix of G which has one row for each k -chord tieset with respect to t of G and one column for each edge of G is called a *k-chord tieset matrix* with respect to t of G , denoted by $B^{(k)}(t)$. Note that in $B^{(k)}(t)$ we do not include a tieset that is obtained merely by reversing the direction, the reason being that by reversing a direction it only changes the sign of a row of $B^{(k)}(t)$. The rank of $B^{(k)}(t)$ is denoted by $\text{rank}[B^{(k)}(t)]$. Now let us denote the rank and the nullity of G by ρ and μ , respectively. Then

$$\text{rank}[B^{(k)}(t)] \leq \mu.$$

In particular, if $k = 1$, then the equality in this relation is always satisfied for every t of G with $\mu \geq 1$. However, if $k \geq 2$, then the equality in the relation is not always satisfied. This brings us to the following question:

“What kind of structure does G have if

$$\text{rank}[B^{(k)}(t)] = \mu$$

is satisfied for at least one tree t of G ?”

The purpose of this paper is to prove that

$$\text{rank}[B^{(2)}(t)] = \mu$$

is satisfied for at least one tree t of G if and only if G is a non-series-parallel graph. This result is an algebraic characterization of the non-series-parallel graphs.

2. Algebraic characterization of non-series-parallel graphs. Let $B_r^{(2)}(t)$ be a tieset matrix of G obtained from $B^{(2)}(t)$ by deleting all rows that are not contained in a maximal set of linearly independent rows of $B^{(2)}(t)$. $B_r^{(2)}(t)$ is called a *reduced two-chord tieset matrix* with respect to t of G . From this definition, it follows that

$$(1) \quad \text{rank}[B_r^{(2)}(t)] = \text{rank}[B^{(2)}(t)].$$

The complete undirected graph on four vertices is denoted by K_4 . A graph obtained from K_4 by assigning arbitrary directions to its edges is called an *oriented*

* Received by the editors July 17, 1981, and in revised form July 26, 1982.

† Department of Electrical Engineering, Chuo University, Tokyo 112, Japan.

‡ Department of Electrical Engineering and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60680.

§ Department of Electrical Engineering and Computer Science, University of Santa Clara, Santa Clara, California 95053.

¹ The term “circuit” is sometimes used instead of “tieset”.

K_4 . For any two subsets E_s and E_0 of E such that $E_s \cap E_0 = \emptyset$, a graph obtained from G by deleting all edges in E_0 and contracting all edges in E_s is called a *minor* of G , denoted by $G(E_s; E_0)$. In particular, $G(\emptyset; E_0)$ is called a *subgraph* of G and $G(E_s; \emptyset)$ is called a *contraction* of G . G is called a *non-series-parallel graph* if an oriented K_4 is a minor of G .

THEOREM 1. *If G is a non-series-parallel graph, then there is a tree t of G such that*

$$(2) \quad \text{rank } [B^{(2)}(t)] = \mu.$$

Proof. Since the rank of a two-chord tieset matrix with respect to a tree t of G remains unchanged under the operation of reversing the directions of edges and two-chord tiesets with respect to t , the directions of edges and two-chord tiesets with respect to t can therefore be assigned arbitrarily in the following proof.

Suppose that G is a non-series-parallel graph. From [1] we can choose a tree t of G in such a way that a minor of G is an oriented K_4 shown in Fig. 1, where thin lines represent chords of t and thick lines represent branches of t . Note that throughout this paper, unless otherwise specified, the directions of some or all of the edges may be omitted for convenience when graphs are drawn.

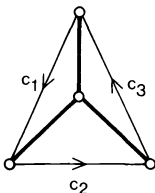


FIG. 1. An orientation of K_4 in which the directions of branches are omitted.

In order to prove this theorem, we have only to show that we can construct a reduced two-chord tieset matrix $B_r^{(2)}(t)$ with respect to t of G which has the form

$$(3) \quad B_r^{(2)}(t) = \begin{bmatrix} S & T \end{bmatrix} \begin{array}{l} \swarrow \text{columns corresponding to chords} \\ \searrow \text{columns corresponding to branches} \end{array}$$

such that

- (a) the zero-nonzero pattern of S is

$$(4) \quad S = \begin{bmatrix} c_1 & c_2 & c_3 & c_4 & \cdots & c_\mu \\ 1 & 1 & 0 & & & \\ 0 & 1 & 1 & & & \\ 1 & 0 & 1 & & & \\ \hline & & & \pm 1 & & \\ & & & & \ddots & \\ & & & & & 0 \\ & & & & & & \ddots & \\ & & & & & & & \pm 1 \end{bmatrix},$$

- (b) every row of S contains exactly two nonzero elements and
- (c) the nonzero elements in S are $+1$ or -1 , where ± 1 stands for either $+1$ or -1 and c_i 's stand for the chords of t .

Now, for a tree t of G , let $B_c(t)$ be the fundamental tieset defined by a chord c with respect to t , and let $Q_b(t)$ be the fundamental cutset defined by a branch b with

respect to t . If $b \in B_c(t) - \{c\}$, then $t \cup \{c\} - \{b\}$ is a tree of G which is different from t . Next, we consider the following algorithm:

Step 1. Set $i \leftarrow 0$ and $C^{(0)} = \{c_1, c_2, c_3\}$, where c_1, c_2, c_3 are the chords of t shown in Fig. 1.

Step 2. Set $i \leftarrow i + 1$.

Step 3. Set $C^{(i)} = \{c' \mid c' \in Q_c(t \cup \{c\} - \{b\}) - C^{(i-1)}, c \in C^{(i-1)}, b \in B_c(t) - \{c\}\}$.

Step 4. If $C^{(i)} \neq \emptyset$, go to Step 2; otherwise stop.

At step 4, if $C^{(i)} \neq \emptyset$, then for each chord $c' \in C^{(i)}$ there exists the fundamental tieset $B_{c'}(t \cup \{c\} - \{b\})$, which is nothing but a two-chord tieset with respect to t . Such a two-chord tieset is called a two-chord tieset defined by chords c and c' with respect to t , denoted by $B_{cc'}(t)$, where

$$(5) \quad B_{cc'}(t) = B_{c'}(t \cup \{c\} - \{b\}).$$

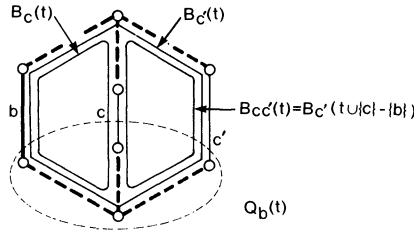


FIG. 2. $B_{cc'}(t)$ and $B_{c'}(t \cup \{c\} - \{b\})$.

Accordingly, we see that we can construct a reduced two-chord tieset matrix $B_r^{(2)}(t)$ satisfying the conditions (a), (b) and (c) if

$$(6) \quad E - t = \bar{t} \subseteq C^{(0)} \cup C^{(1)} \cup \dots \cup C^{(I)},$$

where $I + 1$ is the value of i at the time when the algorithm is terminated. Thus the next problem is to check whether or not (6) is satisfied. Let c_i and c_j be a pair of chords of t , and let B be an elementary tieset of G which contains not only a pair of chords c_i and c_j but also a minimal number of other chords of t . Assume that B is not a two-chord tieset with respect to t . Also, let u_1 and v_1 be the endnodes of c_i and let u_α and v_β be the endnodes of c_j . Then, without loss of generality, $B - \{c_i, c_j\}$ is assumed to be partitioned into two node-disjoint paths P_1 and P_2 such that P_1 and P_2 are paths from u_1 to u_α and from v_1 to v_β , respectively. Also, for convenience, the nodes from u_1 to u_α on P_1 are assumed to be $u_1, u_2, \dots, u_\alpha$, the nodes from v_1 to v_β on P_2 are assumed to be v_1, v_2, \dots, v_β and P_1 and P_2 are denoted by $u_1 u_2 \dots u_\alpha$ and $v_1 v_2 \dots v_\beta$, respectively. Taking into consideration the minimality of the number of chords in B , we easily see that the following are true:

(i) If P_1 contains a chord of t , then for the endnodes, say u_x and u_{x+1} , of the chord there exists no tree-path² from one of nodes u_1, u_2, \dots, u_x to one of nodes u_{x+1}, \dots, u_α . Also, if P_2 contains a chord of t , then a similar conclusion can be reached.

(ii) If P_1 contains a chord of t and P_2 contains at least one edge, then for the endnodes, say u_x and u_{x+1} , of the chord and for any node $v_y (2 \leq y \leq \beta)$ there do not

² By a tree-path of "tree t " is meant a path consisting of branches of t .

exist a pair of node-disjoint paths P'_1 and P'_2 such that P'_1 is a tree-path from one of the nodes u_1, u_2, \dots, u_x to one in v_y, \dots, v_β and P'_2 is a tree-path from one of the nodes u_{x+1}, \dots, u_α to one in v_1, v_2, \dots, v_{y-1} . Also, if P_2 contains a chord of t and P_1 contains at least one edge, then a similar conclusion can be reached.

Now, let G^* be a graph obtained from G by contracting all edges in $B \cap t$. Then it is evident that $t - B$ is a tree of G^* and that $B \cap \bar{t}$ is an elementary tieset of G^* , which contains not only a pair of chords c_i and c_j but also a minimal number of other chords of $t - B$. Also, as is easily seen from the property (ii), any tieset and path in G can be revived uniquely from their corresponding tieset and path in G^* . By taking account of $B \cap t$ in G^* , we have the following two cases to consider.

Case 1. The chords c_i and c_j have a common endnode in G^* . In this case, a tieset $B \cap \bar{t}$ in G^* contains at least three chords c_i, c_j and c_k as illustrated in (a) of Fig. 3, where u'_1 is a common endnode of c_i and c_j , v'_1 is a common endnode of c_i and c_k , v'_2 is the other endnode of c_k , and there exists a path from node v'_2 to node v'_β whose edges are all in $B \cap \bar{t} - \{c_i, c_j, c_k\}$. Since any two nodes can be uniquely connected by a unique tree-path it follows from the property (i) that there exists a pair of tree-paths \tilde{P}_1 and \tilde{P}_2 of t in G such that $\tilde{P}_1 \cap (\bar{t} \cup (t - B))$ and $\tilde{P}_2 \cap (\bar{t} \cup (t - B))$ are tree-paths in G^* from u'_1 to v'_1 and from v'_2 to u'_1 , respectively. Hence we see that G has not only

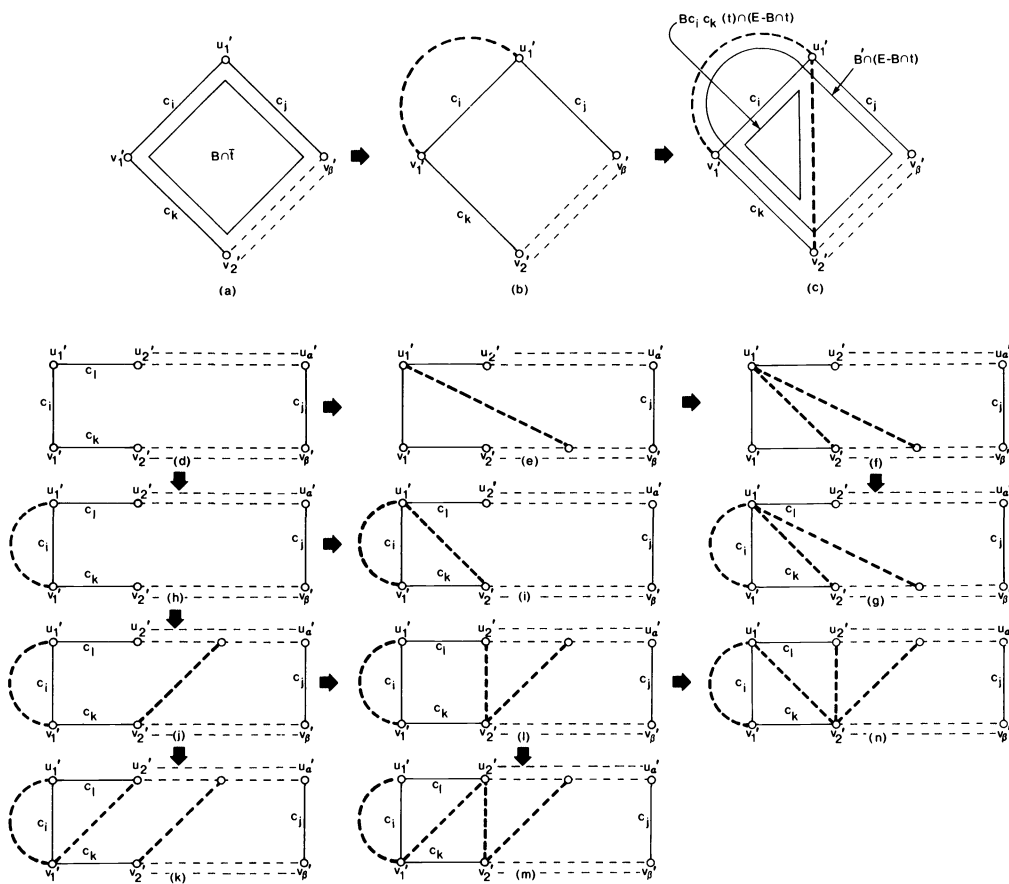


FIG. 3. Graphs used for the proof of Theorem 2.

a two-chord tieset $B_{c_i c_k}(t)$ but also a tieset B' containing a pair of chords c_k and c_j such that the number of chords in B' is smaller than the number of chords in B by 1. Note that such a tieset B' is also a tieset containing not only a pair of chords c_k and c_j but also a minimal number of other chords of t .

Case 2. The chords c_i and c_j have no common endnode in G^* . In this case, a tieset $B \cap \bar{t}$ in G^* contains at least four distinct chords c_i, c_j, c_k and c_l as illustrated in (d) of Fig. 3, where u'_1 and v'_1 are the endnodes of c_i, u'_α and v'_β are the endnodes of $c_j, u'_1 u'_2 \cdots u'_\alpha$ is a path from u'_1 to u'_α and $v'_1 v'_2 \cdots v'_\beta$ is a path from v'_1 to v'_β . From the property (i), it follows that there exists either a tree-path from u'_1 to v'_1 as shown in (h) of Fig. 3 or a tree-path from u'_1 to one of nodes v'_2, \dots, v'_β as shown in (e) of Fig. 3.

Subcase 2.1. There exists a tree-path from u'_1 to one of nodes v'_2, \dots, v'_β as shown in (e) of Fig. 3. In this case, it follows from the properties (i) and (ii) that there exist a pair of tree-paths \tilde{P}_1 and \tilde{P}_2 of t in G such that $\tilde{P}_1 \cap (\bar{t} \cup (t - B))$ and $\tilde{P}_2 \cap (\bar{t} \cup (t - B))$ are tree-paths in G^* from v'_1 to u'_1 and from v'_2 to v'_1 , respectively, as shown in (e), (f) and (g) of Fig. 3. Hence G has not only a two-chord tieset $B_{c_i c_k}(t)$ with respect to t but also a tieset B' containing a pair of chords c_k and c_j such that the number of chords in B' is smaller than the number of chords in B by 1.

Subcase 2.2. There exists a tree-path from u'_1 to v'_1 as shown in (h) of Fig. 3. In this case, from the property (i), it follows that there exists either a tree-path from u'_2 to u'_1 as shown in (i) of Fig. 3 or a tree-path from u'_2 to one of nodes $u'_2 \cdots, u'_\alpha$ as shown in (j) of Fig. 3. If there exists a tree-path from u'_2 to u'_1 as shown in (i) of Fig. 3, then a similar conclusion as in the Subcase 2.1 is reached. If there exists a tree-path from u'_2 to one of nodes u'_2, \dots, u'_α as shown in (j) of Fig. 3, then it follows from the properties (i) and (ii) that there exists either a tree-path from u'_2 to v'_1 as shown in (k) of Fig. 3 or a tree-path from u'_2 to v'_2 as shown in (l) of Fig. 3. If there exists a tree-path from u'_2 to v'_1 as shown in (k) of Fig. 3, then this case is handled in the same way as Subcase 2.1. On the other hand, if there exists a tree-path from u'_2 to v'_2 as shown in (l) of Fig. 3, it follows from the properties (i) and (ii) that there exists either a tree-path from v'_1 to u'_2 as shown in (m) of Fig. 3 or a tree-path from v'_2 to u'_1 as shown in (n) of Fig. 3. However, these two situations as shown in (m) and (n) of Fig. 3 are also handled in the same way as the Subcase 2.1.

Thus, it follows from the mathematical induction that there exists a consecutive sequence of two-chord tiesets such as $B_{c_i c_k}(t), B_{c_k c_l}(t), \dots, B_{c_m c_j}(t)$ for any pair of chords c_i and c_j of t . Hence we can construct a reduced two-chord tieset matrix $B_r^{(2)}(t)$ satisfying the conditions (a), (b) and (c). This completes the proof of the theorem.

The converse of this theorem will be proved in the following way.

THEOREM 2. *If there exists a tree t of G such that*

$$(7) \quad \text{rank } [B^{(2)}(t)] = \mu,$$

then G is a non-series-parallel graph.

Proof. Suppose that there exists a tree t of G such that (7) is satisfied. Then, without loss of generality, we can assume that a reduced two-chord tieset matrix $B_r^{(2)}(t)$ has the form

$$(8) \quad B_r^{(2)}(t) = \begin{bmatrix} S & T \end{bmatrix}$$

\swarrow
 \searrow

\rightarrow columns corresponding to chords of t
 \rightarrow columns corresponding to branches of t

such that

$$(9) \quad \tilde{S} = \begin{bmatrix} \tilde{S}^{(i_1)} & 0 & \cdots & 0 & 0 \\ 0 & \tilde{S}^{(i_2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \tilde{S}^{(i_q)} & 0 \\ \text{hatched} & \text{hatched} & \cdots & \text{hatched} & \begin{matrix} \pm 1 & 0 \\ \vdots & \ddots \\ 0 & \pm 1 \end{matrix} \end{bmatrix},$$

where every row of \tilde{S} contains exactly two nonzero elements (+1 or -1), ± 1 represents either +1 or -1 and the k th block-diagonal square submatrix $\tilde{S}^{(i_k)}$ is

$$(10) \quad \tilde{S}^{(i_k)} = \left. \begin{matrix} \overbrace{\begin{matrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & & 0 \\ \vdots & & \ddots & & \vdots \\ & & & \ddots & 0 \\ 0 & 0 & & 1 & 1 \\ \varepsilon(i_k) & 0 & \cdots & 0 & 1 \end{matrix}}^{i_k \text{ columns}} \right\} i_k \text{ rows},$$

where $\varepsilon(i_k)$ is +1 if i_k is odd and -1 if i_k is even [2]. Note that \tilde{S} does not include $\tilde{S}^{(2)}$ because a pair of chords c_i and c_j with respect to a tree t of G defines a unique two-chord tieset $B_{c_i c_j}(t)$ with respect to the tree of G where we do not consider a two-chord tieset obtained from $B_{c_i c_j}(t)$ merely by reversing its direction.

G is called a series-parallel graph if G is not a non-series-parallel graph. Now, suppose that G is a series-parallel graph. Let c_1, c_2, \dots, c_m be the chords of a tree t of G , and assume that there exists a sequence of two-chord tiesets $B_{c_1 c_2}(t), B_{c_2 c_3}(t), \dots, B_{c_{m-1} c_m}(t), B_{c_1 c_m}(t)$ with respect to t such that the directions of chords c_{j-1} and c_j are chosen to agree with the direction of the two-chord tieset $B_{c_{j-1} c_j}(t)$, where $j = 2, 3, \dots, k$. As will be easily checked, a necessary and sufficient condition for a two-chord tieset defined by a pair of chords c_i and c_j with respect to a tree t to exist in G is that there exists a fundamental cutset with respect to t which contains the pair of chords c_i and c_j . Therefore, $B_{c_1 c_2}(t)$ can be drawn as shown in (a) of Fig. 4, where the thin lines represent chords of t and the thick dotted lines represent tree-paths of t . Now, since G is a series-parallel graph, any elementary cutset of G can be transformed into an incidence set of one of its 2-isomorphic graphs, where G' is said to be 2-isomorphic to G if G' has not only the same set of edges but also the same set of cutsets as G . (See Lemma 1 which will be given in the appendix.) Thus, we can regard the relationship between $B_{c_1 c_2}(t)$ and a pair of chords c_1 and c_2 as shown

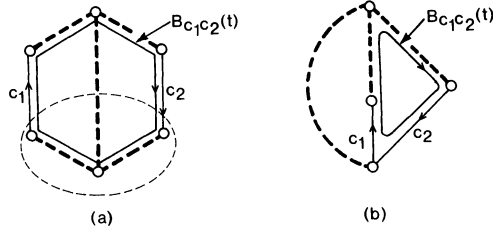


FIG. 4. Relationships between $B_{c_1c_2}(t)$ and a pair of chords c_1 and c_2 .

in (b) of Fig. 4. Also, the possibilities for the existence of $B_{c_1c_2}(t)$ and $B_{c_2c_3}(t)$ are listed in Fig. 5. In any case shown in (a), (b) and (c) of Fig. 5 there exists a two-chord tieset $B_{c_1c_3}(t)$, but in the case shown in (d) of Fig. 5 there does not exist a two-chord tieset $B_{c_1c_3}(t)$. That is to say, in any case shown in (a), (b) and (c) of Fig. 5 there exists a fundamental cutset with respect to t which contains chords c_1 , c_2 and c_3 , but in the case shown in (d) of Fig. 5 there does not exist a fundamental cutset with respect to t which contains chords c_1 and c_3 . If a set of chords c'_1, c'_2, \dots, c'_l , which defines a sequence of two-chord tiesets $B_{c'_1c'_2}(t), B_{c'_2c'_3}(t), \dots, B_{c'_{l-2}c'_{l-1}}(t), B_{c'_{l-1}c'_l}(t)$, is contained in a fundamental cutset with respect to t , then for any two of chords c'_1, c'_2, \dots, c'_l there exists the corresponding two-chord tieset and the sequence $(B_{c'_1c'_2}(t), \dots, B_{c'_{l-1}c'_l}(t))$ is said to be *transitive*. For such a transitive sequence, $B_{c'_1c'_l}(t)$ is called the representative two-chord tieset of the sequence. If the directions of chords c'_{i-1} and c'_i are chosen to agree with the direction of the two-chord tieset $B_{c'_{i-1}c'_i}(t)$, where $i = 2, 3, \dots, l$, then c'_1 and c'_l have the same direction in $B_{c'_1c'_l}(t)$ if l is even and the opposite directions in $B_{c'_1c'_l}(t)$ if l is odd. For a sequence $(B_{c_1c_2}(t), B_{c_2c_3}(t), \dots, B_{c_{m-1}c_m}(t), B_{c_1c_m}(t))$ of two-chord tiesets, let us obtain a sequence

$$(B_{c_1c_{k_2}}(t), B_{c_{k_2}c_{k_3}}(t), \dots, B_{c_{k_{r-1}}c_m}(t), B_{c_1c_m}(t))$$

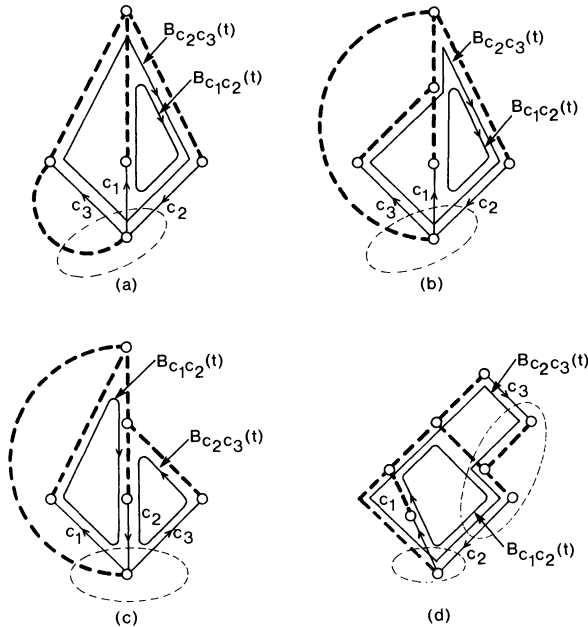


FIG. 5. Relationships between $B_{c_1c_2}(t)$ and $B_{c_2c_3}(t)$.

of two-chord tiesets by replacing maximal transitive subsequences by their representative two-chord tiesets, and let $B^{(2)}(t)$ have the rows corresponding to the two-chord tiesets $B_{c_1c_2}(t), \dots, B_{c_{m-1}c_m}(t)$ and $B_{c_1c_m}(t)$. Then we see that $B^{(2)}(t)$ has

$$(11) \quad S_x = \left[\begin{array}{cccccc} c_1 & c_2 & c_3 & & c_{m-1} & c_m \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & & & 1 & 1 \\ \varepsilon_x & 0 & \cdots & & 0 & 1 \end{array} \right] \left. \vphantom{\begin{array}{cccccc} c_1 & c_2 & c_3 & & c_{m-1} & c_m \end{array}} \right\} m \text{ rows}$$

and

$$(12) \quad S'_x = \left[\begin{array}{cccccc} c_1 & c_{k_2} & c_{k_3} & & c_{k_{r-1}} & c_m \\ 1 & (-1)^{k_2} & 0 & \cdots & 0 & 0 \\ 0 & 1 & (-1)^{k_3-k_2+1} & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & & & 1 & (-1)^{m-k_{r-1}+1} \\ \varepsilon_x & 0 & \cdots & & 0 & 1 \end{array} \right] \left. \vphantom{\begin{array}{cccccc} c_1 & c_{k_2} & c_{k_3} & & c_{k_{r-1}} & c_m \end{array}} \right\} r \text{ rows}$$

as its square submatrices. Now by reversing the direction of chords and two-chord tiesets appropriately, S'_x in $B^{(2)}$ can be changed into

$$(13) \quad S''_x = \left[\begin{array}{cccccc} c_1 & c_{k_2} & c_{k_3} & \cdots & c_{k_{r-1}} & c_m \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & & & 1 & 1 \\ \varepsilon'_x & 0 & \cdots & & 0 & 1 \end{array} \right] \left. \vphantom{\begin{array}{cccccc} c_1 & c_{k_2} & c_{k_3} & \cdots & c_{k_{r-1}} & c_m \end{array}} \right\} r \text{ rows,}$$

where

$$(14) \quad \varepsilon'_x = (-1)^{k_2} \times (-1)^{k_3-k_2+1} \times \cdots \times (-1)^{m-k_{r-1}+1} \varepsilon_x = (-1)^{m+r-2} \varepsilon_x.$$

By repeated application of replacing maximal transitive subsequences by their representative two-chord tiesets, we finally obtain a sequence $(B_{c_1c_p}(t), B_{c_p c_m}(t), B_{c_1c_m}(t))$ of two-chord tiesets. Thus we see that $B^{(2)}(t)$ has

$$(15) \quad S'''_x = \left[\begin{array}{ccc} c_1 & c_p & c_m \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ \varepsilon''_x & 0 & 1 \end{array} \right]$$

as its square submatrix, where

$$(16) \quad \varepsilon''_x = (-1)^{m+r-2} \times (-1)^{r+s-2} \times \cdots \times (-1)^{w+3-2} \varepsilon_x = (-1)^{m+1} \varepsilon_x.$$

Since G is a series-parallel graph, it follows from (a), (b) and (c) of Fig. 5 that ε''_x is always equal to -1 . Thus we get

$$(17) \quad \varepsilon_x = (-1)^m$$

from which it follows that there holds

$$\det [S_x] = 0$$

regardless of whether m is even or odd. This is in contradiction to the hypothesis that $\tilde{S}^{(m)}$ for some m ($m \geq 3$) is contained in $B_r^{(2)}(t)$. This completes the proof of this theorem.

Combining Theorems 1 and 2, we get the main theorem of this paper as follows.

THEOREM 3. *A necessary and sufficient condition for G to be a non-series-parallel graph is that there exists a tree t of G such that*

$$\text{rank } [B^{(2)}(t)] = \mu.$$

As a consequence, the following corollary is obtained.

COROLLARY 1. *A necessary and sufficient condition for G to be a series-parallel graph is that for every tree t of G*

$$\text{rank } [B^{(2)}(t)] < \mu.$$

3. Conclusions. In this paper, we have given a necessary and sufficient condition for a graph G to be non-series-parallel by use of the rank condition of a two-chord tieset matrix. This condition is a new characterization of the non-series-parallel graphs. Dually, if we use the concept of two-branch cutsets which are defined to be cutsets containing exactly two branches of a tree of G , we obtain another condition for G to be a non-series-parallel graph.

Appendix. **LEMMA 1.** *Any elementary cutset of a graph G can be transformed into an incidence set of one of its 2-isomorphic graphs if and only if G is a series-parallel graph.*

Proof. G is a series-parallel graph if and only if every minor of G is not an oriented K_4 . Since an elementary cutset consisting of exactly four edges of any orientation of K_4 cannot be transformed into an incidence set of one of its 2-isomorphic graphs, "only if" part is evident. Then "if" part will be proved below.

Suppose that G is a series-parallel graph. Then, since G is planar, G has its dual, denoted by G^* , and G^* is also a series-parallel graph. Accordingly, any elementary cutset of G is an elementary tieset of G^* and vice versa. Let Q be an elementary cutset of G . Then Q is an elementary tieset of G^* .

Now assume that G^* cannot be drawn on a plane without intersection of its edges so that Q is a mesh. Then there must exist four distinct nodes u, v, w and x on Q in G^* such that a path from u to w and a path from v to x intersect in the outside of a closed region surrounded by Q . This means that a minor of G^* is an oriented K_4 . This is a contradiction. Hence G^* can be drawn on a plane without intersection of its edges so that Q is a mesh. Hence Q is an incidence set of a graph 2-isomorphic to G , which is a dual of G^* .

We note that another proof of this lemma is given in [4].

REFERENCES

- [1] R. J. DUFFIN, *Topology of series-parallel graphs*, J. Math. Appl., 10 (1965), pp. 303–318.
- [2] S. SHINODA, K. ONAGA AND W. MAYEDA, *Graph-theoretic properties of a pseudo-incidence matrix with an application to network diagnosis*, Conference Record, 12th Asilomar Conference on Circuits, Systems and Computers, November 6–8, 1978, Pacific Grove, CA, pp. 749–753.
- [3] W. K. CHEN, *Applied Graph Theory*, 2nd ed., North-Holland, Amsterdam, 1976.
- [4] H. KAKITANI AND O. KAKUSHO, *Modularity of tieset matrices and a characterization of series-parallel graphs*, Papers of Technical Group on Circuit and System Theory of IECE of Japan, CST78-21, 1978, pp. 23–30.

THE MAXIMUM COVERAGE LOCATION PROBLEM*

NIMROD MEGIDDO,[†] EITAN ZEMEL[‡] AND S. LOUIS HAKIMI[§]

Abstract. In this paper we define and discuss the following problem which we call the maximum coverage location problem. A transportation network is given together with the locations of customers and facilities. Thus, for each customer i , a radius r_i is known such that customer i can currently be served by a facility which is located within a distance of r_i from the location of customer i . We consider the problem from the point of view of a new company which is interested in establishing new facilities on the network so as to maximize the company's "share of the market." Specifically, assume that the company gains an amount of w_i in case customer i decides to switch over to one of the new facilities. Moreover, we assume that the decision to switch over is based on proximity only, i.e., customer i switches over to a new facility only if the latter is located at a distance less than r_i from i . The problem is to locate p new facilities so as to maximize the total gain.

The maximum coverage problem is a relatively complicated one even on tree-networks. This is because one aspect of the problem is the selection of the *subset* of customers to be taken over. Nevertheless, we present an $O(n^2p)$ algorithm for this problem on a tree. Our approach can be applied to other similar problems which are discussed in the paper.

1. Introduction. We shall discuss in this paper problems in which establishing new facilities in an existing network is aimed at attracting a maximum number of customers. There is thus some competitive flavor to such problems in that the existing facilities may belong to one company while a second company is trying to extract the maximum profit by locating its own facilities on the same network. For further discussion of this and related problems the reader is referred to [H2].

Consider a graph $G = (V, E)$ with edge-lengths d_{ij} . We identify each edge (i, j) of G with a line-segment of length d_{ij} so that we can speak of points (not necessarily vertices) on the edges of G . Each such point x is identified by its distances from the endpoints of the appropriate edge. For every two points x, y of G let $d(x, y)$ be the length of the shortest path between them along the edges of G .

We associate a "customer" with each vertex of G . We assume that there exists a threshold radius r_i for each customer i so that if a new facility is established within a distance of r_i from i then this customer would start using the new facility (unless, of course, an even closer facility is established; in any case, a customer uses one of the closest facilities). We say in such a case that customer i is "covered," with a resulting gain which we denote by w_i (the "weight" of i). The results of this paper are valid for any set of positive constants r_i . However, several simplifications are possible if the threshold radii are derived from distances to old facilities already positioned on G . This topic is addressed in Appendix 2.

Assume that p facilities can be located anywhere on G (including points on the edges other than vertices). We wish to locate these facilities so as to maximize the total weight of the customers which are covered. We will show that it is easy to identify a fairly small subset of points $Y = \{y_1, \dots, y_m\}$ such that we need consider only points

* Received by the editors September 1, 1981, and in revised form July 1, 1982. This work was supported in part by the National Science Foundation under grants ECS7909724, SOC7905900, ENG7902506 and ECS8121741.

[†] Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel.

[‡] J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60201. The research of this author was partially supported by the J. L. Kellogg Center for Advanced Studies in Managerial Economics.

[§] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60201.

of Y as possible location for facilities. Alternatively, the problem may be originally specified with respect to a finite set of feasible location sites.

Our maximum coverage problem is obviously NP-hard on a general graph since the problem of minimum dominating set (see [GJ]) can be formulated as that of minimizing the number of facilities required to gain $W = n$ (where n is the number of vertices). To that end we set $r_i = 1.5$, $w_i = 1$, $d_{ij} = 1$ for all i, j .

We shall present a polynomial-time algorithm for the maximum coverage problem on a tree. We note that unlike the problem of minimum dominating set on a tree (which is easily solvable in linear time), or even that of gaining the total weight (i.e., covering all the vertices), the existence of a polynomial-time algorithm for our problem is nontrivial. The relative difficulty is due to the fact that here we do not require covering of *all* vertices but only k of them (in the special case of unit weights and $W = k$, say). Thus, the large number of different combinations of k out of n complicates the problem. Further evidence to the difficulty of the problem even on a tree is given by the fact that, unlike in many other locational problems on trees (see [T], [K]), the integer linear programming problem associated with ours is *not* solvable as a regular linear program (as we demonstrate in Appendix 1).

In § 2 we discuss the set of potential points for the construction of new facilities. The basic routines of the algorithm are described in § 3. The algorithm itself is explained in § 4. In § 5 we discuss the complexity of the algorithm. Section 6 discusses further problems, related to the maximum coverage problem, which are solvable by a modified version of our algorithm. Appendix 1 describes the linear programming aspects and Appendix 2 discusses simplifications in the case where all threshold radii are implied from distances to (old) existing facilities.

2. Potential locations for new facilities. As we stated in the introduction, a new facility may be constructed at any point of the graph. We denote by $d(x, y)$ the distance (along the shortest route) between any pair of points (x, y) . However, we shall identify a fairly small finite set from which an optimal combination can be selected. Our algorithm can also be applied to problems in which new facilities can be constructed at designated points, y_1, \dots, y_m , only.

Let U_i be the r_i -neighborhood of vertex i , i.e., U_i is the set of all points x such that the distance between i and x is *less* than r_i . For every set S of vertices, let $U_S = \bigcap_{i \in S} U_i$. We say that U_S is *maximal* if $U_S \neq \emptyset$ and $U_T = \emptyset$ for every $T \supsetneq S$. Obviously, we may assume without loss of generality that a new facility will always belong to some maximal U_S . Moreover, we may select in advance a single point $y_S \in U_S$ from each maximal U_S and consider only these points y_S for locations of new facilities. We shall now prove that the number of maximal U_S 's is not too large.

THEOREM. *On a general graph with e edges and n vertices the number of maximal U_S 's is $O(en)$ while on a tree this number is $O(n)$.*

Proof. First, note that if x is a boundary point of a set U_S then there exists a vertex i such that $d(i, x) = r_i$. Thus, each vertex i can contribute at most two boundary points on each edge of the graph. Moreover, i cannot contribute the same boundary point to more than one maximal U_S . It follows that a single edge contains no more than $2n$ boundary points and hence the total number of boundary points is at most $2en$. This establishes the first claim of the theorem.

Consider now the case where G is a tree T . Note first that there are at most n maximal U_S 's which contain a vertex, since these sets are pairwise disjoint. It thus suffices to show that the number of maximal U_S 's which do not contain any vertex is $O(n)$. Every such U_S has precisely two boundary points x_i, x_j such that i is a vertex

with $d(i, x_i) = r_i > d(i, x_j)$ and j is a vertex with $d(j, x_j) = r_j > d(j, x_i)$. The set U_S could thus be identified with an interval (x_i, x_j) . By removing this interval from T , our tree decomposes into two subtrees T_i, T_j , such that for every point $x \neq x_i$ in $T_i, d(i, x) > r_i$ and for every point $x \neq x_j$ in $T_j, d(j, x) > r_j$ (see Fig. 1). Thus, if vertex i contributes

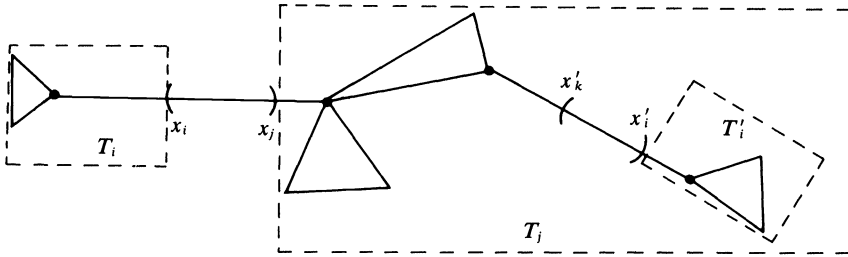


FIG. 1

a boundary point to another such interval, then that interval must be completely contained in T_j . Suppose that there is such an interval (x'_i, x'_k) , where $d(i, x'_i) = r_i > d(i, x'_k)$ and k is a vertex such that $d(k, x'_k) = r_k > d(k, x'_i)$. It is easy to verify that there can be no interval to which both j and k contribute two distinct boundary points. In general, consider a graph G^* on the vertices of T such that vertices u and v are linked with an edge in G^* if and only if there exists an “interval” to which both u and v contribute distinct boundary points as previously described. Then, it can be proved by induction that G^* has no cycles. This implies that the number of those intervals is not greater than $n - 1$ and that completes the proof.

We note that the determination of all the boundaries of the maximal U_S 's can be easily carried out in $O(n^2)$ time.

3. The basic routines. The algorithm for the maximum coverage problem on a tree works in general as follows. Suppose that the potential locations are at the points y_1, \dots, y_m . To simplify the presentation, let us assume that the potential locations are precisely the vertices themselves. Since $m \leq 2n$ it follows that by adding all the y_j 's as vertices we do not lose in terms of the asymptotic complexity.

Let the tree now be rooted at an arbitrary vertex u . If u is selected for a location of a new facility then we proceed, recursively, by looking at the subtrees rooted at the neighbors of u , taking into consideration the fact that there has been a new facility established at u . This requires, however, the solution of a “resource allocation” problem, i.e., optimizing the distribution of the $p - 1$ remaining facilities among the subtrees rooted at the neighbors of u . If u is not selected as a location of a new facility, we proceed, recursively, to the subtrees and then we have to consider the interactions between these subtrees caused by the fact that vertices of one subtree may be covered by a facility located in another.

In order to overcome all these difficulties we define the following routines:

1. **EXT** (T, π, r). Here T is a rooted tree with parameters d_{ij}, r_i, w_i as explained in the introduction, π is an integer and r is a nonnegative number. The routine **EXT** finds the *maximal* total weight of vertices in T that can be gained by locating π new facilities in T , given that there is one additional facility outside of T at a distance r from the root. Thus, this gain consists of the total weight of vertices i , such that the distance between i and the root is less than $r_i - r$, plus the total weight of *other* vertices i such that the distance between i and one of the π new facilities is less than r_i .

2. $\text{Int}(T, \pi, r)$. With T and π as in $\text{EXT}(T, \pi, r)$ this routine solves the maximum coverage problem on T with π new facilities but with an additional requirement that at least one new facility *must* be located at a distance less than or equal to r from the root of T .

It is easy to verify that the routine EXT has at most n critical values for the parameter r , namely, the differences $\delta_i = r_i - d(i, u_T)$ (where u_T is the root of T), for all vertices i . In other words, it suffices to know the output of EXT for each δ_i in order to know that output for all values of r . Analogously, INT has just the distances $d(i, u_T)$ as the critical values for the parameter r .

3. $\text{ALLOC}(f_1, \dots, f_k; \pi)$. This is a routine that solves the resource allocation problem with concave returns. Specifically, let f_1, \dots, f_k be monotone concave functions of discrete variables, and let π be a nonnegative integer. Then, ALLOC solves the following:

$$\begin{aligned} &\text{Maximize} && \sum_{i=1}^k f_i(p_i) \\ &\text{subject to} && \sum_{i=1}^k p_i = \pi \\ &&& p_i \text{ is a nonnegative integer.} \end{aligned}$$

Fast algorithms for ALLOC have been proposed by Galil and Megiddo [GM] and by Frederickson and Johnson [FJ]. The latter requires $O(k \log \pi)$ evaluations of the functions f_i . On the other hand, if one wishes to know the solutions of all problems for $\pi = 1, 2, \dots, p$ (with f_1, \dots, f_k fixed), then only kp evaluations are required in addition to $O(p \log k)$ time for running the greedy algorithm.

4. The algorithm. The routines EXT and INT described in § 3 require recursive calls to each other. Our maximum coverage problem could be solved by either of these routines, if r is chosen sufficiently large.

We use the following notation. The tree T is rooted at the vertex u whose “sons” are u_1, \dots, u_k . For $i = 1, \dots, k$, T_i is the subtree rooted at u_i . Let n_i denote the number of vertices in T_i and let $d_1^i \leq \dots \leq d_{n_i}^i$ denote the distances of vertices of T_i from u_i .

We first describe the routine $\text{INT}(T, \pi, r)$. There are two cases to examine:

Case (i). A facility is established at u . Let $f_i(p_i) = \text{EXT}(T_i, p_i, d(u_i, u))$, $i = 1, \dots, k$. It is easy to verify that the f_i 's are monotone and concave. Obviously, in this case the total gain is $w_u + \text{ALLOC}(f_1, \dots, f_k; \pi - 1)$.

Case (ii). No facility is established at u . Here, the routine INT considers k different subcases. In a typical subcase, a subtree $T_j (1 \leq j \leq k)$ is selected and for each $\rho \in \{d_1^j, \dots, d_{n_j}^j\}$ such that $\rho + d(u_j, u) \leq r$, the following is considered. For every $i \neq j$ ($i = 1, \dots, k$) let $f_i(p_i) = \text{EXT}(T_i, p_i, \rho + d(u_j, u_i))$. Also, let $f_j(p_j) = \text{INT}(T_j, p_j, \rho)$. Again, the functions f_i are monotone and concave. Note that these functions depend on the parameter ρ . Now define

$$A_j(\rho) = \text{ALLOC}(f_1, \dots, f_k; \pi) + w_u \delta(\rho),$$

where $\delta(\rho) = 1$ if $\rho + d(u_j, u) < r_u$ and $\delta(\rho) = 0$ otherwise. Let $A_j = \text{Max}_\rho \{A_j(\rho)\}$, $j = 1, \dots, k$.

The routine INT returns either the maximum of the A_j 's or the optimal value of case (i), whichever is the larger.

The computation of $\text{EXT}(T, \pi, r)$ is analogous. Again, two cases are distinguished.

Case (i). A facility is established within a distance of r from u . This, by definition, is identical with the situation solved by $\text{INT}(T, \pi, r)$.

Case (ii). No facility can be established within a distance of r from the root. Note that in this case, since there is a facility already located at a distance r from the root, there are no interactions among the subtrees. On the other hand, such a constrained problem cannot be solved directly by our routines. Instead, we solve the relaxed problem (i.e., we remove the requirement of *not* constructing a facility within a distance of r from the root) but ignore the interactions among the subtrees. Specifically, let $f_i(p_i) = \text{EXT}(T_i, p_i, d(u_i, u) + r)$, $i = 1, \dots, k$. Let $A = \text{ALLOC}(f_1, \dots, f_k; \pi) + w_u \cdot \delta(r)$ (where the δ is as in the description of INT).

We now claim that $\text{EXT}(T, \pi, r) = \max\{A, \text{INT}(T, \pi, r)\}$. To see this note that if one of the π optimal locations for the problem solved by $\text{EXT}(T, \pi, r)$ is at a distance of less than or equal to r from u then $\text{EXT}(T, \pi, r) = \text{INT}(T, \pi, r)$ and $A \leq \text{INT}(T, \pi, r)$. Otherwise, we are in Case (ii) and no interactions exist among the subtrees. Therefore, $\text{EXT}(T, \pi, r) = A$ and $\text{INT}(T, \pi, r) \leq A$.

5. The complexity of the algorithm. Suppose that $0 = d_0 \leq d_1 \leq \dots \leq d_{n_u}$ are the distances of vertices of T from u . Consider the function $g(r) = \text{INT}(T, \pi, r)$, where π is fixed. Obviously, g is a step-function with jumps only at $r = d_s (0 \leq s \leq n_u)$. Moreover, if $d_s (s \geq 1)$ is a distance from a vertex in T_j , then

$$\text{INT}(T, \pi, d_s) = \max[\text{INT}(T, \pi, d_{s-1}), \text{ALLOC}(f_1, \dots, f_k; \pi) + w_u \cdot \delta(u)],$$

where $f_j(p_j) = \text{INT}(T_j, p_j, d_s - d(u, u_j))$ and for $i \neq j$, $f_i(p_i) = \text{EXT}(T_i, p_i, d_s + d(u, u_i))$. This implies that when π is given, the evaluation of $\text{INT}(T, \pi, r)$ for all critical values of r takes $O(n)$ computations of resource allocation. Similarly, if $\delta_0 \leq \delta_1 \leq \dots \leq \delta_{n_u}$ are the sorted values of $r_x - d(x, u)$, where x is a vertex in T , then the critical values of r in $\text{EXT}(T, \pi, r)$ are in the set $\{d_0, \dots, d_{n_u}, \delta_0, \dots, \delta_{n_u}\}$; i.e., at a critical value either $r = d(x, u)$ or $d(x, u) + r = r_x$ for some vertex x in T . Since $\text{EXT}(T, \pi, r) = \max[\text{INT}(T, \pi, r), A]$, it follows that the evaluation of $\text{EXT}(T, \pi, r)$ for all critical values of r (where π is fixed) also requires only $O(n)$ computations of resource allocation.

If the algorithm is recursively run as stated in § 4, then it requires superpolynomial time. However, this is only because the same subproblems are being solved over and over again in such an implementation. To avoid that, one just has to be careful not to compute the same thing more than once. Specifically, if we store the results of all computations then we run in polynomial-time by the following argument. The number of different problems that either EXT or INT has to solve is $O(n^2 p)$. This is because there are $O(n)$ subtrees to be considered, each with $O(n)$ critical values of r , and π may take on only the values $0, 1, \dots, p$. When we have to compute $\text{INT}(T, \pi, r)$, say, then it takes only one computation of resource allocation if all the necessary values that are returned recursively are known. This establishes a bound of $O(n^3 p \log n)$.

A more careful analysis shows that the algorithm can be implemented much faster. Consider the computation of $\text{INT}(T, \pi, r)$ for example, where T is rooted at u and u has k sons. If we solve the necessary allocation problem only for one value of π , then it takes $O(k \log \pi)$ time, once the necessary EXT and INT values are known. However, we can solve the problem relative to all values of π in $O(k \min(p, \log k) + p \log k)$ time. For, once the values $f_i(1) - f_i(0)$, $i = 1, \dots, k$ are sorted, it takes $O(p \log k)$ time to find the p largest marginal gains of the form

$f_i(m+1) - f_i(m)$ ($i = 1, \dots, k, m = 0, \dots, p-1$); the initial sort should be eliminated if $kp < k \log k$, in which case those p largest values can be found in kp steps. Now, the solution of $\text{EXT}(T, \pi, r)$ and $\text{INT}(T, \pi, r)$ for all values of π and r takes $O(n(k \min(p, \log k) + p \log k))$, once the necessary values are known. The total effort is therefore $O(n \sum_u (\deg(u) \min(p, \log \deg(u)) + p \log \deg(u)))$, where the summation is over all the vertices u and $\deg(u)$ is the degree of u . This is however $O(n^2 p)$.

6. Related problems. A natural generalization of the problem treated in § 1 is as follows. Suppose that each vertex i has an additional parameter, c_i , which represents the cost of establishing a new facility at vertex i . We now replace the number p by some budget B and seek to find the maximum coverage subject to this budget. This problem, however, is NP-hard even on chain networks since the knapsack problem can be easily formulated as such a coverage problem. On the other hand, our algorithm can be modified to solve this problem on a tree in pseudo-polynomial time (see [GJ]), i.e., in polynomial time in terms of n and B . Also, the problem of covering a maximum number of vertices given a fixed budget can be solved in polynomial time on a tree by considering the equivalent problem of covering at least q vertices given a budget B . The latter is easily seen to be amendable to an algorithm similar to the one proposed in § 4.

Another related problem is that of covering all the vertices at minimum cost. More formally, we wish to select points at which facilities will be established such that every vertex i has a facility located within a distance r_i from i , and such that the total construction cost is minimized. Tamir [T] solves this problem in $O(n^3)$ time. Kolen [K] solves a related problem, where the threshold radii are associated with the facilities (rather than the demand points) in $O(nm)$ time, where m is the number of potential locations of facilities. Our approach easily provides $O(nm)$ algorithms for both these problems as follows. Define $\text{INT}(T, r)$ to be the minimum total cost of facilities established in T so as to cover all the vertices of T , subject to a constraint that at least one of them has to be located within a distance of r from the root. Let $\text{EXT}(T, r)$ denote the minimum total cost of covering all the vertices of T that are not covered by a facility located outside of T at a distance r from the root. For every u , these functions evaluated at the subtree rooted at u , have $O(m)$ critical values of r . These are simply the distances from u to any potential location of a facility. If $d_0 \leq d_1 \leq \dots \leq d_t$ are the distances from u to such locations in T from which u itself would be covered and if $d_s (s \geq 1)$ is a distance from a vertex in T_j to u , then

$$\text{INT}(T, d_s) = \min [\text{INT}(T, d_{s-1}), \text{INT}(T_j, d_s - d(u, u_j)) + \sum_{\substack{i=1 \\ i \neq j}}^k \text{EXT}(T_i, d_s + d(u, u_i))].$$

For the routine EXT , let $A = \sum_{i=1}^k \text{EXT}(T_i, r + d(u_i, u))$ if $r < r_u$, and $A = +\infty$ otherwise. Then, $\text{EXT}(T, r) = \min [A, \text{INT}(T, r)]$. It follows from the structure of these formulae that it takes $O(mk)$ to solve the problems at u for all values of r . The total effort is therefore $O(m \sum_v \deg(v)) = O(nm)$.

Finally, consider the problem of maximizing the net gain; i.e., the total revenue (resulting from covering nodes) minus the total construction cost. This problem too is NP-hard on a general network since the minimum dominating set (see [GJ]) can be reduced to this problem by taking $w_j = 2, c_j = 1, j = 1, \dots, n$ and $B = k$. However, on a tree the solution is similar to that of the min-cost coverage of all vertices, i.e., the parameter p and the ALLOC routine can be eliminated. That leads to the same $O(nm)$ time bound.

Naturally, the same methods can be used to solve other types of problems defined on tree networks such as those which seek to locate facilities as far apart as possible from the various vertices. For instance, consider the problem of minimum coverage of obnoxious facilities. Here we are seeking to minimize the total weight of vertices who are “damaged” because an “obnoxious” facility [CG], [CT] is located too close, given that we have to locate p such facilities and the interfacility distances are also bounded from below. This can be solved in a way which is very similar to the one developed in the present paper.

Appendix 1. Linear programming considerations. Tamir [T] and Kolen [K] have recently shown that a fairly large class of location problems on trees can be solved by linear programming. Specifically, let $a_{ij} = 1$ if $d(i, j) < r_i$ and $a_{ij} = 0$ otherwise. Let $x_j = 1$ be interpreted as establishing a facility at j and $x_j = 0$ otherwise. Thus, the program

$$\begin{aligned} &\text{Minimize } \sum_{j=1}^n c_j x_j \\ &\text{s.t. } \quad Ax \geq 1 \\ &\quad \quad x_j \in \{0, 1\}. \end{aligned}$$

($A = (a_{ij})$) solves the problem of covering *all* vertices with minimum cost. It is known [G], [K], [T] that for a tree network the matrix A is balanced and hence the polytope $\{x : Ax \geq 1, x \geq 0\}$ has only integral extreme points. On the other hand, our coverage problem can be formulated as maximizing the number of vertices that would be covered by at most p facilities. Thus, if $y_i = 0$ is interpreted as “vertex i is covered” and $y_i = 1$ otherwise, then our problem is in fact

$$\begin{aligned} &\text{Minimize } \sum_{i=1}^n y_i \\ &\text{subject to } Ax + y \geq 1 \\ &\text{s.t. } \quad \sum_{j=1}^n x_j \leq p \\ &\quad \quad x_j, y_i \in \{0, 1\}. \end{aligned}$$

Now, even though the underlying matrix in this problem, namely,

$$\begin{bmatrix} A & I \\ 1 \cdots 1 & 0 \end{bmatrix}$$

is still balanced, a linear programming solution may lead to a nonintegral solution, as we show in the following example. Consider the tree in Fig. 2. All weights $w_i = 1$ and r_i 's and d_{ij} 's are shown in the figure. There are four potential points denoted by arrows (i.e., for every other point y there is one of the four points that covers at least those vertices covered by y).

Consider the problem of maximizing the number of vertices covered by two new facilities. With two “integral” facilities, namely, the center and another point of the four, we can cover at most nine vertices. However, by selecting one “half” of a facility to be located in each one of these four vertices, we manage to cover nine and a “half” vertices.

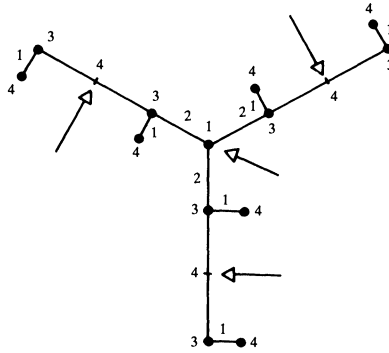


FIG. 2

Appendix 2. Threshold radii arising from distances to existing facilities. Suppose that the radii r_i are in fact the current distances from each vertex i to the nearest existing facility (which belongs to the first company) and the customer located at i would use a new facility if one were established by the second company at a distance less than r_i from i .

We claim that the problem can be decomposed in this case as follows. Consider the connected components of the tree induced by the locations of the existing facilities. Specifically, two points belong to the same component if there is no existing facility on the path which connects them. Let these components be denoted by T_1, \dots, T_k . Obviously, a customer located in T_j would use a new facility only if it is located inside T_j . Thus, it suffices to consider maximum coverage problems on the components and then solve a resource allocation problem as follows. Suppose that $f_i(p_i)$ is the maximum gain possible by locating p_i new facilities in T_i . Then, the solution to our problem is by maximizing $\sum f_i(p_i)$ subject to $\sum p_i = p$ ($p_i \geq 0$ and integral). We note that in each component an existing facility is always located at a leaf. Furthermore, we can assume without loss of generality that every leaf contains a facility. For, assume on the contrary, that a customer i is located at a leaf in which no existing facility is located. Such a customer will switch to a new facility inside the component if and only if its unique neighbor does. Thus, we can eliminate the leaf i from the tree and add its weight to that of its neighbor. Continuing with this process, we eventually get components in which the existing facilities coincide with the leaves.

The case of a chain tree is extremely simple. Here we split the chain into subchains whose boundary points are the locations of the existing facilities. Each subchain should be assigned either 0, 1, or 2 new facilities. Thus, the resource allocation problem that has to be solved in this case is very simple and the problem can be solved by the greedy algorithm in linear time.

REFERENCES

- [CG] R. L. CHURCH AND R. S. GARFINKEL, *Locating an obnoxious facility on a network*, *Transport Sci.*, 1 (1978), pp. 107–118.
- [CT] R. CHANDRASEKARAN AND A. TAMIR, *Locating obnoxious facilities*, Dept. Statistics, Tel-Aviv University, Tel Aviv, Israel, 1979.
- [FJ] G. N. FREDERICKSON AND D. B. JOHNSON, *Optimal algorithms for generating quantile information in $X + Y$ and matrices with sorted columns*, Proc. 13th Ann. Conf. on Information Science and Systems, the Johns Hopkins Univ., Baltimore, 1979, pp. 47–52.

- [GM] 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.
- [GJ] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.
- [G] R. GILES, *A balanced hypergraph defined by subtrees of a tree*, Ars Combinatoria, 6 (1978), pp. 179–183.
- [H1] S. L. HAKIMI, *Optimal location of switching centers and the absolute centers and medians on a graph*, Oper. Res., 12 (1964), pp. 450–459.
- [H2] S. L. HAKIMI, *On locating new facilities in a competitive environment*, presented at ISOLDEII June 15–20, 1981, Skodsborg, Denmark.
- [KH] O. KARIV AND S. L. HAKIMI, *An algorithmic approach to network location problems, Part II: p -Medians*, SIAM J. Appl. Math., 37 (1979), pp. 539–560.
- [K] A. KOLEN, *An $O(nm)$ algorithm for the minimum cost covering problem on a tree with n vertices and m neighborhood subtrees*, ISOLDEII (see [H2]).
- [T] A. TAMIR, *A class of balanced matrices arising from location problems*, Dept. Statistics, Tel Aviv University, Tel Aviv, Israel, 1980.

DETERMINACY IN LINEAR SYSTEMS AND NETWORKS*

J. SCOTT PROVAN†

Abstract. We study interdependent and determinate behavior between variables subject to a system of linear equalities. For each pair of variables in such a system, four definitions of “correlation” are introduced which relate the behavior of the variables to a chosen set of “basic” variables for the system. These definitions correspond directly to such terms as statistical correlation, rates of substitution in economics, sensitivity in linear programming, and sign-solvability in linear algebra. For each definition of correlation, there is a stronger property of *determinacy* between two variables, established by the consistency in sign of the correlation between the two variables over every set of basic variables. We show that the property of determinacy is independent of which definition of correlation is used. We also examine correlation and determinacy in systems related to networks, and derive good characterizations of determinacy in terms of properties of the underlying networks.

Key words. correlation, determinacy, network, sensitivity analysis

1. Introduction. The purpose of this paper is to investigate the concept of “correlation” between variables which are subject to linear constraints. Specifically, given the system of variables x_1, \dots, x_n subject to the linear constraints

$$(1.1) \quad Ax = b,$$

where A is a $m \times n$ matrix and b is an m -vector, what can we say about the activity of variable x_i in relation to that of variable x_j , and can we derive any sense of interdependent or determined behavior between the activity of these variables? The concept of correlation will certainly vary depending upon the context in which such a term is to be defined, and in such diverse fields as linear algebra, operations research, statistics or economics, much attention has been given to formalizing concept of correlated behavior in mathematical models. Many authors, [10], [15], [17], [18] and [20], for example, have looked at determined relationships between the variables and right-hand sides of linear systems along the lines of “qualitative analysis” suggested by Samuelson [22]. The only attempt to describe internal correlation between variables in these systems, however, appears to be by Greenberg in [11]. His definition of “qualitative determinacy” is based on activity occurring in the basic tableaux associated with the linear system. We present in this paper several definitions of correlation and determinacy in the spirit of—and including that of—Greenberg, which appear in various contexts in all of the fields mentioned above. These definitions turn out to be related to such a surprising extent that techniques used to study one type of correlation can be applied to all of them. In particular, we examine correlation and determinacy in linear systems related to networks where graph theoretic techniques can be used effectively in uncovering correlation in these systems.

We first need to establish some notation. Define $P(A, b)$ as the system given by (1.1). Denote by a_{ij} , A_i and A^j the elements, rows and columns of A for $i = 1, \dots, m$ and $j = 1, \dots, n$, and denote by A^T the transpose of A . A *basis* for $P(A, b)$ consists of a subset $B = (A^{B_1}, \dots, A^{B_m})$ of m columns of A which forms a nonsingular matrix. The set $x_B = (x_{B_1}, \dots, x_{B_m})$ is called the set of *basic variables*. The set x_N of remaining variables is called the set of *nonbasic variables*, and N will denote the corresponding

* Received by the editors February 25, 1982, and in revised form August 2, 1982. This research was supported by an NRC/NAS Postdoctoral Associateship.

† National Bureau of Standards, Washington, DC 20234, and University of North Carolina at Chapel Hill, North Carolina 27514.

matrix of columns. We will often take B and N to represent index sets as well as sets of columns. The *basic solution* (x_B^*, x_N^*) corresponding to B in $P(A, b)$ consists of setting

$$x_B^* = B^{-1}b, \quad x_N^* = 0.$$

Finally, the *basic tableau* corresponding to B is defined to be the matrix

$$\bar{A} = B^{-1}A$$

so that an equivalent system to $P(A, b)$ (in terms of the solution set of X) is

$$\bar{A}x = \bar{b} = B^{-1}b$$

or

$$(1.2) \quad x_B + B^{-1}Nx_N = B^{-1}b.$$

This amounts to solving for x_B in terms of x_N , and indicates the dependence of x_B or x_N through the matrix $B^{-1}N$.

We are interested in detecting consistency or correlation of sign patterns between variables in $P(A, b)$ over the set of basic tableaus of the system $P(A, b)$. There are four types of correlation of interest to us, which are defined as follows. From a basis B and respective tableau \bar{A} , form the $n \times n$ *extended tableau* \hat{A} by (1) appending to \bar{A} the $n - m$ rows corresponding to the negative i th unit vectors for $i \in N$ and (2) reordering the rows to correspond to the columns. Let x_i and x_j be two variables in $P(A, b)$, and assume throughout the paper that i and j are distinct. Define the *row correlation* $\rho_B(i, j)$ between x_i and x_j to be the dot product $\hat{A}_i \cdot \hat{A}_j$ between the i th and j th rows of \hat{A} , and define the *column correlation* $\gamma_B(i, j)$ between x_i and x_j to be the negative dot product $-\hat{A}^i \cdot \hat{A}^j$ between the i th and j th columns of \hat{A} . It follows immediately that when x_i is basic, $i = B_q$, and when x_j is nonbasic, then $\rho_B(i, j) = \gamma_B(i, j) = -\bar{a}_{qj}$. We call x_i and x_j *strongly row (column) correlated* or *sign row (column) correlated*, if each term of the dot product $\rho_B(i, j)$ (respectively $\gamma_B(i, j)$) has the same sign or is zero. Finally, x_i and x_j are said to be *weakly row (column) determinate*, or simply *row (column) determinate*, if $\rho_B(i, j)$ (respectively $\gamma_B(i, j)$) has the same sign or is zero over all bases of $P(A, b)$, and *strongly row (column) determinate* if they are weakly row (column) determinate and strongly row (column) correlated over all bases. The system $P(A, b)$ is *totally row (column) determinate*, in either the weak or strong sense, if every pair of variables is row (column) determinate.

Section 2 of this paper interprets the four definitions of correlation and of determinacy in the context of geometry, operations research, linear algebra, statistics and economics. We show the dual nature of row and column correlation and show that in fact all four forms of determinacy are equivalent. In § 3 we restrict ourselves to systems related to transportation, transshipment and network flow problems and give simple characterizations of determinacy in terms of underlying graph structures. In § 4 we look at generalized network systems and apply the results of § 3 to derive strong necessary conditions for determinacy in these systems. This in turn solves, up to sign equivalent matrices, a problem introduced by Greenberg in [11].

2. General results. This section relates the notion of correlation and determinacy to geometric, pivotal, sign-solvable, statistical and economic properties associated with linear systems. We begin with a series of geometric justifications for the notion of correlation. The first set of results concerns strong correlation and its relation to “lines” of a linear system. For a basis B of the linear system $P(A, b)$ and any nonbasic

variable x_k , define the *line* associated with B and k to be

$$L(B, k) = \{x : x \text{ satisfies } P(A, b) \text{ and } x_l = 0 \text{ for } l \in N - \{k\}\}.$$

The line $L(B, k)$ thus describes the linear system when activity is restricted to the variables $x_{B_1}, \dots, x_{B_m}, x_k$. In fact, if b is in general position with respect to A (that is, b is linearly independent of any set of $m - 1$ columns of A), then every one-dimensional set of the type

$$X(S) = \{x : x \text{ satisfies } P(A, b) \text{ and } x_l = 0 \text{ for } l \in S\},$$

where S is a subset of indices, is a line of $P(A, b)$. Further, for any two variables x_i and x_j with x_j not constant on $L(B, k)$, we have a unique change in x_i with respect to x_j along $L(B, k)$, which we denote $\Delta x_i / \Delta x_j |_{L(B, k)}$. Its value can be computed using the following result:

PROPOSITION 2.1. *For any basis B of $P(A, b)$, any nonbasic variable x_k and any two variables x_i and x_j , we have*

$$\frac{\Delta x_i}{\Delta x_j} \Big|_{L(B, k)} = \begin{cases} 1, & i = j = k, \\ -\bar{a}_{rk}, & i = B_r, \quad j = k, \\ \bar{a}_{rk} / \bar{a}_{sk}, & i = B_r, \quad j = B_s, \quad \bar{a}_{sk} \neq 0, \\ \text{undefined} & \text{otherwise,} \end{cases}$$

where \bar{A} is the basic tableau corresponding to B .

Proof. Solving for x_B in terms of x_N , we have

$$x_B = B^{-1}b - \bar{N}x_N,$$

where $\bar{N} = B^{-1}N$ denotes the matrix of columns of \bar{A} associated with x_N . In particular, if $x_l = 0$ for $l \in N - \{k\}$, then x_B can be expressed as a function of x_k by

$$x_B = B^{-1}b - \bar{A}^k x_k$$

so that $\Delta x_i / \Delta x_j$, when defined, results from solving this system. The proposition follows.

As a corollary we obtain a characterization of strong row correlation by describing the relative change of two variables on lines of the linear system.

COROLLARY 2.2. *Two variables x_i and x_j in the system $P(A, b)$ are strongly row correlated with respect to a basis B if and only if $\Delta x_i / \Delta x_j |_L$ has the same sign, is zero or is undefined over every line L of $P(A, b)$ associated with B . The variables x_i and x_j are strongly row determinate if and only if $\Delta x_i / \Delta x_j |_L$ has the same sign, is zero, or is undefined over every line L of $P(A, b)$.*

Proof. Follows directly from Theorem 2.1 and the definition of \hat{A} .

Remark 2.3. There is obvious information contained in basic tableaus of a linear system that concerns the pivoting structure of that system (see, for instance [4, Chap. 7]). In particular, if $P(A, b)$ is *nondegenerate*, that is, distinct bases correspond to distinct basic solutions, then the lines of $P(A, b)$ describe all activity associated with pivoting in $P(A, b)$ and vice versa. Thus the properties of strong correlation and strong determinacy allow us to make powerful general statements concerning sensitivity analysis in $P(A, b)$. In addition, Greenberg [11], [12] and later Greenberg, Lundgren and Maybee [13] have established the importance of this type of correlation as a tool in computer-aided analysis of linear programming models. In the case where nonnegativity constraints are added to $P(A, b)$, the new system becomes the feasible region of a linear program. The lines of $P(A, b)$ associated with bases whose basic solutions are nonnegative—i.e., basic feasible solutions—correspond to standard simplex

algorithm pivots—and therefore a description of activity along these lines describes the activity along the edges of the polyhedron associated with the constrained system. Determinacy in linear systems with nonnegativity constraints is studied in [21].

Remark 2.4. The concepts developed in this paper begin to address some general questions related to “sign-solvability” of linear systems. This topic has elicited considerable research, for example in [10], [15], [17], [18], [19], [20] and [22]. The problem addressed here is: Given a solution x to the system $P(A, 0)$ (this can be thought of as the system of “feasible directions” for $P(A, b)$), when are the signs of certain components of x sufficient to determine the signs of the remaining components? In the context of the definitions and results thus far, we can say that x_i and x_j are strongly row determinate in $P(A, b)$ if and only if the sign of x_j determines uniquely the sign of x_i in every solution that is on some line of $P(A, 0)$ and for which x_i and x_j are nonzero. Further, $P(A, b)$ is totally row determinate in the strong sense if and only if the signs of all nonzero components of any solution that is on some line of $P(A, 0)$ are determined uniquely up to negation of the solution vector.

Column correlation takes a dual role to row correlation. If we embed the system $P(A, b)$ into the linear program

$$\begin{aligned} \min cx, \\ Ax = b, \\ x \geq 0 \end{aligned}$$

for some n -vector c , then the *dual* program becomes

$$\begin{aligned} \max yb, \\ yA + z = c, \\ z \geq 0 \end{aligned}$$

and the variable z_i is called the *marginal* or *reduced cost* associated with $x_i, i = 1, \dots, n$ (see [15, p. 95]). Define $P^*(A, c)$ to be the linear system

$$yA + z = c.$$

In other words, $P^*(A, c) = P(A^*, c)$, where $A^* = (A^T, I)$. Now corresponding to any basis B for $P(A, b)$ there is a dual basis B^* for $P^*(A, c)$ consisting of the variables (y, z_N) , where z_N is the set of reduced cost variables associated with x_N . The equivalent system to $P^*(A, c)$, corresponding to the dual basis B^* as indicated by (1.2), is therefore

$$(1.3) \quad y + z_B B^{-1} = c_B B^{-1}, \quad y - z_B B^{-1} N + z_N + c_N - c_B B^{-1} N,$$

where c_B and z_B are those sets of components of c and z corresponding to x_B , and c_N is that set of components of c corresponding to x_N . From (1.3) we can, for any two reduced cost variables z_i and z_j , define the row correlation $\rho^*_{B^*}(i, j)$ between z_i and z_j with respect to B^* in $P^*(A, c)$. The relationship of this correlation to that in $P(A, b)$ is found in the following theorem.

THEOREM 2.5. *Let B be a basis for $P(A, b)$, with corresponding dual basis B^* for $P^*(A, b)$, and let z_i and z_j be two reduced cost variables in $P^*(A, b)$. Then*

$$\rho^*_{B^*}(i, j) = -\gamma_B(i, j),$$

where x_i and x_j are the variables in $P(A, b)$ corresponding to z_i and z_j .

Proof. From (1.3), the tableau in $P^*(A, c)$ corresponding to B^* is

$$\bar{A}^* = \begin{pmatrix} I_m & (B^{-1})^T & 0 \\ 0 & -(B^{-1}N)^T & I_{n-m} \end{pmatrix},$$

where I_j is the $j \times j$ identity matrix. The extended tableau corresponding to \bar{A}^* is therefore

$$\hat{A}^* = \begin{pmatrix} I_m & (B^{-1})^T & 0 \\ 0 & -(B^{-1}N)^T & I_{n-m} \\ 0 & -I_m & 0 \end{pmatrix} = \begin{pmatrix} I_m & 0 & 0 \\ B^{-1} & 0 & 0 \\ 0 & -\hat{A} & 0 \end{pmatrix}^T$$

where \hat{A} is the extended tableau of $P(A, b)$ corresponding to the original basis B . It follows that

$$\rho^*_{B^*}(i, j) = \hat{A}^*_i \cdot \hat{A}^*_j = \hat{A}^i \cdot \hat{A}^j = -\gamma_B(i, j),$$

and this completes the proof.

The economic significance of studying column correlation in order to determine correlation between reduced cost variables of a linear system now becomes apparent. Theorem 2.5 simply reiterates the economic fact that goods which “substitute” for one another (i.e., are negatively correlated) tend to have positively correlated reduced costs. We can now characterize strong column correlation between two variables in terms of their reduced costs.

COROLLARY 2.6. *Two variables x_i and x_j in $P(A, b)$ are strongly column correlated with respect to a basis B if and only if the associated reduced cost variables z_i and z_j are strongly row correlated in $P^*(A, c)$ with respect to the dual basis B^* . The variables x_i and x_j are column determinate in $P(A, b)$ if and only if the row correlation of z_i and z_j has the same sign, is zero or is undefined over every dual basis of $P^*(A, c)$.*

We can use Corollary 2.6, together with Corollary 1.6, to give a geometric interpretation for column determinacy. For any dual basis B^* and any $k \in B$ we can define the *dual line* associated with B^* and z_k to be

$$L^*(B^*, k) = \{(y, z) : (y, z) \text{ satisfies } P^*(A, c) \text{ and } z_l = 0 \text{ for } l \in B - \{k\}\}.$$

(Note that dual lines do not constitute all lines of $P^*(A, c)$.) Dual to Corollary 1.6 we have

COROLLARY 2.7. *Two variables x_i and x_j in the system $P(A, b)$ are strongly column correlated with respect to a basis B if and only if $\Delta z_i / \Delta z_j|_{L^*}$ has the same sign, is zero, or is undefined over every dual line L^* of $P(A, b)$ associated with the dual basis B^* . The variables x_i and x_j are strongly column determinate if and only if $\Delta z_i / \Delta z_j|_{L^*}$ has the same sign, is zero, or is undefined over every dual line L^* of $P^*(A, c)$.*

The next pair of results links the correlation measures ρ_B and γ_B to statistical and economic concepts. We begin with a lemma concerning statistical correlation between variables related by a system of linear equations. For the proof and a reference to the statistical terms, see [23, § 2.4].

LEMMA 2.8. *Let M be an $m \times n$ matrix, d an m -vector, and let X and Y be n - and m -vectors of variables related by the equation*

$$Y = MX - d.$$

If X comprises independent, identically distributed random variables with common variance σ^2 , then the covariance matrix of Y is $\sigma^2 MM^T$.

Using Lemma 2.8, we can characterize weak row correlation in terms of statistical correlation.

THEOREM 2.9. *Let B be a basis for the system $P(A, b)$, and consider the variables x_B to be dependent on the variables x_N as indicated by (1.2). If the variables in x_N are independent and identically distributed with common variance σ^2 , then for any two variables x_i and x_j , the covariance between x_i and x_j is precisely $\sigma^2 \rho_B(i, j)$.*

Proof. By setting $X = x_N$, $Y = x$, $M = \begin{pmatrix} -B \\ I \end{pmatrix}^{-1N}$ and $d = \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix}$, we have $Y = MX + d$. Applying Lemma 2.8 we have that the covariance matrix for $Y = x$ is

$$MM^T = \hat{A}\hat{A}^T - \begin{pmatrix} I_B & 0 \\ 0 & 0 \end{pmatrix},$$

where I_B is the identity matrix on the rows and columns of B . Thus the off-diagonal elements of MM^T are identical to those of $\hat{A}\hat{A}^T$, and this completes the proof.

Therefore, we can say that variables x_i and x_j are weakly row determinate if and only if the covariance between x_i and x_j has the same sign, or is zero, regardless of which basis is chosen to define the independent and dependent variables for the stochastic structure in Theorem 2.9.

The above result can be applied in a dual sense to justify the definition of weak column correlation. There is a more direct justification, however, which is important for both economic and geometric reasons. It is basically due to Greenberg [11], but we extend the definition slightly in order to apply it in the context of this paper. For basis B of $P(A, b)$ and any two variables x_k and x_l , define the flat associated with B , k and l to be

$$F(B, k, l) = \{x : x \text{ satisfies } P(A, b) \text{ and } x_p = 0 \text{ for } p \in N - \{k, l\}\}.$$

The flat $F(B, k, l)$, then, describes the activity of the linear system when activity is restricted to the variables $x_{B_1}, \dots, x_{B_m}, x_k, x_l$. Again, if b is in general position with respect to A , then all two-dimensional sets of the type $X(S)$ are flats in $P(A, b)$. Flats can have dimension 0, 1 or 2 depending on whether both, one, or neither of x_k and x_l are basic. Now for any variables x_i and x_j , any direction v in $F(B, k, l)$ with $v_j \neq 0$ and any magnitude λ , a displacement of $x^* = (x_B^*, x_N^*)$ to the point $x' = x^* + \lambda v \in F(B, k, l)$ will cause a relation change $\Delta x_i / \Delta x_j|_v$ of x_i with respect to x_j and a relative change $\Delta d / \Delta x_j|_v$ of the distance of x' to x^* with respect to x_j which are dependent only on v . The result is

THEOREM 2.10 (Greenberg). *Let $F(B, i, j)$ be a flat in $P(A, b)$ such that x_j is nonbasic with respect to B . Then that direction v^* in $F(B, i, j)$ which minimizes $\Delta d / \Delta x_j|_v$ has*

$$\frac{\Delta x_i}{\Delta x_j} \Big|_{v^*} = \frac{2}{1 + \|\bar{A}^i\|^2} \gamma_B(i, j),$$

when \bar{A} is the tableau associated with B .

Proof. See [11] for the case when i is also nonbasic. If $i = B_q$ is basic, then $F(B, i, j) = L(B, j)$, and the unique direction v has $v_B = \lambda A^j$, $v_j = \lambda$ and $v_l = 0$ for $l \in N - \{i, j\}$, where λ is any nonzero scalar. Further, A^i is the i th unit vector, and so

from Proposition 2.1

$$\begin{aligned} \frac{2}{1 + \|\bar{A}^i\|^2} \gamma_B(i, j) &= -\frac{2}{1 + 1} \bar{a}_{qj} \\ &= -\Delta x_i / \Delta x_j|_{v=v^*}. \end{aligned}$$

The theorem follows.

The economic significance of Theorem 2.10 is derived from the fact that under reasonable assumptions about personal utility in an economic model, the displacement from a point of economic equilibrium due to a change in any system factors tends to be in the direction of least distance. Thus when change is restricted to a flat $F(B, k, l)$ we establish a sense of economic correlation between variables x_i and x_j by noting that the *marginal rate of substitution* of x_i to x_j on $F(B, k, l)$ is defined to be $\bar{\gamma}_{B,k,l}(i, j) = \Delta x_i / \Delta x_j|_{v^*}$ for v^* as defined in Theorem 2.10 and that this value is uniquely determined whenever x_j is not constant on $F(B, k, l)$ (see [11] for details). As a corollary we have

Corollary 2.11. Two variables x_i and x_j in $P(A, b)$ are column determinate if and only if the marginal rate of substitution of x_i to x_j has the same sign, is zero or is undefined over all flats of $P(A, b)$.

Proof. (\Leftarrow) Follows from Theorem 2.10 and the definition of column correlation.

(\Rightarrow) Suppose $\bar{\gamma}_{B,k,l}(i, j)$ is positive for the flat $F(B, k, l)$. Let \bar{A} be the tableau corresponding to B . We take four cases.

Case 1 ($i, j \in N$). We have $F(B, k, l) = F(B, i, l)$, and so from Theorem 2.10, we have

$$\bar{\gamma}_{B,k,l}(i, j) = \frac{2}{1 + \|\bar{A}^i\|^2} \gamma_B(i, j)$$

so that $\gamma_B(i, j) > 0$.

Case 2 ($i = B_q \in B, j = l \in N, \bar{a}_{qk} \neq 0$). By pivoting on \bar{a}_{qk} we form new basis B' with i and j nonbasic, so that $F(B, k, l) = F(B', i, j)$. This reduces to Case 1.

Case 3 ($i = B_q \in B, j = l \in N, \bar{a}_{qk} = 0$). We have x_i not dependent on x_k in $F(B, k, j)$, so that $\Delta x_i / \Delta x_j|_v = -\bar{a}_{qj}$ for any v with $v_j \neq 0$ and thus $\bar{\gamma}_{B,k,j}(i, j) = -\bar{a}_{qj} = \gamma_B(i, j) > 0$.

Case 4 ($j = B_r \in B$). Since x_j cannot be constant on $F(B, k, l)$, then either \bar{a}_{rk} or \bar{a}_{rl} is nonzero. By pivoting on the appropriate element, we form new basis B' with x_j nonbasic, and this reduces to one of the above three cases.

Thus if $\bar{\gamma}_{B,k,l}(i, j)$ is positive, then there must be some basis B' with $\gamma_{B'}(i, j)$ positive. Similarly $\bar{\gamma}_{B,k,l}(i, j) < 0$ implies $\gamma_{B'}(i, j) < 0$ for some basis B' . Therefore, if the marginal rate of substitution of x_i with respect to x_j has the opposite sign for two flats of $P(A, b)$, then x_i and x_j are not column determinates, and this completes the proof.

It turns out, surprisingly, that the property of determinacy in a linear system is independent of which of the four types of correlation is considered.

THEOREM 2.12. *Let x_i and x_j be variables of the linear system $P(A, b)$. Then*

(1) *x_i and x_j are strongly row correlated with respect to every basis if and only if they are strongly column correlated with respect to every basis;*

(2) *the following are equivalent:*

- (a) *x_i and x_j are weakly row determinate,*
- (b) *x_i and x_j are weakly column determinate,*
- (c) *x_i and x_j are strongly row determinate,*
- (d) *x_i and x_j are strongly column determinate,*
- (e) *\bar{a}_{qj} has the same sign, or is zero, for every basic tableau \bar{A} with $i = B_q$ basic.*

Proof. (1) Suppose that x_i and x_j are not strongly row correlated with respect to basis B . Then x_i and x_j must both be basic, $i = B_q, j = B_r$, and the corresponding tableau

\bar{A} must look like

$$\bar{A} = \begin{bmatrix} & i & j & k & l \\ & \vdots & \vdots & \vdots & \vdots \\ q & \cdots & 1 & \cdots & 0 & \cdots & \bar{a}_{qk} & \cdots & \bar{a}_{ql} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r & \cdots & 0 & \cdots & 1 & \cdots & \bar{a}_{rk} & \cdots & \bar{a}_{rl} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

where \bar{a}_{qk} , \bar{a}_{ql} , \bar{a}_{rk} and \bar{a}_{rl} are nonzero with exactly three having the same sign. But now the matrix

$$C = \begin{pmatrix} \bar{a}_{qk} & \bar{a}_{ql} \\ \bar{a}_{rk} & \bar{a}_{rl} \end{pmatrix}$$

has nonzero determinate, and so by pivoting consecutively on the (q, k) th and (r, l) th entries of \bar{A} , we get new basis B and corresponding tableau \bar{A}^* which looks like

$$\bar{A}^* = \begin{bmatrix} & i & j & k & l \\ & \vdots & \vdots & \vdots & \vdots \\ p & \cdots & \bar{a}_{rl}/\det C & \cdots & -\bar{a}_{ql}/\det C & \cdots & 1 & \cdots & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q & \cdots & -\bar{a}_{rk}/\det C & \cdots & \bar{a}_{qk}/\det C & \cdots & 0 & \cdots & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

which has \bar{a}_{pi}^* , \bar{a}_{pj}^* , \bar{a}_{qi}^* and \bar{a}_{qj}^* nonzero and exactly three with the same sign. Thus x_i and x_j are not strongly column correlated with respect to B . The only if part of (1) is symmetric.

(2) From the definitions we have that any of (2a) through (2d) implies (2e). We proceed to prove (2e) implies (2a) through (2d).

(2e) \Rightarrow (2a). If x_i and x_j are not weakly row determinate, then there must be bases B and B' for which $\rho_B(i, j) > 0$ and $\rho_{B'}(i, j) < 0$. For the basis B , at least one of x_i and x_j is basic. If x_j is basic, $j = B_q$ and x_i is nonbasic, then it must be that $\bar{a}_{qi} < 0$. Thus by pivoting on \bar{a}_{qi} we get new tableau \bar{A}^* with x_i basic, $i = B_q^*$, x_j nonbasic and $\bar{a}_{qj}^* = 1/\bar{a}_{qi} < 0$. If both x_i and x_j are basic, $i = B_q$, $j = B_r$, then there must be at least one column $k \neq i$ with $\bar{a}_{qk} \cdot \bar{a}_{rk} > 0$. By pivoting on $\bar{a}_{rk} \neq 0$ we get new tableau \bar{A}^* with x_i basic, $i = B_q$ and $\bar{a}_{qj}^* = -\bar{a}_{qk}/\bar{a}_{rk} < 0$. In either case $\rho_B(i, j) > 0$ implies that there exists a basic tableau \bar{A} with $i = B_q$ basic and $\bar{a}_{qj} < 0$. An identical argument shows that if $\rho_{B'}(i, j) < 0$ then there exists a basic tableau \bar{A}^* with $i = B_q$ basic and $\bar{a}_{qj}^* > 0$, and thus (2e) does not hold.

(2e) \Rightarrow (2b). Follows by a dual argument to the one above.

(2e) \Rightarrow (2c). Suppose x_i and x_j are not strongly row determinate. In view of the argument above, the only case left to consider is when x_i and x_j are not strongly correlated for some basis B . Then as in the proof of (1), x_i and x_j must both be basic, $i = B_q$ and $j = B_r$, and there must be columns k and l such that column k has $\bar{a}_{qk} \cdot \bar{a}_{rk} > 0$ and column l has $\bar{a}_{ql} \cdot \bar{a}_{rl} < 0$. Now exactly as in the argument above we may pivot respectively on \bar{a}_{rk} and \bar{a}_{rl} to produce distinct basic tableaus \bar{A} and \bar{A}' , with $i = B_q = B'_q$ basic, such that $\bar{a}_{qj} > 0$ and $\bar{a}'_{qj} < 0$. Thus (2e) does not hold.

(2e)⇒(2d). Follows again by as dual argument to the one above. This completes the proof of the theorem.

The final result is of interest in the study of “hidden structure” in linear systems (see [1] and [5], as well as the references in § 3). A matrix A' is *projectively equivalent* [3] to matrix A if there exist $m \times m$ nonsingular matrix B and $n \times n$ nonsingular matrix D such that $A' = BAD$. For $P(A, b)$, the transformation to $P(A', Bb)$ involves essentially a nonzero scaling of the variables in $P(A, b)$ and otherwise no change in the set of solutions to the system. It follows immediately

PROPOSITION 2.13. *Determinacy of variables in a linear system is invariant under projective equivalence of the underlying matrices.*

We shall say more about this in the next section.

3. Network systems. Define a *transshipment matrix* to be a $(0, \pm 1)$ matrix with exactly one $+1$, exactly one -1 , or exactly one $+1$ and one -1 in each column. A *transshipment system* is any system $P(A, b)$ for which A is a transshipment matrix. Transshipment systems occur in numerous network related problems, most notably transportation and network flow problems. Associated with any transshipment matrix A is a directed network $G(A) = (V, E)$ whose node set V corresponds to the rows of A together with an additional *source node* r and whose arc set E corresponds to columns of A , or equivalently, the variables of $P(A, b)$, where, for $k = 1, \dots, n$, the arc associated with x_k is

$$e_k = \begin{cases} (v_i, v_j) & \text{if } a_{ik} = -1 \text{ and } a_{jk} = +1, \\ (r, v_j) & \text{if } a_{jk} = 1, a_{ik} = 0 \text{ for } l \neq j, \\ (v_i, r) & \text{if } a_{ik} = -1 a_{lk} = 0 \text{ for } l \neq i. \end{cases}$$

The first result is based on the fact that A is a *totally unimodular* matrix, that is, all square submatrices of A have determinant 0, $+1$ or -1 . It is stated in terms of totally unimodular matrices.

PROPOSITION 3.1. *Let A be totally unimodular. Then for any basis B of $P(A, b)$, every pair of variables is strongly correlated.*

Proof. It is a well-known fact that all basic tableaus for a totally unimodular matrix are totally unimodular. Now suppose $r_B(i, j)$ does not have all of its terms the same sign. Then it must be the case that i and j are both nonbasic and that \bar{A} must look like

$$\bar{A} = \begin{pmatrix} & \vdots & & \vdots & \\ \cdots & \bar{a}_{pi} & \cdots & \bar{a}_{pj} & \cdots \\ & \vdots & & \vdots & \\ \cdots & \bar{a}_{qi} & \cdots & \bar{a}_{qj} & \cdots \\ & \vdots & & \vdots & \end{pmatrix},$$

where $\bar{a}_{pi}, \bar{a}_{pj}, \bar{a}_{qi}$ and \bar{a}_{qj} are ± 1 and exactly three have the same sign. But this means that the 2×2 submatrix

$$\begin{pmatrix} \bar{a}_{pi} & \bar{a}_{pj} \\ \bar{a}_{qi} & \bar{a}_{qj} \end{pmatrix}$$

has determinant ± 2 , contradicting the fact that A is totally unimodular. The proposition follows.

We next turn to the description of the network structures related to correlation in network systems. A *path* in $G = G(A)$ is an alternating sequence $C = v_0, e_i, v_i, \dots, v_{k-1}, e_k, v_k$ of distinct nodes and arcs for which $e_i = (v_{i-1}, v_i)$ or $(v_i, v_{i-1}), i = 1, \dots, k$. A *circuit* is a path whose endpoints are the same. Arcs e_i and e_j are said to have the *same sense* in a circuit C if they are directed the same way in C and to have the *opposite sense* otherwise. A *spanning tree* in G is a set of arcs which covers all nodes of G and which contain no circuits. An equivalent definition of spanning tree is a set of arcs for which every two nodes are joined by a unique path.

The proof of the first result can be found in [4, Chap. 17].

PROPOSITION 3.2. *If $P(A, b)$ is a transshipment system, then a set B of columns of A forms a basis for $P(A, b)$ if and only if the corresponding arcs in $G(A)$ form a spanning tree.*

Remark 3.3. From Proposition 3.2 it follows that a transshipment matrix A has rank m if and only if $G(A)$ is path connected, that is, every two nodes of $G(A)$ are connected by a path. We henceforth assume this to be the case.

With the aid of Proposition 3.2 we can describe, for any basis B of a transshipment system, the corresponding basic tableau A and hence the value of $\rho_B(i, j)$. Let B be a basis for $P(A, b)$ with T_B the corresponding tree in $G = G(A)$, and let x_k be a nonbasic variable with corresponding arc $e_k \in V - T_B$. It follows that $T_B \cup \{e_k\}$ contains exactly one circuit $C(T_B, e_k)$. If we start with basic solution (x_B^*, x_N^*) corresponding to B and increase x_k by some amount ε , with all other nonbasic variables remaining zero, then the effect on the basic variables is to increase “flow” around the edges of $C(T_B, e_k)$. More precisely, we increase by ε those variables whose edges have the same sense as e_k with respect to $C(T_B, e_k)$ and decrease by ε those variables whose edges have the opposite sense as e_k with respect to $C(T_B, e_k)$, while all other basic variables remain constant (see also [4, Chap. 17]). Thus, for any basic variable x_i with $i = B_q$ we have

$$\bar{a}_{ik} = \frac{\Delta x_i}{\Delta x_k} \Big|_{L(B,k)} = \begin{cases} 1 & \text{if } x_i \text{ and } x_k \text{ have the same sense in } C(T_B, e_k), \\ -1 & \text{if } x_i \text{ and } x_k \text{ have the opposite sense in } C(T_B, e_k), \\ 0 & \text{otherwise.} \end{cases}$$

We have immediately

LEMMA 3.4. *Let B be a basis for $P(A, b)$, and let x_i and x_j be two variables in $P(A, b)$. Then*

- (1) $\rho_B(i, j) > 0$ if e_i and e_j have the same sense for at least one (and hence by Proposition 3.1 for all) circuits $C(T_B, e_k)$ that contain both e_i and e_j ,
- (2) $\rho_B(i, j) < 0$ if x_i and x_j have the opposite sense for at least one (and hence all) circuits $C(T_B, e_k)$ that contain both x_i and x_j ,
- (3) $\rho_B(i, j) = 0$ otherwise.

The next result, due to Duffin [6], plays a critical role in describing determinacy in network models. It concerns the relationship between three properties in networks. First, a pair e_i and e_j are said to be *confluent* if there do not exist circuits C_1 and e_i and e_j in the same sense and C_2 meets e_i and e_j in the opposite sense. Second, a *series-parallel* network is a network which can be obtained from a single arc by performing any sequence of the following three operations

- (1) replace an arc (v_i, v_j) with a pair of identical arcs $(v_i, v_j), (v_i, v_j)$;
- (2) replace an arc (v_i, v_j) with a pair of arcs $(v_i, v_k), (v_k, v_j)$, where v_k is a new node;
- (3) for an existing node v_i add the arc (v_i, v_k) , where v_k is a new node; and then arbitrarily redirecting arcs.

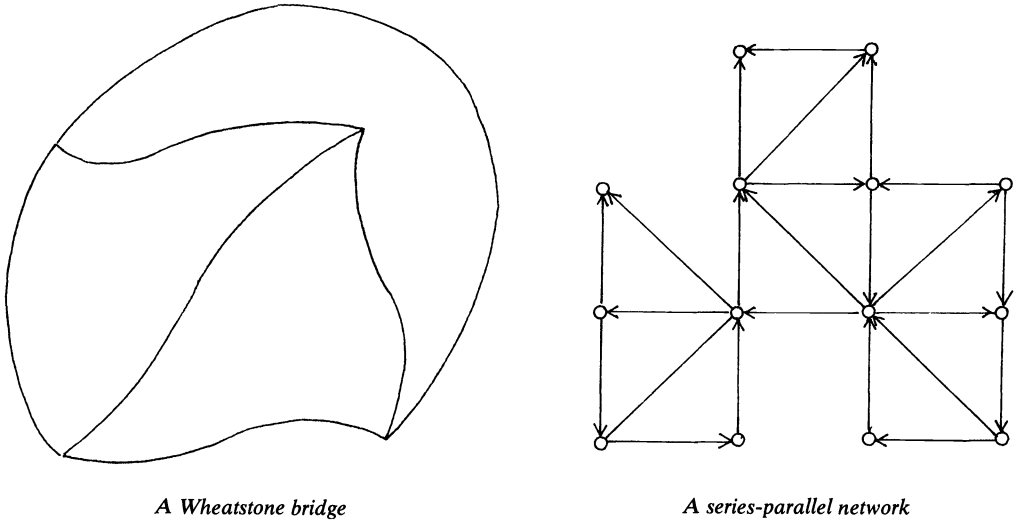


FIG. 1

Finally, a *Wheatstone bridge* is a set of six paths which join every pair of four “corner” nodes and which are otherwise node disjoint. Figure 1 shows a Wheatstone bridge and a series-parallel network. Duffin’s result is now

THEOREM 3.5. (Duffin). *For network G , the following are equivalent:*

- (1) every pair of edges is confluent,
- (2) G is a series parallel network,
- (3) G contains no subset of edges which form a Wheatstone bridge.

In fact, the proof of Theorem 1 in [6] can be modified to show that edges which are not confluent lie on disjoint paths of some Wheatstone bridge. The connection between Theorem 3.5 and determinacy in networks lies in the following lemma.

LEMMA 3.6. *Two variables x_i and x_j of a transshipment system $P(A, b)$ are determinate if and only if e_i and e_j are confluent in $G(A)$.*

Proof. By Lemma 3.4, x_i and x_j are determinate if and only if there do not exist bases B and B' and respectively nonbasic indices k and k' for which e_i and e_j have the same sense in $C(T_B, e_k)$ and the opposite sense in $C(T_{B'}, e_{k'})$. But these circuits correspond to C_1 and C_2 of the definition of confluence and this establishes the necessary part of the lemma. Conversely, given any circuits C_1 and C_2 for which e_i and e_j have the same sense in C_1 and the opposite sense in C_2 , we can remove some edge e_k from C_1 and $e_{k'}$ from C_2 . Now $C_1 - \{e_k\}$ and $C_2 - \{e_{k'}\}$ can be extended to spanning trees T and T' in $G(A)$ which do not contain e_k and $e_{k'}$, respectively. For the corresponding bases B and B' , $\rho_B(i, j) > 0$ and $\rho_{B'}(i, j) < 0$. This establishes the sufficient part of the lemma.

The following characterizations of determinacy and total determinacy in transshipment matrices follow immediately from Theorem 3.3 and Lemma 3.6.

THEOREM 3.7. *Two variables x_i and x_j in a transshipment system $P(A, b)$ are determinate if and only if e_i and e_j are not contained in disjoint paths of some Wheatstone bridge in $G(A)$.*

COROLLARY 3.8. *The transshipment system $P(A, b)$ is totally determinate if and only if $G(A)$ is a series-parallel network.*

We can use the results of Theorem 3.7 and Corollary 3.8 to state a simple characterization of determinacy in a class of transshipment systems related to the classical transportation problem ([4, Chap. 14]) and to the physical flows model described in § 4. A transshipment matrix A is called an *extended transportation matrix* if the rows of A can be partitioned into subsets S_1 and S_2 so that

- (1) If column A^i has a single nonzero entry, then this entry occurs in a row of S_1 if it is $+1$ and in a row of S_2 if it is -1 ;
- (2) If column A^i has two nonzero entries, then the -1 entry of A^i occurs in a row of S_1 and the $+1$ entry of A^i occurs in a row of S_2 ;
- (3) for each row i of A there is column of A whose only nonzero entry is in row i .

$P(A, b)$ is called a *transportation system* if A is an extended transportation matrix. The variables whose columns contain only $+1$ are called *supply* variables, those whose columns contain only -1 are called *demand* variables and those whose columns contain both a $+1$ and a -1 are called *transportation* variables. Statement (3) implies that there is an arc in $G(A)$ from the root r to every node of S_1 and an arc from every node of S_2 to r . The structure of $P(A, b)$ can therefore be described, up to duplicate columns of A , completely by the bipartite network $G'(A)$ whose nodes correspond to rows of A and whose arcs are those of $G(A)$ which are not adjacent to the source node r of $G(A)$. It follows that all arcs of $G'(A)$ are elements of $S_1 \times S_2$. Since multiple copies of edges in $G'(A)$ denote identical and hence obviously determinate corresponding variables, we will assume that $G'(A)$ has no multiple edges.

Define a *lasso* $C \cup \Gamma$ in a network G to consist of a circuit C together with a path Γ whose initial endpoint u is on C and which is otherwise node disjoint from C . Two arcs e_i and e_j , or an arc e_i and node v_j are said to be on *opposite ends* of the lasso $C \cup \Gamma$ if e_i is in C and not adjacent to u , and if e_j is the final arc, or v_j is the final node, of Γ . Figure 2 shows such a lasso. For transportation systems we have the following characterization.

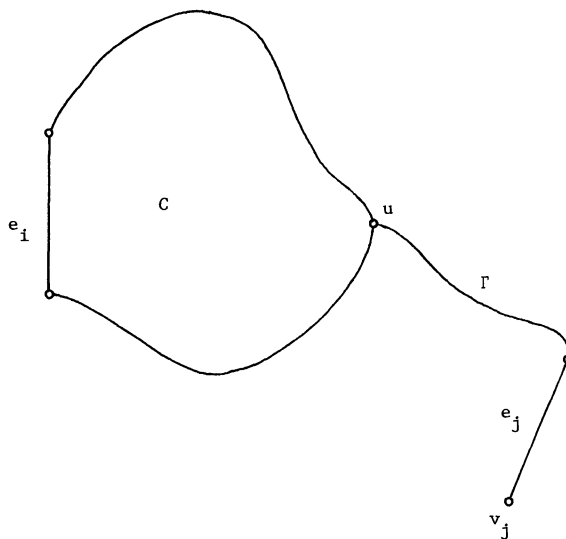


FIG. 2. A lasso.

THEOREM 3.9. *Two variables x_i and x_j in a transportation system $P(A, b)$ are determinate if and only if they do not correspond to two arcs or an arc and a node which are on opposite ends of a lasso in $G'(A)$.*

Proof. (\Rightarrow) Suppose the x_i is a transportation variable with corresponding arc $e_i = (w, z)$ and x_j is either a transportation variable with corresponding arc e_j or a supply/demand variable with corresponding node v_j . Suppose further that e_i and e_j , or e_i and v_j , are on opposite ends of a lasso $C \cup \Gamma$ in $G'(A)$ made up of circuit C containing e_i and path Γ which meets C at node u , with e_j or v_j being at the opposite end of Γ . Let v_j also denote the node of e_j at the end of Γ . Now add to $C \cup \Gamma$ the additional edges $e(v_j)$, $e(w)$ and $e(z)$ of $G(A)$, where

$$e(x) = \begin{cases} (r, x), & x \in S_1, \\ (x, r), & x \in S_2 \end{cases}$$

(see Fig. 3). This forms a Wheatstone bridge in $G(A)$, with corner nodes u, w, z and r and paths $\Gamma \cup \{e(v_j)\}$, e_i , $e(w)$, $e(z)$ and the two paths in $C - \{e_i\}$ from u to v and w respectively. Thus x_i and x_j are not determinate.

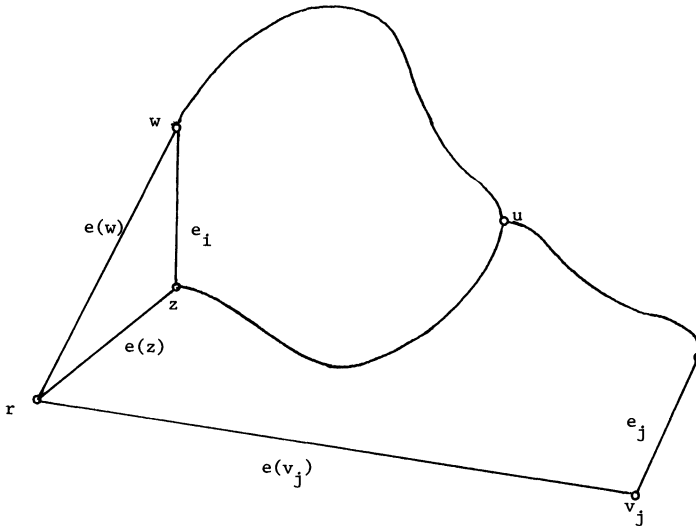


FIG. 3

(\Leftarrow) Suppose that x_i and x_j are not determinate. Then they must be on disjoint paths of some Wheatstone bridge in $G(A)$. In particular one variable, say x_i , must be a transportation variable. Let $\Gamma_1, \dots, \Gamma_6$ be the six paths, and u, v, w, z the corner points, arranged as in Fig. 4. First, suppose that r is one of the corner points and by symmetry we may suppose $r = w$. Then the circuit $\Gamma_1 \cup \Gamma_3 \cup \Gamma_6$, plus the path comprised of the portion of Γ_2 from e_j (or from the endpoint v_j of e_j in $G'(A)$) to z forms a lasso in $G'(A)$ with e_i and e_j or v_j at opposite ends. Second, suppose that r is not one of the four endpoints. Then five of the paths $\Gamma_1, \dots, \Gamma_6$ lie entirely in $G'(A)$, and so there will always be a circuit in $G'(A)$ containing one of e_i and e_j , say e_i , along with a path joining the edge e_j , or node v_j if $e_j = (r, v_j)$ or (v_j, r) , to one of the corner points

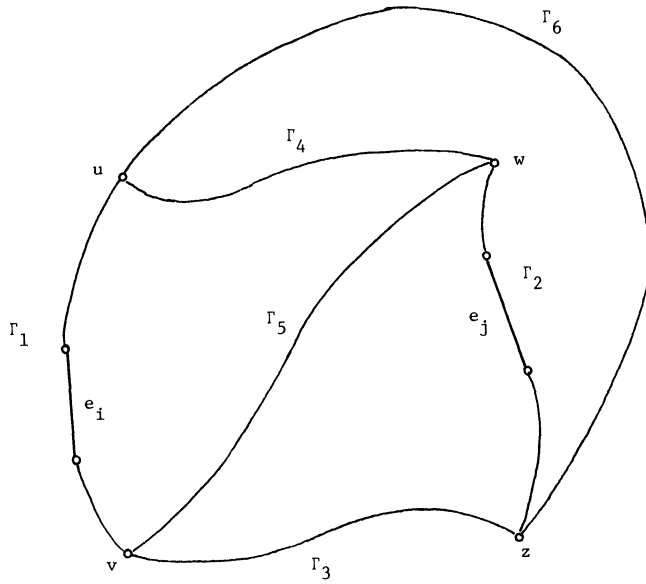


FIG. 4

w or z. This also forms a lasso in $G'(A)$ with e_i and e_j or v_j at opposite ends and the proof is complete.

COROLLARY 3.10. *A transportation system $P(A, b)$ is totally determinate if and only if $G'(A)$ contains no circuits.*

Proof. If $P(A, b)$ contains a circuit then it automatically contains a lasso. Since the circuit must have at least three nodes (we have assumed no duplicate edges), then there must be an edge e_i and a node v_j which is not on that edge. Thus the e_i and v_j are on opposite ends of a lasso, implying that the corresponding variables are not determinate. Conversely, the existence of any lasso in $G'(A)$ implies the existence of a circuit in $G'(A)$. This proves the corollary.

Remark 3.11. We mention here briefly the computational complexity of establishing determinacy in networks. By using Lemma 3.6, along with the algorithm of Shiloach in [24] for finding disjoint paths between two pairs of nodes in a graph, we can produce an $O(mn)$ algorithm for determining whether or not two variables in a transshipment system are determinate. Using Corollary 3.8, along with the algorithm of Valdes, Tarjan and Lawler [25] for recognizing series-parallel graphs, we can produce an $O(n)$ algorithm for determining whether a transshipment system is totally determinate. We do not give the algorithms here. For general linear systems, it is not known whether a polynomial algorithm exists for solving either of these two problems.

We end this section by discussing a final class of systems to which any of the results of this section can be applied. A *hidden network* system [1] is a system $P(A, b)$ for which A is projectively equivalent to a transshipment matrix. There has been considerable recent study of hidden network systems, for instance [2], [8], [14], [16] and Bixby in [1] notes that there are countless examples of problems which can be transformed into network systems in this way. We remark here that by Proposition 2.12 determinacy in any hidden network system can be characterized by looking at

the corresponding transshipment systems and applying the appropriate results from this section. A further generalization of transshipment systems is investigated in § 4, and the appropriate hidden network systems can be evaluated accordingly.

4. Generalized network systems. We now generalize transshipment and transportation systems and apply Theorems 3.7 and 3.9 and their corollaries to derive some powerful necessary conditions for determinacy in these systems. A *generalized network matrix* is a matrix A for which the $(0, \pm 1)$ matrix A^* defined by $a_{ij}^* = \text{sign } a_{ij}$ for $i = 1, \dots, m, j = 1, \dots, n$, is a transshipment matrix. The matrix A is a *physical flows matrix* [11] if A^* is an extended transportation matrix. We associate with A the graph $G(A) = G(A^*)$. The system $P(A, b)$ is referred to as a *generalized network system* or *physical flows system* accordingly.

Generalized network systems have been studied extensively (see, for example, [4, Ch. 21], [7] and [9]) and constitute an important extension of transshipment systems encountered frequently in practical problems. The basis structure in generalized network systems is very similar to that of transshipment systems. In particular, we have

LEMMA 4.1. *Let $P(A, b)$ be a generalized network system, and let B be a set of columns of A for which the corresponding edges in $G(A)$ form a spanning tree. Then B is a basis for $P(A, b)$.*

Proof. It is well known (for instance, in the proof of Theorem 4 in [4, § 17-1]) that if the matrix B corresponds to a spanning tree in $G(A)$, then B can be diagonalized, and since the diagonal elements are nonzero, then B is nonsingular.

We next state a lemma which gives a partial description of the basic tableau A corresponding to a spanning tree basis. For any spanning tree T in $G = G(A)$ and any arc e_k , we note that there exists a unique cycle C in $T \cup \{e_k\}$ and a unique path Γ from some point in C to r , that is, a unique lasso in $T \cup \{e_k\}$ adjacent to r .

LEMMA 4.2. *Let B be a basis of the generalized network system which corresponds to a spanning tree in $G(A)$, and let x_k be a nonbasic variable. Let $C \cup \Gamma$ be the unique lasso in $T \cup \{e_k\}$ adjacent to r , with C and Γ defined:*

$$C: v_{i_0}, e_k, v_{i_1}, \dots, v_{i_s}, e_{j_s}, u, e_{j'_s}, v_{i'_s}, e_{j_{t-1}}, \dots, e_{j'_1}, v_{i'_1} = v_{i_0};$$

$$\Gamma: u = v_{i_{s+1}}, e_{j_{s+1}}, v_{i_{s+2}}, \dots, v_{j_{s+q}}, e_{j_{s+q}}, r.$$

Then the k th column of the basic tableau corresponding to B has

$$a_{jl k} = (-1)^l \frac{a_{i_1 k} a_{i_2 j_1} \dots a_{i_l j_{l-1}}}{a_{i_1 j_1} a_{i_2 j_2} \dots a_{i_l j_l}}, \quad l = 1, \dots, s,$$

$$a_{j k k} = (-1)^l \frac{a_{i_1 k} a_{i_2 j'_1} \dots a_{i_l j'_{l-1}}}{a_{i_1 j'_1} a_{i_2 j'_2} \dots a_{i_l j'_l}}, \quad l = 1, \dots, t.$$

Proof. By Proposition 2.1, we can find the values of the elements of \bar{A}^k by increasing x_k by one unit and observing the corresponding change $\Delta x_i / \Delta x_k$ in the basic variables, that is, by solving the system

$$(4.1) \quad Bx_B + A^k = 0.$$

Now by setting $x_j = 0$ for e_j not on the lasso $C \cup \Gamma$ and then solving successively

equations $i_1, \dots, i_s, i'_1, \dots, i'_t, i_{s+1}, \dots, i_q$ in (4.1) we obtain values

$$\begin{aligned} x_{j_1} &= -\frac{a_{i_1 k}}{a_{i_1 j_1}}, \\ x_{j_2} &= -\frac{a_{i_2 j_1}}{a_{i_2 j_2}} x_{j_1} = \frac{a_{i_1 k} a_{i_2 j_1}}{a_{i_1 j_1} a_{i_2 j_2}}, \\ &\vdots \\ x_{j_s} &= -\left(\frac{a_{i_s j_{s-1}}}{a_{i_s j_s}}\right) x_{j_{s-1}} \\ &= (-1)^s \frac{a_{i_1 k} a_{i_2 j_1} \cdots a_{i_s j_{s-1}}}{a_{i_1 j_1} a_{i_2 j_2} \cdots a_{i_s j_s}}, \\ x_{j'_1} &= -\frac{a_{i'_1 k}}{a_{i'_1 j'_1}}, \\ x_{j'_2} &= \frac{a_{i'_1 k} a_{i'_2 j'_1}}{a_{i'_1 j'_1} a_{i'_2 j'_2}}, \\ &\vdots \\ x_{j'_t} &= (-1)^t \frac{a_{i'_1 k} a_{i'_2 j'_1} \cdots a_{i'_t j'_{t-1}}}{a_{i'_1 j'_1} a_{i'_2 j'_2} \cdots a_{i'_t j'_t}}, \\ x_{j_{s+1}} &= -\frac{a_{i_{s+1} j_s} x_{j_s} + a_{i_{s+1} j'_t} x_{j'_t}}{a_{i_{s+1} j_{s+1}}} \equiv -\gamma, \\ x_{j_{s+2}} &= -\frac{a_{i_{s+2} j_{s+1}}}{a_{i_{s+2} j_{s+2}}} \gamma, \\ &\vdots \\ x_{j_{s+q}} &= (-1)^q \frac{a_{i_{s+2} j_{s+1}} \cdots a_{i_{s+q} j_{s+q-1}}}{a_{i_{s+2} j_{s+2}} \cdots a_{i_{s+q} j_{s+q}}} \gamma. \end{aligned}$$

Since $\bar{a}_{j_1 k} = x_{j_1}$ and $\bar{a}_{j'_1 i'_1} = x_{j'_1}$ the lemma follows.

The crucial thing to note about Lemma 4.2 is that the sign of \bar{a}_{k_j} for any edge e_{B_k} on the circuit C of the lasso is dependent only on the signs of the elements of A . (This is not necessarily true for e_j on the path Γ of the lasso.) Therefore, using Theorem 3.1 and Lemma 4.2, we get that the sign of $\rho_B(i, j)$, for B a basis corresponding to a tree in $G(A)$ and e_i and e_j on some cycle $C(B, k)$, is the same as that of $\rho_B(i, j)$ in the system $P(A^*, b)$. We have immediately

THEOREM 4.3. *Two variables x_i and x_j are determinate in the generalized network system $P(A, b)$ only if they are determinate in the corresponding transshipment system $P(A^*, b)$. $P(A, b)$ is totally determinate only if $P(A^*, b)$ is totally determinate.*

COROLLARY 4.3. *If two variables x_i and x_j are determinate in the generalized network system $P(A, b)$, then e_i and e_j cannot lie on disjoint paths of a Wheatstone bridge in $G(A)$. If $P(A, b)$ is totally determinate, then $G(A)$ is series parallel.*

Finally, for physical flows systems

COROLLARY 4.4. *If two variables x_i and x_j are determinate in the physical flows system $P(A, b)$, then x_i and x_j cannot correspond to two edges or an edge and a vertex*

which lie on opposite ends of a lasso in $G'(A)$. If $P(A, b)$ is totally determinate, then $G'(A)$ is acyclic.

Since the necessary conditions of Corollaries 4.3 and 4.4 are sufficient if the corresponding matrices are in fact transshipment matrices, then we can say that these corollaries are the strongest statements about determinacy in a generalized network system or physical flows system $P(A, b)$ based on *qualitative* (that is solely sign dependent) properties of A . The results in Corollary 4.4, moreover, constitute a significant extension of the physical flows theorem in [11].

REFERENCES

- [1] R. E. BIXBY (1981), *Hidden and embedded structure in linear programs*, Computer-Aided Analysis and Model Simplification, H. Greenberg and J. Maybee, eds., Academic Press, New York, pp. 327–360.
- [2] R. E. BIXBY AND W. H. CUNNINGHAM (1980), *Converting linear programs to network problems*, Math. Oper. Res., 5, pp. 321–357.
- [3] T. H. BRYLAWSKI AND P. LUCAS (1975), *Uniquely representable combinatorial geometrics*, Proceedings International Colloquia on Combinatorial Theory, Rome, Italy.
- [4] G. B. DANTZIG (1963), *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, NJ.
- [5] G. B. DANTZIG AND A. F. VEINOTT (1978), *Discovering hidden totally Leontief substitution systems*, Math. Oper. Res., 3, pp. 102–103.
- [6] R. J. DUFFIN (1965), *Topology of series-parallel networks*, J. Math. Anal. Appl., 10, pp. 303–318.
- [7] J. ELAM, F. GLOVER AND D. KLINGMAN (1979), *A strongly convergent primal simplex algorithm for generalized networks*, Math. Oper. Res., 4, pp. 39–59.
- [8] J. R. EVANS (1976), *A combinatorial equivalence between a class of multicommodity flow problems and the capacitation transportation problems*, Math. Programming, 10, pp. 401–404.
- [9] F. GLOVER, J. HULTZ, D. KLINGMAN AND J. STUTZ (1977), *Generalized networks: a fundamental computer-based planning tool*, Res. Rep. CCS 307, Center for Cybernetic Studies, Univ. Texas at Austin, Austin, TX.
- [10] W. M. GORMAN (1964), *A wider scope for qualitative economics*, Rev. Econ. Studies, 31.
- [11] H. J. GREENBERG (1979), *Measuring complementarity and qualitative determinacy in matricial forms*, in Proceedings of the Symposium on Computer-Assisted Analysis and Model Simplification, H. Greenberg and J. Maybee, eds., Academic Press, New York, pp. 497–522.
- [12] — (1981), *A functional description of ANALYZE: A Computer-Assisted Analysis System for Linear Programming Models*, ACM Trans. on Math. Software, to appear.
- [13] H. J. GREENBERG, J. R. LUNDGREN AND J. S. MAYBEE (1981), *Graph theoretic methods for the qualitative analysis of rectangular matrices*, this Journal, 2, pp. 227–239.
- [14] I. HELLER (1964), *On linear programs equivalent to the transportation problem*, SIAM J. Appl. Math., 12, pp. 31–42.
- [15] V. KLEE, R. LADNER AND R. MANBER (1982), *Sign solvability revisited*, Tech. Rep. 82-04-04, Dept. Computer Science, Univ. of Washington, Seattle.
- [16] D. KLINGMAN (1977), *Equivalent network formulations for constrained networks*, Management Sci., 23, pp. 737–744.
- [17] G. LADY (1982), *The Structure of qualitatively determinate relationships*, Econometrics, to appear.
- [18] K. L. LANCASTER (1965), *The theory of qualitative linear systems*, Econometrics, 33, pp. 395–408.
- [19] J. MAYBEE (1980), *Sign solvability*, Symposium on Computer-Assisted Analysis and Model Simplification, J. Maybee and H. Greenberg, eds., Academic Press, New York.
- [20] J. MAYBEE AND J. QUIRK (1969), *Qualitative problems in matrix theory*, SIAM Rev., 11, pp. 30–51.
- [21] J. S. PROVAN AND A. S. KYDES (1980), *Correlation and determinacy in network models*, BNL Rep. 51243, Brookhaven National Laboratory, Upton, NY.
- [22] P. A. SAMUELSON (1947), *The Foundations of Economic Analysis*, Harvard Univ. Press, Cambridge, MA, pp. 23–28.
- [23] S. R. SEARLE (1971), *Linear Models*, John Wiley, New York.
- [24] Y. SHILOACH (1980), *A polynomial solution to the undirected two paths problem*, J. Assoc. Comput. Mach., 27, pp. 445–456.

EQUIVALENCE CLASSES OF HERMITIAN MATRICES AND THEIR SCHUR PARAMETRIZATION*

PHILIPPE DELSARTE,[†] Y. GENIN[†] AND Y. KAMP[†]

Abstract. A natural equivalence relation on Hermitian matrices is introduced to analyze the concepts of Toeplitz distance and displacement rank. Two Hermitian block-matrices are said to be equivalent when they are congruent under a block-triangular Toeplitz transformation. The matrices having the smallest Toeplitz distance within a given equivalence class are identified. This minimum distance equals the displacement rank minus twice the block size.

Some Σ -unitary transfer functions $S(z)$ and some Schur-like functions $\Phi(z)$ are constructed from the Lyapunov relation defining the displacement rank. These functions are used to characterize the equivalence classes and especially their minimum-distance representatives. A canonical factorization of $S(z)$ corresponds to a Schur-like decomposition of $\Phi(z)$, involving a sequence of generalized Schur parameters E_k . The sets of functions $S(z)$, $\Phi(z)$ and of Schur sequences (E_k) characterizing an equivalence class are described in detail. Some results are obtained concerning the Schur parameters of the inverse of a given matrix.

1. Introduction. The concept of the *displacement rank* of a Hermitian matrix was introduced by Kailath, Kung and Morf [8] to measure how the complexity of the inversion problem depends on the distance of the given matrix to the Toeplitz structure. A closely related concept is the *Toeplitz distance*, used for the same purpose by Friedlander, Morf, Kailath and Ljung [5] in their generalized Levinson algorithm.

Two matrices are called *equivalent* when they are congruent under a triangular Toeplitz transformation. Although equivalent matrices clearly have the same displacement rank, they do not generally have the same Toeplitz distance. Hence the question arises of identifying those representatives of a given equivalence class that have the smallest Toeplitz distance. This question was settled by the authors [2] in the simple case where the displacement rank equals two; it was shown that such "minimal representatives" are nothing but the Toeplitz matrices. The general problem was briefly evoked in [3], in connection with the generalized Levinson algorithm, and is studied in full detail in the present paper. The minimum value of the Toeplitz distance within a given equivalence class is shown to depend only on the displacement rank. In addition, the minimal representatives are explicitly identified.

Each equivalence class can be characterized by two types of functions, denoted by $S(z)$ and $\Phi(z)$, both deduced from the *Lyapunov relation* defining the displacement rank. Functions $S(z)$ of the first type occur as *transfer functions* derived from certain embeddings of the Lyapunov relation [3], [6]. They are Σ -unitary on the unit circle in the general case and Σ -lossless in the positive definite case. As for functions $\Phi(z)$ of the second type, which are defined directly from the Lyapunov relation, they can be viewed as formal generalizations of the classical Schur functions [1], [13]. These *generalized Schur functions* were first introduced by Lev-Ari and Kailath [10]. (See also [3], [7], [9], [11].) It turns out that a canonical factorization of $S(z)$, actually equivalent to the recurrence relation of the generalized Levinson algorithm, follows directly from an extension of the Schur algorithm; see [3], [4], [7], [9], [10], [11]. As a result, the equivalence classes of Hermitian matrices are represented by some sequences of *generalized Schur parameters*. The present paper contains a detailed description of the whole family of functions $S(z)$ and $\Phi(z)$ associated with a given equivalence class. (The reader is referred to a survey by Kailath [7] concerning this subject.) Some remarkable members of the families $\{S(z)\}$ and $\{\Phi(z)\}$ are identified

* Received by the editors February 19, 1982, and in revised form August 9, 1982.

[†] Philips Research Laboratory Brussels, Av. Van Becelaere 2, Box 8, B-1170 Brussels, Belgium.

in correspondence with the minimal representatives of the equivalence classes. The problem of defining “canonical” Schur parameters of minimal matrices is mentioned in the general case and is solved in the positive definite case.

Another interesting question arises in the theory: is there any relation between the functions $S(z)$, $\Phi(z)$ and the Schur parameters associated with a given matrix and with its inverse? A very special case was treated in [2], where the canonical Schur sequences of a positive definite Toeplitz matrix and of its inverse were shown to be reciprocal of each other. A generalization of this result is obtained here for arbitrary positive definite matrices.

It should be mentioned that the present paper deals always with the *block*-displacement rank and the *block*-Toeplitz distance of Hermitian matrices. Finally, it is worth noting that a similar theory can be developed for non-Hermitian complex matrices and, more generally, for square matrices over an arbitrary field.

2. Embeddings and transfer functions. For given integers p and n , with $p \geq 1$ and $n \geq 0$, let F denote the *left p-shift matrix* of order $\alpha = (n + 1)p$, i.e., the matrix $F = [F_{i,j}; 0 \leq i, j \leq n]$ with $p \times p$ blocks $F_{i,j} = 0$ for $j \neq i - 1$ and $F_{i,i-1} = I_p$ for $i = 1, \dots, n$. Throughout this paper we consider a Hermitian matrix $P = [P_{i,j}; 0 \leq i, j \leq n]$, of order α , with $p \times p$ blocks $P_{i,j} = \tilde{P}_{j,i}$. (Here and in the sequel the tilde stands for the conjugate transpose.) Let us define integers β^+ and β^- in terms of P by the following expression:

$$(1) \quad \beta^\pm(P) = \max \{p, r^\pm(P - FP\tilde{F})\},$$

where $r^+(X)$ and $r^-(X)$ stand for the number of positive and negative eigenvalues of the Hermitian matrix X . The sum $\beta = \beta^+ + \beta^-$ will be referred to as the *displacement rank* of P and the terms β^+ and β^- as the *positive* and *negative constituents* of β . Note that these definitions coincide with the classical ones [3], [8] in the “normal case” $r^\pm(P - FP\tilde{F}) \geq p$, which implies $\beta^\pm(P) = r^\pm(P - FP\tilde{F})$ and thus $\beta(P) = \text{rk}(P - FP\tilde{F})$.

Two Hermitian matrices P and \tilde{P} are said to be *equivalent* if there exists a nonsingular matrix L of order α commuting with F such that $\tilde{P} = LP\tilde{L}$. It appears that the commutativity condition $FL = LF$ exactly means that L is a *lower block-Toeplitz matrix*, i.e., $L = [L_{i,j}; 0 \leq i, j \leq n]$ with $L_{i,j} = L_{i-j}$ for $i \geq j$ and $L_{i,j} = 0$ for $i < j$. By definition, $\tilde{P} - FP\tilde{F} = L(P - FP\tilde{F})\tilde{L}$, so that (1) immediately yields $\beta^\pm(\tilde{P}) = \beta^\pm(P)$ in view of Sylvester’s law of inertia. As a conclusion, *the constituents of the p-displacement rank are constant for all matrices in the same equivalence class.*

Let us now examine the equivalence relation just defined from the viewpoint of the Lyapunov equation and its embedding [3], [6]. The starting point is an identity of the form

$$(2) \quad P - FP\tilde{F} = G\Sigma\tilde{G},$$

where $\Sigma = \text{diag}(\pm 1)$ is a signature matrix of order β ($= p$ -displacement rank of P), containing β^\pm diagonal elements ± 1 , while G is a suitable $\alpha \times \beta$ matrix. In the sequel, (2) is called a *Lyapunov relation* for P . In view of $F^{n+1} = 0$ it appears that P is uniquely determined from Σ and G in the form

$$(3) \quad P = \sum_{k=0}^n F^k G \Sigma \tilde{G} \tilde{F}^k.$$

In case P is *nonsingular*, the Lyapunov relation (2) is known to be embeddable into a $(P + \Sigma)$ -unitary relation [6]. This means that there exists a $\beta \times \alpha$ matrix H and

a $\beta \times \beta$ matrix J yielding a $(P \dagger \Sigma)$ -unitary matrix

$$(4) \quad X = \begin{bmatrix} F & G \\ H & J \end{bmatrix}.$$

We recall that X is said to be $(P \dagger \Sigma)$ -unitary when it satisfies $X(P \dagger \Sigma)\tilde{X} = P \dagger \Sigma$. Such a matrix (4) will be referred to as a $(P \dagger \Sigma)$ -unitary embedding of the pair (F, G) . From (4) we construct the $\beta \times \beta$ rational matrix

$$(5) \quad S(z) = J + H(zI_\alpha - F)^{-1}G.$$

Thus $S(z)$ is the *transfer function* admitting X as state space realization. (Note that $S(z)$ is a polynomial of formal degree $n + 1$ in the variable z^{-1} .) Using the fact that X is $(P \dagger \Sigma)$ -unitary, with F nilpotent and P nonsingular, one can easily show that the pairs (F, G) and (F, H) are controllable and observable, respectively, so that X is a *minimal realization* of $S(z)$. Hence $S(z)$ has McMillan degree α . On the other hand, it is easily verified that $S(e^{i\theta})$ is Σ -unitary for all real θ . In addition, $S(z)$ is Σ -lossless in case P is positive definite [3], [6].

To progress further into the question of equivalence one is led to make a weakly restrictive assumption of *nondegeneracy* concerning P , namely

$$(6) \quad \max \{r^+(P - FP\tilde{F}), r^-(P - FP\tilde{F})\} \geq p.$$

(It suffices for example that $P_{0,0}$ be positive or negative definite.) Let now $\tilde{P} = LP\tilde{L}$ be a matrix equivalent to P and consider a Lyapunov relation $\tilde{P} - F\tilde{P}\tilde{F} = \tilde{G}\Sigma\tilde{G}$ for \tilde{P} . Comparing with (2) one obtains

$$(7) \quad \tilde{G}\Sigma\tilde{G} = (LG)\Sigma(\tilde{G}\tilde{L}).$$

Our assumption (6) means that $r^+(G\Sigma\tilde{G}) = r^+(\Sigma)$ or $r^-(G\Sigma\tilde{G}) = r^-(\Sigma)$. In such a situation it turns out that (7) implies the existence of a Σ -unitary matrix U satisfying

$$(8) \quad \tilde{G} = LGU.$$

Let us give an elementary proof of this result (see [12]), in the case $r^+(G\Sigma\tilde{G}) = r^+(\Sigma)$. The starting point is a factorization $G\Sigma\tilde{G} = R_0\Delta\tilde{R}_0$, where R_0 is an $\alpha \times \rho$ matrix of full column rank $\rho = \text{rk}(G\Sigma\tilde{G})$ and Δ is a signature matrix of order ρ such that $r^+(\Delta) = r^+(\Sigma)$. Thus one can write $\Delta = I_\mu \dagger (-I_{\rho-\mu})$ and $\Sigma = I_\mu \dagger (-I_{\beta-\mu})$, without loss of generality. From R_0 construct a nonsingular matrix $R = [R_0, R_1]$ of order α . Defining the $\alpha \times \beta$ matrices $A = R^{-1}G$ and $B = R^{-1}L^{-1}\tilde{G}$ one obtains $A\Sigma\tilde{A} = B\Sigma\tilde{B} = \Delta \dagger 0$ by use of (7). Hence, in view of the structure of Δ and Σ , there exist Σ -unitary matrices U_a and U_b such that the first μ rows of both AU_a and BU_b constitute the matrix $[I_\mu, 0]$. Next, let us interchange the last $\beta - \mu$ columns of AU_a and BU_b ; thus, from the $(\mu, \beta - \mu)$ partitioning $AU_a = [A_0, A_1]$ and $BU_b = [B_0, B_1]$, define the matrices $A' = [A_0, B_1]$ and $B' = [B_0, A_1]$. The property $A\Sigma\tilde{A} = B\Sigma\tilde{B}$ clearly becomes $A'\tilde{A}' = B'\tilde{B}'$, which implies the existence of a unitary matrix Ω of order β satisfying $A'\Omega = B'$. By construction, the first block-row of this identity reduces to $[I_\mu, 0]\Omega = [I_\mu, 0]$, which clearly forces $\Omega = I_\mu \dagger V$ for a suitable unitary matrix V of order $\beta - \mu$. The conclusion is $AU_a(I_\mu \dagger V) = BU_b$; this proves the result (8) for $U = U_a(I_\mu \dagger V)U_b^{-1}$.

Assume again P to be nonsingular. It follows from (8) that the *set* of all transfer functions (5) relative to P remains unchanged when P is replaced by any equivalent matrix $\tilde{P} = LP\tilde{L}$. Indeed, if X is a $(P \dagger \Sigma)$ -unitary embedding of (F, G) , then $\tilde{X} = (L \dagger I_\beta)X(L^{-1} \dagger I_\beta)$ is a $(\tilde{P} \dagger \Sigma)$ -unitary embedding of $(F, \tilde{G}U^{-1})$. Hence, applying (5) to \tilde{X} and \tilde{X} yields $S(z) = \tilde{S}(z)$ as claimed.

Next, we examine the relationship between any two transfer functions $S(z)$ and $\bar{S}(z)$ relative to the same matrix P . Let X and \bar{X} denote the embeddings from which $S(z)$ and $\bar{S}(z)$ originate. By construction, \bar{X} is $(P^{-1} \dagger \Sigma)$ -unitary, so that one can write

$$(9) \quad \begin{bmatrix} \tilde{H} \\ \tilde{j} \end{bmatrix} \Sigma [H \ J] = \begin{bmatrix} P^{-1} & 0 \\ 0 & \Sigma \end{bmatrix} - \begin{bmatrix} \tilde{F} \\ \tilde{G} \end{bmatrix} P^{-1} [F \ G].$$

Applying (8) yields $\bar{G} = GU$ for a certain Σ -unitary matrix U . Hence, substituting \bar{X} for X in (9), one obtains, by direct comparison,

$$(10) \quad \begin{bmatrix} \tilde{H} \\ \tilde{U}^{-1} \tilde{j} \end{bmatrix} \Sigma [\bar{H} \ \bar{J}U^{-1}] = \begin{bmatrix} \tilde{H} \\ \tilde{j} \end{bmatrix} \Sigma [H \ J].$$

As the matrices $[H, J]$ and $[\bar{H}, \bar{J}]$ have full row-rank, (10) implies the existence of a Σ -unitary matrix V such that $[\bar{H}, \bar{J}U^{-1}] = V[H, J]$. Thus the realization \bar{X} has the form $\bar{X} = (I_\alpha \dagger V)X(I_\alpha \dagger U)$, so that the transfer functions $S(z)$ and $\bar{S}(z)$ are simply related by

$$(11) \quad \bar{S}(z) = VS(z)U.$$

We now turn to a converse approach. Let $S(z)$ be a $\beta \times \beta$ rational matrix, of McMillan degree α , admitting the left p -shift matrix F of order α as minimal state transition matrix. Assume $S(e^{i\theta})$ to be Σ -unitary for all real θ , for a given signature matrix Σ . A minimal realization X of $S(z)$ is known to yield a unique Hermitian matrix P of order α such that X is $(P \dagger \Sigma)$ -unitary [3]. (In general, P is not guaranteed to be nonsingular. However, if $S(z)$ is Σ -lossless, then P turns out to be positive definite [6].) On the other hand, any minimal realization \bar{X} of $S(z)$, with $\bar{F} = F$, has the form $\bar{X} = (L \dagger I_\beta)X(L^{-1} \dagger I_\beta)$ for some lower block-Toeplitz matrix L ; hence \bar{X} corresponds to the matrix $\bar{P} = LPL^*$. As a result, $S(z)$ yields a whole class of equivalent matrices. Note that, for U and V varying over the group of Σ -unitary matrices, all functions $\bar{S}(z)$ defined by (11) yield the same equivalence class.

For future use we define a rational $p \times q$ matrix function $\Phi(z)$, with $q = \beta - p$, in the following manner. From the block-rows of the $\alpha \times \beta$ matrix G occurring in the Lyapunov relation (2) construct the $p \times \beta$ matrix polynomial

$$(12) \quad G(z) = [I_p, zI_p, \dots, z^n I_p]G.$$

Consider then the partition $G(z) = [G_0(z), G_1(z)]$, where $G_0(z)$ has p columns and $G_1(z)$ has q columns (with $q = \beta - p$). Assuming $G_0(0)$ to be nonsingular, define the $p \times q$ matrix

$$(13) \quad \Phi(z) = G_0(z)^{-1}G_1(z).$$

Let now $L = [L_{i-j}; 0 \leq i, j \leq n]$ be a nonsingular lower block-Toeplitz matrix (with $L_k = 0$ for $k < 0$). The polynomial $\bar{G}(z)$ obtained by substituting (8) for the matrix G in (12) is given by

$$(14) \quad \bar{G}(z) \equiv L(z)G(z)U \pmod{z^{n+1}},$$

with $L(z) = \sum_{k=0}^n L_k z^k$. It is easily verified by use of (14) that, in case $\bar{G}_0(0)$ is nonsingular, the matrix function $\bar{\Phi}(z) = \bar{G}_0(z)^{-1}\bar{G}_1(z)$ is related to $\Phi(z)$ by

$$(15) \quad \bar{\Phi}(z) \equiv [U_{00} + \Phi(z)U_{10}]^{-1}[U_{01} + \Phi(z)U_{11}] \pmod{z^{n+1}},$$

where $U = [U_{ij}; 0 \leq i, j \leq 1]$ is the (p, q) -partitioned form of U . Thus the functions $\Phi(z)$ corresponding to a given equivalence class of matrices P are deduced from each other by Σ -unitary homographic transformation (15).

It is interesting to see how $\Phi(z)$ can be determined from $S(z)$, in the case where P is nonsingular. To that end we construct the $\beta \times p$ matrix polynomial

$$(16) \quad H(z) = H[z^n I_p, \dots, z I_p, I_p]^T$$

from the block-columns of the $\beta \times \alpha$ matrix H occurring in a $(P \mp \Sigma)$ -unitary embedding (4) of the pair (F, G) . Applying (5) immediately yields

$$(17) \quad z^{n+1} S(z) \equiv H(z) G(z) \pmod{z^{n+1}}.$$

Assume $H(0)$ to have full column-rank and denote by K any $p \times \beta$ matrix such that the product $KH(0)$ is nonsingular. Consider then the column-partition $z^{n+1} K S(z) = [W_0(z), W_1(z)]$, thus defining a $p \times p$ matrix polynomial $W_0(z)$ and a $p \times q$ matrix polynomial $W_1(z)$. From (17) one deduces $W_i(z) \equiv KH(z) G_i(z) \pmod{z^{n+1}}$, so that (13) yields

$$(18) \quad \Phi(z) \equiv W_0(z)^{-1} W_1(z) \pmod{z^{n+1}}.$$

3. Minimal representatives of an equivalence class. We first recall the definition of the Toeplitz distance [3], [5], which is closely related to the displacement rank. Let $P_{\text{inf}} = [P_{i,j}: 1 \leq i, j \leq n]$ and $P_{\text{sup}} = [P_{i,j}: 0 \leq i, j \leq n-1]$ denote the Hermitian matrices of order np obtained by dropping the first and last p rows and columns of P . The *Toeplitz distance* $\gamma(P)$ is defined to be the rank of the matrix $P_{\text{inf}} - P_{\text{sup}}$. The *positive* and *negative constituents* of the Toeplitz distance of P are the integers $\gamma^+(P)$ and $\gamma^-(P)$ defined by

$$(19) \quad \gamma^\pm(P) = r^\pm(P_{\text{inf}} - P_{\text{sup}}).$$

It turns out that the constituents of the displacement rank and Toeplitz distance obey the inequalities

$$(20) \quad \beta^\pm(P) - p \leq \gamma^\pm(P) \leq \beta^\pm(P).$$

To establish (20) let us start from the identity

$$(21) \quad P - FP\tilde{F} = \begin{bmatrix} P_{0,0} & \tilde{T} \\ T & P_{\text{inf}} - P_{\text{sup}} \end{bmatrix},$$

with $\tilde{T} = [P_{0,1}, \dots, P_{0,n}]$. Put $Q = P - FP\tilde{F}$ and $Q_1 = P_{\text{inf}} - P_{\text{sup}}$. By definition, $r^+(Q)$ is the dimension of the largest linear space X of complex α -vectors x such that the real number $\tilde{x}Qx$ is positive for all nonzero x in X . Consider then the linear space X_1 consisting of the $(\alpha - p)$ -vectors x_1 such that $(0, x_1^T)^T$ belongs to X . One clearly has $\dim(X_1) \geq \dim(X) - p$. Hence $r^+(Q_1) \geq r^+(Q) - p$, which yields $\gamma^+(P) \geq \beta^+(P) - p$ in view of definitions (1) and (19). The result $\gamma^-(P) \geq \beta^-(P) - p$ follows from the same argument. The right-hand side inequalities (20) are immediate.

Our next objective is to show that, given a Hermitian matrix P of order α with a nonsingular first block $P_{0,0}$, there exists a matrix \bar{P} equivalent to P such that the Toeplitz distance of \bar{P} achieves the lower bound (20), i.e.,

$$(22) \quad \gamma^\pm(\bar{P}) = \beta^\pm(\bar{P}) - p.$$

To that end, we shall transform an $\alpha \times \beta$ matrix G occurring in the Lyapunov relation (2) into a matrix \bar{G} of the form (8) in such a way that the corresponding matrix $\bar{P} = L\bar{P}\tilde{L}$ satisfies (22). Consider a factorization $P_{0,0} = K\Delta_0\tilde{K}$ with Δ_0 a signature matrix and K

a nonsingular matrix of order p . Since $\beta^\pm \geq p$ one can write, without loss of generality,

$$(23) \quad \Sigma = \Delta_0 \mp (-\Delta_1) \quad \text{with } \Delta_1 = \Delta_0 \mp (-\Gamma),$$

where Γ is a signature matrix of order $\beta - 2p$.

We now prove that there exists a Σ -unitary matrix U and a nonsingular lower block-Toeplitz matrix L such that the product LGU has the form

$$(24) \quad LGU = \begin{bmatrix} \Delta_0 & 0 & 0 \\ T & -T & D \end{bmatrix},$$

where T is an $(\alpha - p) \times p$ matrix and D an $(\alpha - p) \times (\beta - 2p)$ matrix. Let G_0 be the $p \times \beta$ matrix consisting of the first p rows of G . Defining the $p \times \beta$ matrix $R = [K, 0]$ one has $R\Sigma\tilde{R} = K\Delta_0\tilde{K} = P_{0,0} = G_0\Sigma\tilde{G}_0$, via (2). This implies the existence of a Σ -unitary matrix U satisfying $R = G_0U$. It is then easily verified that there exists a unique lower block-Toeplitz matrix $L = [L_{i-j}; 0 \leq i, j \leq n]$ such that LGU has the structure (24). Indeed, writing

$$(25) \quad GU = \begin{bmatrix} K & 0 & 0 \\ A_1 & B_1 & C_1 \\ \vdots & \vdots & \vdots \\ A_n & B_n & C_n \end{bmatrix},$$

where A_k and B_k are $p \times p$ matrices while C_k is a $p \times (\beta - 2p)$ matrix, one successively determines the blocks L_0, L_1, \dots, L_n from the conditions $L_0K = \Delta_0$, $L_1K + L_0(A_1 + B_1) = 0$, $L_2K + L_1(A_1 + B_1) + L_0(A_2 + B_2) = 0$, etc. Hence the desired result is proved.

For the choice of L just explained, define the matrix $\bar{P} = LP\tilde{L}$, equivalent to the given P . Using (2) yields $\bar{P} - F\bar{P}\tilde{F} = \bar{G}\Sigma\tilde{G}$ with $\bar{G} = LGU$. Hence, in view of (23) and (24), one has

$$(26) \quad \bar{P} - F\bar{P}\tilde{F} = \begin{bmatrix} \Delta_0 & \tilde{T} \\ T & D\Gamma\tilde{D} \end{bmatrix}.$$

Thus $\bar{P}_{\text{inf}} - \bar{P}_{\text{sup}} = D\Gamma\tilde{D}$, which implies $\gamma^\pm(\bar{P}) = r^\pm(D\Gamma\tilde{D}) \leq r^\pm(\Gamma) = r^\pm(\Sigma) - p = \beta^\pm(P) - p$. As a consequence, (20) forces the desired equality (22).

A Hermitian matrix P with a normalized nonsingular block $P_{0,0} = \Delta_0 = \text{diag}(\pm 1)$ is said to be *minimal* if its Toeplitz distance achieves the lower bound (20), i.e., if $\gamma^\pm(P) = \beta^\pm(P) - p$. We have just proved that *any Hermitian matrix P with a nonsingular block $P_{0,0}$ is equivalent to a minimal matrix.*

We now consider the problem of describing the whole set of minimal representatives of a given equivalence class obeying the nondegeneracy condition (6). Given a minimal matrix P let Γ be a signature matrix of order $\gamma = \beta - 2p$, containing $\gamma^\pm = \beta^\pm - p$ diagonal elements equal to ± 1 , such that one has

$$(27) \quad P_{\text{inf}} - P_{\text{sup}} = D\Gamma\tilde{D},$$

for some $(\alpha - p) \times (\beta - 2p)$ matrix D . Within permutation, the signature matrix Σ associated with P is then given by (23), where we can choose $\Delta_0 = P_{0,0}$. As it appears from (21) and (27), a solution G to (2) is provided by the right member of (24), i.e.,

$$(28) \quad G = \begin{bmatrix} \Delta_0 & 0 & 0 \\ T & -T & D \end{bmatrix},$$

with $\tilde{T} = [P_{0,1}, \dots, P_{0,n}]$. Let us briefly examine the function $\Phi(z)$ in the present situation. Applying (13) to the special case (28) yields

$$(29) \quad \Phi(z) = [\Delta_0 + T(z)]^{-1}[-T(z), D(z)],$$

with $T(z) = (zI_p, \dots, z^n I_p)T$ and $D(z) = (zI_p, \dots, z^n I_p)D$. As $T(0) = 0$ and $D(0) = 0$, the function $\Phi(z)$ is analytic and vanishes in $z = 0$. Note that G can be reconstructed from $\Phi(z)$. Indeed, (29) implies

$$(30) \quad G(z) = \Delta_0[I_p + \Phi_a(z)]^{-1}[I_p, \Phi_a(z), \Phi_b(z)],$$

where $\Phi = [\Phi_a, \Phi_b]$ is the (p, γ) -partition of the columns of Φ . Furthermore, given a $p \times q$ matrix function $\Phi(z)$ analytic and vanishing in $z = 0$, the polynomial $G(z)$ defined from truncating the Maclaurin expansion of the right member of (30) produces a matrix G of the form (28).

Let us similarly define \bar{G} with the structure (28) for a given minimal matrix $\bar{P} = LP\tilde{L}$ equivalent to P . In this case, the first block-row of (8) reads $[\Delta_0, 0] = L_0[\Delta_0, 0]U$. Hence the block U_{01} must vanish. As U is Σ -unitary, this implies $U_{10} = 0$, so that U has the form

$$(31) \quad U = U_0 \dagger U_1,$$

where $U_i (= U_{ii})$ is a Δ_i -unitary matrix. As a result, the homographic transformation (15) reduces to

$$(32) \quad \bar{\Phi}(z) \equiv U_0^{-1} \Phi(z) U_1 \pmod{z^{n+1}}.$$

On the other hand, the lower block-Toeplitz matrix L is uniquely determined from G and U . In fact, straightforward computation based on (14) and (28) yields

$$(33) \quad L(z)^{-1} \Delta_0 \equiv \Delta_0 U_0 + T(z)(U_0 - U_{1,a}) + D(z)U_{1,b},$$

where $U_{1,a}$ consists of the first p rows and columns of U_1 while $U_{1,b}$ consists of the last γ rows and first p columns of U_1 .

Conversely, given a minimal matrix P one obtains an equivalent minimal matrix $\bar{P} = LP\tilde{L}$ by constructing L from (33) in terms of any Σ -unitary matrix U of the form (31). Indeed, (33) expresses the fact that if G has the structure (28) then so has $\bar{G} = LGU$, which implies that \bar{P} is minimal. Equivalently, $\bar{G}(z)$ can be constructed from U by substituting (32) for $\Phi(z)$ in (30). As a conclusion, the *minimal representatives of a given equivalence class are parametrized by the direct sums of Δ_0 -unitary matrices and Δ_1 -unitary matrices*. The parametrization appears best in the expression (32).

4. Factorization and Schur parameters. In this section we make the classical assumption [3], [5] that the submatrix $P_k = [P_{i,j}: 0 \leq i, j \leq k]$ consisting of the first $(k+1)p$ rows and columns of P is nonsingular for $k = 0, 1, \dots, n$. It is then known that the transfer function (5) associated with P can be factorized in the form

$$(34) \quad S(z) = WR_n(z)R_{n-1}(z) \cdots R_1(z)R_0(z),$$

where $R_k(z)$ is a matrix polynomial of formal degree 1 in z^{-1} and of McMillan degree p , having the property that $R_k(e^{i\theta})$ is Σ -unitary for all θ , while W is a constant Σ -unitary matrix (which is introduced here to simplify further notations). Moreover, the matrices $R_k(z)$ are uniquely determined within constant Σ -unitary left and right factors. Without going into details, let us mention that, for a minimal matrix P , the factorization (34) is closely related to the recurrence relations underlying the *generalized Levinson algorithm* [3], [5]. In fact, it turns out that the function $S_k(z) = R_k(z) \cdots R_0(z)$ is

associated with P_k in the same manner as $S(z)$ with P . The identity $S_k(z) = R_k(z)S_{k-1}(z)$ is then equivalent to the three-term recurrence relations just mentioned.

Let us assume henceforth that the Schur complement of P_{k-1} with respect to P_k is congruent to Δ_0 for $k = 1, \dots, n$. (In case $\Delta_0 = I_p$ this exactly means that P is positive definite.) As explained below, the factorization (34) can then be deduced from a generalization of the Schur algorithm [1], [13] applied to the matrix function $\Phi(z)$ given by (13). The essence of the algorithm in question is the recurrence relation

$$(35) \quad \Phi_{k+1}(z) = z^{-1} \Delta_0 \tilde{D}_{0k}^{-1} [I_p - \Phi_k(z) \Delta_1 \tilde{E}_k]^{-1} [\Phi_k(z) - \Delta_0 E_k] \tilde{D}_{1k} \Delta_1$$

for $k = 0, 1, \dots, n$, with the initialization $\Phi_0(z) = \Phi(z)$. The $p \times q$ matrices E_0, E_1, \dots, E_n occurring in (35) are recursively defined by

$$(36) \quad E_k = \Delta_0 \Phi_k(0), \quad k = 0, 1, \dots, n,$$

so as to make $\Phi_{k+1}(z)$ analytic at the origin. As for D_{0k} and D_{1k} they are nonsingular matrices of order p and q determined from E_k via the relations

$$(37) \quad \tilde{D}_{0k} \Delta_0 D_{0k} = (\Delta_0 - E_k \Delta_1 \tilde{E}_k)^{-1}, \quad \tilde{D}_{1k} \Delta_1 D_{1k} = (\Delta_1 - \tilde{E}_k \Delta_0 E_k)^{-1}.$$

In fact, it turns out that $\Delta_0 - E_k \Delta_1 \tilde{E}_k$ and $\Delta_1 - \tilde{E}_k \Delta_0 E_k$ are congruent to Δ_0 and Δ_1 , respectively, so that (37) actually admits solutions D_{0k} and D_{1k} . Of course, D_{ik} is only determined within a left Δ_i -unitary factor. In the sequel it is understood that D_{ik} is chosen in a well-defined (but arbitrary) manner, so that the matrices E_k are uniquely determined from $\Phi(z)$; they are called the *Schur parameters* of $\Phi(z)$ resulting from the *generalized Schur algorithm* (35)–(37).

The paragraph above contains a description but no validation of the algorithm. In fact, except for the classical case $\Delta_0 = I_p, \Delta_1 = I_q$, it remains an open question to characterize intrinsically the class of functions $\Phi(z)$ to which the algorithm applies. In the context of our study, the validity of the generalized Levinson algorithm is guaranteed from the origin of $\Phi(z)$ on the basis of the following argument. A typical factor $R_k(z)$ in (34) can be written in terms of a $p \times q$ matrix E_k , with $\Delta_0 - E_k \Delta_1 \tilde{E}_k$ congruent to Δ_0 (hence $\Delta_1 - \tilde{E}_k \Delta_0 E_k$ congruent to Δ_1), in the form

$$(38) \quad R_k(z) = \begin{bmatrix} z^{-1} D_{0k} & 0 \\ 0 & D_{1k} \end{bmatrix} \begin{bmatrix} \Delta_0 & E_k \\ \tilde{E}_k & \Delta_1 \end{bmatrix},$$

where D_{0k} and D_{1k} are determined from E_k as in (37). A proof of this property can be found in [3]. (Note that the expression of $R_k(1)$ defined by (38), together with (37), is nothing but the general expression of a Σ -unitary matrix with a nonsingular upper-left $p \times p$ submatrix.) It turns out that the matrices E_k occurring in (38) can be computed from $\Phi(z)$ via the generalized Schur algorithm (35)–(37). This important result is based on the relationship (18) between $\Phi(z)$ and $S(z)$; the proof is essentially the same as in [3] and [4].

To see more precisely how the Schur parameters E_k characterize an equivalence class of matrices P let us examine how they vary when $\Phi(z)$ and $S(z)$ are replaced by $\bar{\Phi}(z)$ and $\bar{S}(z)$. (This question makes sense because our assumptions concerning P are actually valid for all matrices \bar{P} equivalent to P .) In view of (11) and (34), the transfer function $\bar{S}(z) = VS(z)U$ can be factorized as $\bar{S}(z) = \bar{W}\bar{R}_n(z) \cdots \bar{R}_0(z)$ in terms of the constant factor $\bar{W} = VW(\Omega_0 \dagger \Omega_1)$ and of the first degree factors

$$(39) \quad \bar{R}_0(z) = (\Omega_0 \dagger \Omega_1)^{-1} R_0(z) U, \quad \bar{R}_k(z) = (\Omega_0 \dagger \Omega_1)^{-1} R_k(z) (\Omega_0 \dagger \Omega_1),$$

for $k = 1, \dots, n$, where Ω_i is any Δ_i -unitary matrix. Hence (38) yields the following relation on the Schur parameters:

$$(40) \quad \begin{aligned} \bar{E}_0 &= \Delta_0(\Delta_0 U_{00} + E_0 U_{10})^{-1}(\Delta_0 U_{01} + E_0 U_{11}), \\ \bar{E}_k &= \tilde{\Omega}_0 E_k \Omega_1 \quad \text{for } k = 1, \dots, n \end{aligned}$$

with the choice $\bar{D}_{ik} = \Omega_i^{-1} D_{ik} \tilde{\Omega}_i^{-1}$ for $i = 0, 1$ and $k = 1, \dots, n$. In terms of the Schur algorithm, this corresponds to the transformation $\Phi_k(z) \rightarrow \bar{\Phi}_k(z)$ given by (15) for $k = 0$ and by

$$(41) \quad \bar{\Phi}_k(z) \equiv \Omega_0^{-1} \Phi_k(z) \Omega_1 \pmod{z^{n+1-k}} \quad \text{for } k = 1, \dots, n.$$

Let us now look at the case of minimal representatives, for which the functions $\Phi(z)$ and $\bar{\Phi}(z)$ have the form (29). In this situation, one has $E_0 = \bar{E}_0 = 0$ and one can choose $D_{i0} = \Delta_i$ for $i = 0, 1$, which yields $U = \Omega_0 \dagger \Omega_1$. Hence (41) holds true for $k = 0$; see (32).

It should be noted that the results (39)–(41) make sense only if the choice for the solutions D_{ik} to the relations (37) is adapted in a well-defined manner when passing from the parameters E_k to the parameters \bar{E}_k . This naturally raises the question of defining D_{0k} and D_{1k} as functions of E_k in such a way that, for any Δ_i -unitary matrix U_i ($i = 0, 1$), the substitution $E_k \rightarrow \tilde{U}_0 E_k U_1$ induces the substitution $D_{ik} \rightarrow U_i^{-1} D_{ik} \tilde{U}_i^{-1}$ for $i = 0, 1$. In this situation, the E_k 's are referred to as the *canonical Schur parameters* of the minimal matrix P . For arbitrary signature matrices Δ_0 and Δ_1 the condition above can generally not be fulfilled. (In that respect, it does not seem that the question of identifying some ‘‘remarkable solutions’’ D_{ik} to the equations (37) has yet received much attention in the literature.) However, in the important case of positive definite matrices P , corresponding exactly to $\Delta_0 = I_p$, a suitable definition of D_{0k} and D_{1k} is easily discovered, namely

$$(42) \quad D_{0k} = (I_p - E_k \Delta_1 \tilde{E}_k)^{-1/2}, \quad D_{1k} = \Delta_1 (I_q - \tilde{E}_k E_k \Delta_1)^{-1/2},$$

with $A^{-1/2}$ denoting the inverse of the unique positive square root of the positive matrix A . (Here A is said to be positive whenever all its eigenvalues are positive real numbers.) Note that the definition of D_{1k} differs slightly from that given elsewhere [3], [4]. It is worth mentioning that D_{1k} can be computed from E_k and D_{0k} via the rational expression

$$(43) \quad D_{1k} = \Delta_1 + \Delta_1 \tilde{E}_k D_{0k} (I_p + D_{0k})^{-1} D_{0k} E_k \Delta_1.$$

Let us point out that, for the choice (42), the $(I_p \dagger (-\Delta_1))$ -unitary matrix $R_k(1)$ has the property of being $(I_p \dagger \Delta_1)$ -Hermitian. Finally, let us recall that P can be reconstructed from the sequence of its canonical Schur parameters E_k ; this is explained in [3].

5. Schur parameters of the inverse matrix. Let Q be a nonsingular Hermitian matrix of order $\alpha = (n + 1)p$. Our first point is to show how an embedding of type (4) relative to Q^{-1} can be deduced from an embedding of the same type relative to a permuted version of Q itself. From the symmetric permutation matrix

$$(44) \quad \Pi = \begin{bmatrix} & & & & I_p \\ & & & & \\ & & & I_p & \\ & & \dots & & \\ I_p & & & & \end{bmatrix},$$

of order α , define the matrix $P = \Pi Q \Pi$, with blocks $P_{i,j} = Q_{n-i,n-j}$. Given a $(P \dagger \Sigma)$ -unitary matrix X of the form (4), set then $X^* = (\Pi \dagger I_\beta) \tilde{X} (\Pi \dagger I_\beta)$. In view of the obvious identity $F = \Pi \tilde{F} \Pi$, one has

$$(45) \quad X^* = \begin{bmatrix} F & \Pi \tilde{H} \\ \tilde{G} \Pi & \tilde{J} \end{bmatrix}.$$

Since $\tilde{X}(P^{-1} \dagger \Sigma)X = P^{-1} \dagger \Sigma$ can be written as $X^*(Q^{-1} \dagger \Sigma)\tilde{X}^* = Q^{-1} \dagger \Sigma$, it appears that Q^{-1} has the same p -displacement rank as P and that the substitution $G \rightarrow \Pi \tilde{H}$, $H \rightarrow \tilde{G} \Pi$, $J \rightarrow \tilde{J}$ produces an embedding relative to Q^{-1} from an embedding relative to P . Next, let $S(z)$ and $S^*(z)$ denote the transfer functions (5) admitting the realizations X (associated with $P = \Pi Q \Pi$) and X^* (associated with $P^* = Q^{-1}$), respectively. From (45) one deduces

$$(46) \quad S^*(z) = \tilde{J} + \tilde{G} \Pi (z I_\alpha - F)^{-1} \Pi \tilde{H} = \tilde{S}(\bar{z}).$$

We now examine how the Schur parameters relative to Q^{-1} can be deduced from those relative to P . Here we consider only the case where Q is *positive definite*. Thus let $E_0 = 0, E_1, \dots, E_n$ and $E_0^* = 0, E_1^*, \dots, E_n^*$ denote the canonical Schur parameters of any two minimal matrices equivalent to P and to $P^* = Q^{-1}$, respectively. We shall establish the following identity:

$$(47) \quad E_k^* = U_0 E_{n+1-k} U_1,$$

where U_0 is a unitary matrix and U_1 a Δ_1 -unitary matrix (depending on the specific minimal representatives occurring in the definition). The starting point of the argument is the factorization (34), with the normalization $W = I_\beta$. Using $R_0(z) = z^{-1} I_p \dagger I_q$ and applying (46) one obtains

$$(48) \quad S^*(z) = (z^{-1} I_p \dagger I_q) \tilde{R}_k(\bar{z}) \cdots \tilde{R}_n(\bar{z}).$$

On the other hand, as $R_k(1)$ is $(I_p \dagger \Delta_1)$ -Hermitian, the matrix $\tilde{R}_k(1)$ results from replacing the triple (E_k, D_{0k}, D_{1k}) by the triple $(E_k \Delta_1, D_{0k}, \Delta_1 D_{1k} \Delta_1)$ in the expression (38) of $R_k(1)$. Hence (48) can be written as

$$(49) \quad S^*(z) = R_n^*(z) \cdots R_1^*(z) (z^{-1} I_p \dagger I_q),$$

where $R_k^*(z)$ is the matrix (38) built on the parameter $E_k^* = E_{n+1-k} \Delta_1$. This proves the desired result (47).

6. The non-Hermitian case. The whole theory can be adapted to the case of any square matrix P , over an arbitrary field. In this general situation, the equivalence class containing P consists of all matrices of the form $\bar{P} = L P L'$ with L lower block-Toeplitz and L' upper block-Toeplitz, both nonsingular. Roughly speaking, to apply the results above it suffices to give a purely formal meaning to the tilde symbol. (Of course, the notions of positive and negative constituents of the displacement rank and Toeplitz distance become meaningless, so that the signature matrices disappear from the theory.) We shall not go into details about this subject, except for mentioning that the matter is simpler than in the case of Hermitian matrices.

REFERENCES

[1] P. DELSARTE, Y. GENIN AND Y. KAMP, *Schur parametrization of positive definite block-Toeplitz systems*, SIAM J. Appl. Math., 36 (1979), pp. 34-45.

- [2] ———, *On the class of positive definite matrices equivalent to Toeplitz matrices*, in Proc. International Symposium on Mathematical Theory of Networks and Systems, vol. 4, Santa Monica, CA, August 1981, pp. 40–45.
- [3] ———, *A polynomial approach to the generalized Levinson algorithm*, IEEE Trans. Information Theory, to appear.
- [4] ———, *On the Toeplitz embedding of an arbitrary matrix*, Linear Algebra Appl., to appear.
- [5] B. FRIEDLANDER, M. MORF, T. KAILATH AND L. LJUNG, *New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices*, Linear Algebra Appl., 27 (1979), pp. 31–60.
- [6] Y. GENIN, P. VAN DOOREN, T. KAILATH, J. M. DELOSME AND M. MORF, *On Σ -lossless transfer functions and related questions*, Linear Algebra Appl., to appear.
- [7] T. KAILATH, *Time-variant and time-invariant lattice filters for nonstationary processes*, in Outils et modèles mathématiques pour l'automatique, l'analyse de systèmes et le traitement du signal, vol. 2, CNRS, Paris, 1982, pp. 417–464.
- [8] T. KAILATH, S. Y. KUNG AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979) pp. 395–407.
- [9] T. KAILATH AND H. LEV-ARI, *Generalized Schur parametrization of nonstationary second-order processes*, Proc. Otto Toeplitz Memorial Conference, Tel Aviv, May 1981.
- [10] H. LEV-ARI AND T. KAILATH, *Schur and Levinson algorithms for nonstationary processes*, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, March–April 1981, pp. 860–864.
- [11] ———, *On generalized Schur and Levinson–Szegő algorithms for quasistationary processes*, in Proc. 1981 CDC Conference, San Diego, December 1981.
- [12] M. S. LIVSIC AND A. A. YANTSEVITCH, *Operator Colligations in Hilbert Spaces*, R. G. Douglas, ed., V. H. Winston, New York, 1979.
- [13] I. SCHUR, *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232; 148 (1918), pp. 122–145.

THRESHOLD-BOUNDED INTERVAL ORDERS AND A THEORY OF PICYCLES*

PETER C. FISHBURN†

Abstract. Given $1 \leq m \leq n$ with m and n relatively prime if $m \geq 2$, let $\mathcal{P}[m, n]$ be the class of finite partially ordered sets (A, P) whose points a, b, \dots can be mapped into closed intervals with lengths in $[m, n]$ such that, for all $a, b \in A$, aPb if and only if a 's interval lies completely to the right of b 's interval. A theory of picycles based on a mixture of algebraic and combinatorial ideas leads to the conclusion that each $\mathcal{P}[m, n]$ is axiomatizable by a universal sentence of first-order logic. Necessary and sufficient conditions for membership in $\mathcal{P}[m, n]$ are specified.

The present results lie in sharp contrast to an earlier conclusion that the class \mathcal{P}_n of finite interval orders which can be represented using no more than n interval lengths is not axiomatizable by a universal sentence when $n \geq 2$.

1. Introduction. Let \mathcal{P} be the class of nonempty finite interval orders, henceforth referred to simply as *orders*. Order (A, P) consists of a nonempty finite set A and an asymmetric binary relation P on A such that

$$\forall a, b, x, y \in A: \quad (aPx \text{ and } bPy) \Rightarrow (aPy \text{ or } bPx).$$

The aim of this paper is to show that certain interesting subclasses of \mathcal{P} are axiomatizable by a finite number of universal sentences of first-order logic. The subclasses considered here pertain to threshold models for transitive binary relations, like P , whose symmetric complements need not be transitive. The *symmetric complement* I of P is defined thus: aIb if neither aPb nor bPa .

It is well known [1], [5] that for each order (A, P) there exist real valued functions f and ρ on A with ρ strictly positive such that

$$\forall a, b \in A: \quad aPb \Leftrightarrow f(a) > f(b) + \rho(b).$$

Hence aIb if and only if $f(a) + \rho(a) \geq f(b)$ and $f(b) + \rho(b) \geq f(a)$. We refer to such an (f, ρ) as a *representation* of (A, P) , and to ρ as the *threshold function* or *length function* of the representation. The term "length" will also be used for chains and linkages. A *chain* of length K is a sequence of K contiguous P pairs, say $x_1Px_2P \dots Px_KPx_{K+1}$, and a *linkage* of length K is a sequence of K contiguous I pairs, say $y_1Iy_2I \dots Iy_KIy_{K+1}$.

For each positive integer n let \mathcal{P}_n be the class of orders that have representations whose length functions have no more than n values, i.e., $|\rho(A)| \leq n$. For positive integers $m \leq n$ that are relatively prime when $m \geq 2$, let $\mathcal{P}[m, n]$ be the class of orders that have representations whose length functions are bounded between m and n , i.e., $\rho(A) \subseteq [m, n]$. The class of finite semiorders [4], [9] is \mathcal{P}_1 , or equivalently $\mathcal{P}[1, 1]$. Although $\mathcal{P}_1 = \mathcal{P}[1, 1]$, no \mathcal{P}_k for $k \geq 2$ is identical to any $\mathcal{P}[m, n]$.

It is known [3] that no \mathcal{P}_n for $n \geq 2$ is axiomatizable by a universal sentence in first-order logic. However, we shall prove here that each $\mathcal{P}[m, n]$ is axiomatizable by a universal sentence of first-order logic. In other words, there is a finite list of forbidden orders such that order (A, P) is in $\mathcal{P}[m, n]$ if and only if no restriction of (A, P) is isomorphic to one of the forbidden orders. But no such finite list exists for \mathcal{P}_n when $n \geq 2$.

The next section presents the main theorem, which gives conditions on P and I that are necessary and sufficient for membership in $\mathcal{P}[m, n]$. An easy corollary notes

* Received by the editors January 15, 1982, and in revised form August 12, 1982.

† Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974.

that $\mathcal{P}[m, n]$ is axiomatizable by a universal sentence. The third section introduces the notion of picyles (*PI* cycles) and proves a lemma based on linear solvability theory that is used later to complete the proof of the main theorem. The fourth section develops additional theory about picyles that is needed in the main proof.

2. Main theorem. The composition RS of binary relations R and S on a set A is defined by

$$RS = \{(x, y) \in A \times A : \exists z \in A \text{ such that } xRz \text{ and } zSy\}.$$

We also write $xRSy$ for $(x, y) \in RS$. The k -fold composition of R with itself is R^k : $R^1 = R$ and $R^{k+1} = R^kR^1$. Similarly, $P^\alpha I^\beta P^\gamma$ is an $(\alpha + \beta + \gamma)$ -fold composition, with $x(P^\alpha I^\beta P^\gamma)y$ if and only if there are $a, b \in A$ such that $xP^\alpha a$, $aI^\beta b$ and $bP^\gamma y$. A realization of $P^\alpha I^\beta P^\gamma$ consists of a chain of length α adjoined to a linkage of length β , which in turn is adjoined at its other end to a chain of length γ .

Transitivity for P means that $P^2 \subseteq P$, and the defining properties of orders imply that $PIP \subseteq P$ and $IPI \subseteq (P \cup I)$. The conditions we use for membership in $\mathcal{P}[m, n]$ are similar composition-inclusion conditions. Examples for $\mathcal{P}[1, 3]$ and $\mathcal{P}[2, 3]$ illustrate the approach.

The class of orders whose length functions can be bounded between 1 and 3 is $\mathcal{P}[1, 3]$. Suppose $(A, P) \in \mathcal{P}[1, 3]$ and let (f, ρ) be a representation of (A, P) with $\rho(A) \subseteq [1, 3]$. Then $IP^4 \subseteq P$, for if $xIaPbPcPdPy$ then

$$\begin{aligned} f(x) + \rho(x) &\geq f(a), \\ f(a) &> f(b) + \rho(b), \\ f(b) &> f(c) + \rho(c), \\ f(c) &> f(d) + \rho(d), \\ f(d) &> f(y) + \rho(y). \end{aligned}$$

Addition gives $f(x) > f(y) + \rho(y) + [\rho(b) + \rho(c) + \rho(d) - \rho(x)]$, and since the bracketed ρ sum is nonnegative we get $f(x) > f(y) + \rho(y)$, or xPy . In fact, $IP^4 \subseteq P$ is sufficient as well as necessary for an order to be in $\mathcal{P}[1, 3]$.

Similar arguments show that $I^2P^4 \subseteq P$ and $P^4I^2 \subseteq P$ are necessary for membership in $\mathcal{P}[2, 3]$. However, they are not sufficient, since the order shown in Fig. 1 has no 4-chain but is also not in $\mathcal{P}[2, 3]$: if x_1 through x_8 have lengths in $[2, 3]$, then x_9 's length must exceed 3 to intersect both x_1 and x_8 . The order in the figure violates $P^3I^2P^2I \subseteq P$, which is necessary for $\mathcal{P}[2, 3]$. The main theorem shows that we do not have to go beyond double chain-linkage compositions to obtain sufficient conditions for membership in $\mathcal{P}[2, 3]$.

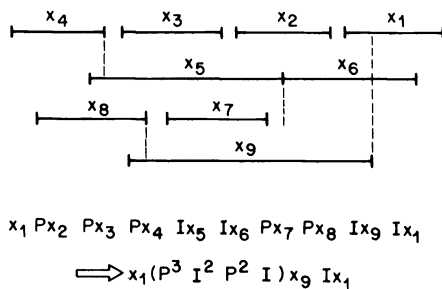


FIG. 1

The general condition for t -fold chain-linkage compositions that we shall use for $\mathcal{P}[m, n]$ is

$A_{t_{mn}}$. For all $(\alpha_1, \beta_1, \dots, \alpha_t, \beta_t) \cong (2, 2, \dots, 2, 1)$ such that $\sum_1^t \alpha_i = n + t$ and $\sum_1^t \beta_i = m + t - 1$:

$$P^{\alpha_1} I^{\beta_1} \dots P^{\alpha_t} I^{\beta_t} \subseteq P,$$

$$I^{\beta_t} P^{\alpha_t} \dots I^{\beta_1} P^{\alpha_1} \subseteq P.$$

Thus $A_{1_{mn}}$ says that $P^{n+1} I^m \subseteq P$ and $I^m P^{n+1} \subseteq P$; $A_{2_{mn}}$ says that $P^\alpha I^\beta P^\gamma I^\delta$ and $I^\delta P^\gamma I^\beta P^\alpha$ are included in P whenever $(\alpha, \beta, \gamma, \delta) \cong (2, 2, 2, 1)$ and $\alpha + \gamma = n + 2$ and $\beta + \delta = m + 1$; and so forth. Our main theorem is

THEOREM 1. Order (A, P) is in $\mathcal{P}[m, n]$ if and only if it satisfies $A_{t_{mn}}$ for $t = 1, \dots, m$.

The proof that $A_{1_{mn}} - A_{m_{mn}}$ are sufficient for $(A, P) \in \mathcal{P}[m, n]$ appears in the next three sections. To demonstrate necessity for $A_{t_{mn}}$, suppose (f, ρ) is a representation of (A, P) with $\rho(A) \subseteq [m, n]$, and let $x_1 P^{\alpha_1} y_1 I^{\beta_1} x_2 P^{\alpha_2} y_2 I^{\beta_2} \dots x_t P^{\alpha_t} y_t I^{\beta_t} x_{t+1}$ be a specific realization of $P^{\alpha_1} I^{\beta_1} \dots P^{\alpha_t} I^{\beta_t}$ with $(\alpha_1, \dots, \beta_t)$ satisfying the hypotheses of $A_{t_{mn}}$. When the (f, ρ) inequalities for the successive pairs in the composition are summed ($f(a) > f(b) + \rho(b)$ for aPb , $f(a) + \rho(a) \geq f(b)$ for aIb) and obvious identical factors on the two sides are cancelled, we get

$$f(x_1) + \left[\text{sum of } \sum_1^t (\beta_i - 1) \text{ terms} + \rho(\cdot) \right]$$

$$> f(x_{t+1}) + \left[\text{sum of } \sum_1^t (\alpha_i - 1) \text{ terms} + \rho(\cdot) \right],$$

or

$$f(x_1) > f(x_{t+1}) + [\text{sum of } n + \rho(\cdot)] - [\text{sum of } (m - 1) + \rho(\cdot)]$$

$$\cong f(x_{t+1}) + nm - (m - 1)n$$

$$= f(x_{t+1}) + n$$

$$\cong f(x_{t+1}) + \rho(x_{t+1}),$$

so that $x_1 P x_{t+1}$. The necessity of $A_{t_{mn}}$'s other conclusion is established in a similar way.

Theorem 1 says that order (A, P) is in $\mathcal{P}[1, n]$ if and only if $(IP^{n+1} \cup P^{n+1}I) \subseteq P$. In fact, our later sufficiency proof shows that only one of the two conclusions of $A_{1_{1n}}$ is needed:

COROLLARY 1. $(A, P) \in \mathcal{P}[1, n]$ if and only if $IP^{n+1} \subseteq P$.

For semiorders we have the known result that order (A, P) is in $\mathcal{P}[1, 1]$ if and only if $IP^2 \subseteq P$.

The main theorem shows exactly what is required for membership in $\mathcal{P}[2, 3]$, namely,

$$(P^4 I^2 \cup I^2 P^4) \subseteq P,$$

$$(P^3 I^2 P^2 I \cup IP^2 I^2 P^3 \cup P^2 I^2 P^3 I \cup IP^3 I^2 P^2) \subseteq P.$$

Because we need only consider the $A_{t_{mn}}$ through $t = m$, and because the hypotheses of $A_{t_{mn}}$ require $\sum \alpha_i = n + t$ and $\sum \beta_i = m + t - 1$ with $(\alpha_1, \dots, \alpha_t, \beta_t) \cong (2, \dots, 2, 1)$, each $\mathcal{P}[m, n]$ is characterized by a finite number of composition inclusions as shown in the statement of $A_{t_{mn}}$. If an order is *not* in $\mathcal{P}[m, n]$ then it must violate one of the

composition inclusions. If it violates a composition inclusion with $K = \sum (\alpha_i + \beta_i)$, then some specific restriction of the order that has $K + 1$ or fewer points must violate that condition. (A restriction of (A, P) is an (A', P') with $\emptyset \subset A' \subseteq A$ and $P' = P \cap (A' \times A')$.) Since there are at most a finite number of nonisomorphic orders with $K + 1$ or fewer points that violate a K -fold composition inclusion, it follows that there is a finite set of orders such that order (A, P) is in $\mathcal{P}[m, n]$ if and only if it has no restriction that is isomorphic to one of the orders in the finite “forbidden” set. Since each composition inclusion is a universal sentence (no existential quantifiers in prenex normal form), it follows from standard definitions [9], [10] for axiomatizations in first-order logic that Theorem 1 implies

COROLLARY 2. *Each $\mathcal{P}[m, n]$ is axiomatizable by a universal sentence of first-order logic.*

As noted earlier, a similar conclusion does not hold for \mathcal{P}_n when $n \geq 2$. In particular, there are orders of arbitrarily large finite cardinality that are not in \mathcal{P}_n but have every proper restriction in \mathcal{P}_n .

3. Forbidden picycles. A constructive sufficiency proof for Theorem 1 would develop a representation with $\rho(A) \subseteq [m, n]$ for any order (A, P) that satisfies $A1_{mn}$ through Am_{mn} . This has been done for semiorders ($m = n = 1$) by Scott and Suppes [9] and others [6], [7], but appears unwieldy for general $[m, n]$. I shall therefore establish sufficiency for Theorem 1 using an indirect approach that is suggested by Scott’s semiorder proof [8].

This approach has two main steps. The first, taken in this section, applies solvability theory for a system of linear inequalities to the inequality system of an (f, ρ) representation to obtain a necessary and sufficient condition on P and I for membership in $\mathcal{P}[m, n]$. The second step, taken in the next two sections, shows that the condition on P and I holds for an order when the order satisfies $A1_{mn} - Am_{mn}$.

Several definitions will be needed. A cycle for (A, P) is any list

$$\mathcal{C} = x_1 R_1 x_2 R_2 \cdots x_K R_K x_1 \quad (K \geq 1)$$

with $x_k \in A$ and $R_k \in \{P, I\}$ for $k = 1, \dots, K$. The cycle is *pure* if all $K x_k$ are different. Cycles $x_1 R_1 x_2 R_2 \cdots x_K R_K x_1$, $x_2 R_2 \cdots x_K R_K x_1 R_1 x_2, \dots$, and $x_K R_K x_1 R_1 \cdots x_{K-1} R_{K-1} x_K$ are viewed as equivalent.

A cycle \mathcal{C} with $R_k = I$ for all k is an *icycle*. Thus, an icycle consists of one linkage that begins and ends at the same point. A cycle is a *picycle* if $R_k = P$ for at least one k . It must have at least one I also since otherwise transitivity of P would give $x_1 P x_1$.

A picycle can always be arranged so that $R_1 = P$ and $R_K = I$. When thus arranged, it consists of a chain, followed by a linkage, followed by a chain, \dots , and ending with a linkage. A picycle \mathcal{C} that is composed of t chains of lengths $\alpha_1, \dots, \alpha_t$, followed alternately with t linkages of lengths β_1, \dots, β_t , will be abbreviated as

$$\mathcal{C} = x_1 P^{\alpha_1} y_1 I^{\beta_1} \cdots x_t P^{\alpha_t} y_t I^{\beta_t} x_1$$

or even more cryptically as $\mathcal{C} = P^{\alpha_1} I^{\beta_1} \cdots P^{\alpha_t} I^{\beta_t}$. The *index* of such a \mathcal{C} is $(\alpha_1, \beta_1, \dots, \alpha_t, \beta_t)$, its *length* is $\sum_1^t (\alpha_i + \beta_i)$, and its *P-excess* e_P and *I-excess* e_I are defined by

$$e_P = \sum (\alpha_i - 1) = \sum \alpha_i - t,$$

$$e_I = \sum (\beta_i - 1) = \sum \beta_i - t.$$

We now connect these ideas with $[m, n]$. For a reason that is made clear in the following lemma, a picycle $\mathcal{C} = P^{\alpha_1} I^{\beta_1} \cdots P^{\alpha_t} I^{\beta_t}$ for which $m e_P \geq n e_I$ will be said to be (m, n) -forbidden.

LEMMA 1. *An order (A, P) is in $\mathcal{P}[m, n]$ if and only if it has no (m, n) -forbidden picycle.*

It is known that there is a 19-point order for $[m, n] = [5, 9]$ that is not in $\mathcal{P}[5, 9]$ and which necessarily has a $(5, 9)$ -forbidden picycle but has no *pure* picycle that is $(5, 9)$ -forbidden. Hence Lemma 1 is not generally true when “pure picycle” is substituted for “picycle”.

The remainder of this section proves Lemma 1. Since it is straightforward to show that (A, P) cannot be in $\mathcal{P}[m, n]$ when it has a picycle with $me_P \cong ne_t$, we turn immediately to the proof that an order not in $\mathcal{P}[m, n]$ must have an (m, n) -forbidden picycle.

To set the stage, suppose for the moment that (A, P) is in $\mathcal{P}[m, n]$ with $|A| = N$ and $A = \{a_1, a_2, \dots, a_N\}$. Then, given any $\tau > 0$, there are f and ρ such that

$$\begin{aligned} f(a_i) &> f(a_j) + \rho(a_j) \quad \text{whenever } a_iPa_j, \\ f(a_i) + \rho(a_i) &\cong f(a_j) \quad \text{whenever } a_iIa_j \text{ and } i \neq j, \end{aligned}$$

and

$$m\tau \leq \rho(a_i) \leq n\tau \quad \text{for } i = 1, \dots, N.$$

Define

$$\omega = (f(a_1), \dots, f(a_N), \rho(a_1), \dots, \rho(a_N), \tau);$$

$a(i, j)$ is a $(2N + 1)$ -vector with 1 in position i , -1 's in positions j and $N + j$, and 0's elsewhere, given a_iPa_j ;

$a(i, j)$ is a $(2N + 1)$ -vector with 1's in positions i and $N + i$, -1 in position j , and 0's elsewhere, given a_iIa_j .

Also let k_j denote k in position j . It then follows that (A, P) is in $\mathcal{P}[m, n]$ if and only if there exists an ω that satisfies

$$\begin{aligned} \omega \cdot (0, \dots, 0, 1_{2N+1}) &> 0, \\ \omega \cdot a(i, j) &> 0 && \text{whenever } a_iPa_j, \\ (*) \quad \omega \cdot a(i, j) &\cong 0 && \text{whenever } a_iIa_j \quad (i \neq j), \\ \omega \cdot (0, \dots, 0, 1_{N+i}, 0, \dots, 0, -m) &\cong 0 && (i = 1, \dots, N), \\ \omega \cdot (0, \dots, 0, -1_{N+i}, 0, \dots, 0, n) &\cong 0 && (i = 1, \dots, N). \end{aligned}$$

Suppose this system has K inequalities, indexed sequentially from 1 to K with a^k the $(2N + 1)$ -vector that ω multiplies in inequality k , and let M be the number of ordered pairs in P . Then the system can be written as

$$\begin{aligned} \omega \cdot a^k &> 0 \quad \text{for } k = 1, \dots, M + 1, \\ \omega \cdot a^k &\cong 0 \quad \text{for } k = M + 2, \dots, K. \end{aligned}$$

Assume henceforth that $(A, P) \notin \mathcal{P}[m, n]$, so that $(*)$ has no ω solution. A standard result in linear solvability theory, e.g., [2, p. 46], says that there are nonnegative integers r_1, \dots, r_K with $r_k > 0$ for some $k \leq M + 1$ such that

$$(1) \quad \sum_{k=1}^K r_k a_i^k = 0 \quad \text{for } i = 1, \dots, 2N + 1,$$

where $a^k = (a_1^k, \dots, a_{2N+1}^k)$. Assume (1) henceforth with $r_k > 0$ for some $k \leq M + 1$.

Let $r = r_1 + r_2 + \dots + r_K$. Replicates of the basic inequalities for those k that have $r_k \geq 2$ along with single instances of the inequalities that correspond to $r_k = 1$ yield the following correspondent of (1):

$$\begin{array}{ll}
 c_1 \text{ equalities} & \tau > 0, \\
 c_2 \text{ inequalities} & +f(a_i) - f(a_j) - \rho(a_j) > 0 \quad (a_i P a_j), \\
 c_3 \text{ inequalities} & +f(a_i) - f(a_j) + \rho(a_i) \geq 0 \quad (a_i I a_j), \\
 c_4 \text{ inequalities} & +\rho(a_i) - m\tau \geq 0, \\
 c_5 \text{ inequalities} & -\rho(a_i) + n\tau \geq 0,
 \end{array}$$

with $c_1 = r_1, c_2 = r_2 + \dots + r_{M+1}, c_1 + c_2 > 0$, and $\sum c_i = r$. According to (1), for each $i \leq N$,

B1. The number of inequalities in the c list with $+f(a_i)$ equals the number with $-f(a_i)$.

B2. The number of inequalities with $+\rho(a_i)$ equals the number with $-\rho(a_i)$.

B3. $c_1 - mc_4 + nc_5 = 0$ [for $i = 2N + 1$ in (1)].

Summation of the r inequalities leaves $0 > 0$ since $c_1 + c_2 > 0$, thus reflecting the lack of an ω solution for (*).

We now work with the P and I pairs for the c_2 and c_3 inequalities, using balance conditions B1–B3 as needed. Suppose first that $c_2 = 0$. Then all ρ terms for c_2 and c_3 are $+\rho(a_i)$, and, by B2, these must be balanced by $-\rho(a_i)$ terms from c_5 . If $c_5 > c_3$, the $c_5 - c_3$ excess terms from c_5 must be balanced by terms from c_4 . Hence, by B2, $c_4 = c_5 - c_3$. By B3, $c_1 = mc_4 - nc_5 = m(c_5 - c_3) - nc_5 = -(n - m)c_5 - mc_3 \leq 0$, which contradicts $c_1 + c_2 > 0$. Therefore $c_2 > 0$.

Given $c_2 > 0$, we arrange the $c_2 x_i P y_i$ pairs and the $c_3 z_i I w_i$ pairs in two rows:

	$c_2 P$ pairs	$c_3 I$ pairs	
row 1:	$x_1 x_2 \dots x_{c_2}$	$z_1 z_2 \dots z_{c_3}$	$+f$
row 2:	$y_1 y_2 \dots y_{c_2}$	$w_1 w_2 \dots w_{c_3}$	$-f$

Condition B1 implies that row 2 is a reordering of row 1. We now form picycles from this array, treating each column as distinct.

We begin the first picycle with $x_1 P y_1$. This is followed by a $y_1 I w_i$ if y_1 is one of the z_i 's, and by $y_1 P y_j$ if $y_1 \notin \{z_1, \dots, z_{c_3}\}$. In constructing this or any later picycle, we shall always follow a P pair by an I pair whenever possible, and always follow an I pair by a P pair whenever possible, except when the current pair completes a picycle—as described shortly. Once a pair (column) is used, it is deleted. The construction of the first picycle continues until we encounter x_1 as the second member of a newly added pair and there are no unused x_i 's in the first row. For example, if x_1 appears three times in each row, then it will initiate and terminate the first picycle and appear twice in the interior of the picycle. As a final step, we rearrange the picycle so that it begins with a P and ends with an I .

With all columns that are used in the first picycle deleted from the array, we construct a second picycle in a similar manner if any P pairs remain. This continues until the P pairs are exhausted. At this point, any remaining I pairs can be formed into cycles since the remainder of row 2 is a reordering of the remainder of row 1.

The construction of picycles follows a P pair by an I pair, and conversely, whenever possible. We shall say that a *transition* occurs each time P is followed by I . Each transition has the form $a P b I c$, with $-\rho(b)$ associated with $a P b$, and $+\rho(b)$ associated with $b I c$. Hence the ρ terms for the transitions balance out as part of B2.

We claim that the y_i in nontransitional P columns are disjoint from the z_j in nontransitional I columns. Suppose to the contrary that

$$\begin{aligned} x_i P y_i P c & \text{ is part of a picycle,} \\ a I y_i I w_j & \text{ is part of a picycle or icycle.} \end{aligned}$$

This contradicts the method of construction, for if $x_i P y_i$ arises in the construction process before $a I y_i$ has been used, then $x_i P y_i$ would be followed by $y_i I w_j$ (or some $y_i I w_k$), and if $a I y_i$ arises before $x_i P y_i$ has been used, then $a I y_i$ would be followed by $y_i P c$ (or some $y_i P y_k$).

Let T be the number of transitions in the constructed picycles. By the preceding paragraph, the $(c_2 - T)$ y_i in nontransitional P columns have $-\rho(y_i)$ terms whose y_i are disjoint from the $(c_3 - T)$ z_j in nontransitional I columns, which are associated with $+\rho(z_j)$ terms. Therefore, by B2, c_4 must include $c_2 - T$ terms $(+\rho)$ to balance the $-\rho(y_i)$ ones, c_5 must include $c_3 - T$ terms $(-\rho)$ to balance the $+\rho(z_j)$ ones. In addition, c_4 and c_5 can each have S other terms that cancel between the two. We then have

$$c_4 = c_2 - T + S, \quad c_5 = c_3 - T + S.$$

Since B3 says that $c_1 = m c_4 - n c_5 \geq 0$, it follows that $m(c_2 - T) \geq n(c_3 - T) + (n - m)S$, hence

$$m(c_2 - T) \geq n(c_3 - T).$$

Let $E_P(E_I)$ be the sum of the excesses $e_P(e_I)$ in the constructed picycles. Then, since each chain in a picycle is associated with one transition, $E_P = c_2 - T$. Similarly, since each linkage in a picycle is associated with one transition, and since icycles could arise, $c_3 - T \geq E_I$. Therefore

$$m E_P \geq n E_I,$$

and hence $m e_P \geq n e_I$ for some picycle. Thus, an order not in $\mathcal{P}[m, n]$ must have an (m, n) -forbidden picycle.

4. Reducible picycles. This section develops a theory of reducibility for forbidden picycles which is then used in the next section along with Lemma 1 to complete the sufficiency proof of Theorem 1. For convenience, the mn designation, as in (m, n) -forbidden and At_{mn} , is often omitted.

We shall say that a forbidden picycle $\mathcal{C} = P^{\alpha_1} I^{\beta_1} \dots P^{\alpha_T} I^{\beta_T}$ is *reducible* if some contiguous segment of the picycle involving two or more P and/or I pairs can be replaced by one P or I pair formed with the first and last elements in the segment so that the picycle \mathcal{C}' obtained from \mathcal{C} by the replacement is also forbidden. For example, when $[m, n] = [2, 3]$, the forbidden picycle

$$\mathcal{C} = x_1 P x_2 P x_3 P x_4 P x_5 P x_6 I x_7 I x_8 I x_1$$

is reducible since $\mathcal{C}' = x_1 P x_3 P x_4 P x_5 P x_6 I x_7 I x_8 I x_1$ has $m e'_P = 2(3) \geq 3(2) = n e'_I$. The segment $x_1 P x_2 P x_3$ is replaced by $x_1 P x_3$ to get \mathcal{C}' from \mathcal{C} .

Here, and later, we shall let $h(\mathcal{C})$ denote the length of picycle \mathcal{C} . If \mathcal{C} is a reducible forbidden picycle and \mathcal{C}' is a reduction of \mathcal{C} obtained by replacing a segment of \mathcal{C} with a single P or I pair as described in the preceding paragraph, then $h(\mathcal{C}') < h(\mathcal{C})$.

Our basic lemma for reducibility is

LEMMA 2. *Suppose $\mathcal{C} = P^{\alpha_1} I^{\beta_1} \dots P^{\alpha_T} I^{\beta_T}$ is a forbidden picycle. Let $\alpha_i = \alpha_{i-T}$ and $\beta_i = \beta_{i-T}$ when $T + 1 \leq i \leq 2T$. \mathcal{C} is reducible if $\alpha_i = 1$ or $\beta_i = 1$ for some $i \in \{1, \dots, T\}$. If $\alpha_i \geq 2$ and $\beta_i \geq 2$ for all i , and if At holds for a specified $t \leq T$, then \mathcal{C} is reducible to*

\mathcal{C}' with $h(\mathcal{C}') = h(\mathcal{C}) - [m + n + 2(t - 1)]$ if there is an $i \in \{1, \dots, T\}$ such that either

- (a) $\alpha_i + \dots + \alpha_{i+t-1} \geq n + t, \alpha_{i+1} + \dots + \alpha_{i+t-1} \leq n + t - 2$ if $t \geq 2, \beta_i + \dots + \beta_{i+t-1} \geq m + t$, and $\beta_i + \dots + \beta_{i+t-2} \leq m + t - 2$ if $t \geq 2$; or
- (b) $\beta_i + \dots + \beta_{i+t-1} \geq m + t, \beta_{i+1} + \dots + \beta_{i+t-1} \leq m + t - 2$ if $t \geq 2, \alpha_{i+1} + \dots + \alpha_{i+t} \geq n + t$, and $\alpha_{i+1} + \dots + \alpha_{i+t-1} \leq n + t - 2$ if $t \geq 2$.

Proof. Given the hypotheses of Lemma 2, suppose first that $\beta_1 = 1$. If $T = 1$, then $\alpha_1 = 1$ is impossible since not $(xPyIx)$, and $\alpha_1 \geq 2$ is impossible since $PIP \subseteq P$ and $P^2 \subseteq P$ would give xPx . Hence $T \geq 2$. But then, since $PIP \subseteq P, \mathcal{C}$ can be reduced to $\mathcal{C}' = P^{\alpha_1 + \alpha_2 - 1} I^{\beta_2} \dots$, which is forbidden since it has $e'_P = e_P$ and $e'_I = e_I$. It follows that \mathcal{C} is reducible if any $\beta_i = 1$.

Assume henceforth that $\beta_i \geq 2$ for all i . Suppose next that some $\alpha_i = 1$. Then $me_P \geq ne_I$ requires $T \geq 2$, so assume for definiteness that $\alpha_2 = 1$. Since $IPI \subseteq (I \cup P)$, the IPI part of $I^{\beta_1} P I^{\beta_2}$ can be replaced by I or P to yield \mathcal{C}' . Since it is easily checked that either replacement gives $me'_P \geq ne'_I, \mathcal{C}$ is reducible.

Assume henceforth that $\alpha_i \geq 2$ for all i . Suppose At holds for some $t \leq T$ and that the α_j and β_j satisfy the inequalities of statement (a) for some $i \leq T$. For notational convenience let $i = 1$. If $t = 1$, then (a) says that $\alpha_1 \geq n + 1$ and $\beta_1 \geq m + 1$, so that the $P^{n+1} I^m$ part of $P^{\alpha_1} I^{\beta_1}$ can be replaced by P according to A1. This changes e_P to $e'_P = e_P - n$ and e_I to $e'_I = e_I - m$, so $me'_P \geq ne'_I$ and \mathcal{C} is reducible with $h(\mathcal{C}') = h(\mathcal{C}) - [m + n]$. If $t \geq 2$ then (a) says that

$$\alpha_1 \geq n + t - \sum_2^t \alpha_i \geq 2 \quad \text{and} \quad \beta_t - 1 \geq m + t - 1 - \sum_1^{t-1} \beta_i \geq 1,$$

so the segment

$$P^{n+t-\sum_2^t \alpha_i} I^{\beta_1} P^{\alpha_2} \dots P^{\alpha_i} I^{m+t-1-\sum_1^{t-1} \beta_i}$$

of length $m + n + 2t - 1$ can be replaced by P according to At to yield picycle \mathcal{C}' with $h(\mathcal{C}') = h(\mathcal{C}) - [m + n + 2(t - 1)]$. Since an I pair immediately follows the replaced segment,

$$e'_P = e_P - \left[\left(n + t - \sum_2^t \alpha_i - 1 \right) + \sum_2^t (\alpha_i - 1) \right] = e_P - n,$$

$$e'_I = e_I - \left[\left(m + t - 1 - \sum_1^{t-1} \beta_i \right) + \sum_1^{t-1} (\beta_i - 1) \right] = e_I - m.$$

Hence $me'_P \geq ne'_I$, so \mathcal{C} is reducible to \mathcal{C}' .

The proof with the inequalities in (b) is similar. \square

The next three lemmas assume that $\mathcal{C} = P^{\alpha_1} I^{\beta_1} \dots P^{\alpha_T} I^{\beta_T}$ is a forbidden picycle, that $\alpha_i \geq 2$ and $\beta_i \geq 2$ for all i , that $m \geq 2$ and that A1 through A_m hold. These lemmas, whose proofs conclude this section, form the basis of our sufficiency proof of Theorem 1 along with Lemma 1.

LEMMA 3. \mathcal{C} is reducible if $\alpha_i \geq n + 1$ for some i .

LEMMA 4. \mathcal{C} is reducible if $\beta_i \geq m + 1$ for some i .

LEMMA 5. Suppose $0 \leq k \leq m - 3, k + 2 \leq T$, and $\sum_{j=i}^{i+k} \alpha_j \leq n + k$ and $\sum_{j=1}^{i+k} \beta_j \leq m + k$ for $i = 1, \dots, T$, where $\alpha_j = \alpha_{j-T}$ and $\beta_j = \beta_{j-T}$ for $j > T$. Then either:

- (a) for each $i \in \{1, \dots, T\}$,

$$\sum_{j=i}^{i+k+1} \alpha_j \leq n + k + 1 \quad \text{and} \quad \sum_{j=i}^{i+k+1} \beta_j \leq m + k + 1; \quad \text{or}$$

- (b) $me_P - ne_I \geq n$, in which case \mathcal{C} is reducible by shortening a chain; or

(c) \mathcal{C} is reducible to \mathcal{C}' via Lemma 2(a) or 2(b) for some $t \in \{k+2, \dots, \min\{T, m\}\}$, with $h(\mathcal{C}') = h(\mathcal{C}) - [m+n+2(t-1)]$.

Proof of Lemma 3. Assume without loss of generality that $\alpha_1 \geq n+1$, and suppose that \mathcal{C} is not reducible. Since $\alpha_1 \geq n+1$, Lemma 2(a) requires $\beta_1 \leq m$. If $\alpha_2 \geq n+1$, then $\beta_2 \leq m$; if $\alpha_2 \leq n$, then $\alpha_1 + \alpha_2 \geq n+2$, and Lemma 2(a) for $t=2$ requires either $\beta_1 + \beta_2 \leq m+1$ or $\beta_1 \geq m+1$: since not $(\beta_1 \geq m+1)$, either

$$\alpha_2 \geq n+1 \text{ and } \beta_2 \leq m; \text{ or}$$

$$\alpha_2 \leq n, \alpha_1 + \alpha_2 \geq n+2, \beta_1 + \beta_2 \leq m+1.$$

If $\alpha_3 \geq n+1$ then $\beta_3 \leq m$; if $\alpha_3 \leq n$ and $\alpha_2 + \alpha_3 \geq n+2$ then $\beta_2 + \beta_3 \leq m+1$ or $\beta_2 \geq m+1$ (which is impossible since $\beta_2 \leq m$ by the preceding sentence); if $\alpha_2 + \alpha_3 \leq n+1$ then $\alpha_1 + \alpha_2 + \alpha_3 \geq n+3$, and then Lemma 2(a) for $t=3$ or A3 requires $\beta_1 + \beta_2 + \beta_3 \leq m+2$ or $\beta_1 + \beta_2 \geq m+2$ (which is precluded by the fact that $\alpha_2 + \alpha_3 \leq n+1$ implies $\alpha_2 \leq n$, hence $\alpha_1 + \alpha_2 \geq n+2$, hence $\beta_1 + \beta_2 \leq m+1$). Therefore either

$$\alpha_3 \geq n+1, \beta_3 \leq m; \text{ or}$$

$$\alpha_3 \leq n, \alpha_2 + \alpha_3 \geq n+2, \beta_2 + \beta_3 \leq m+1; \text{ or}$$

$$\alpha_2 + \alpha_3 \leq n+1, \alpha_1 + \alpha_2 + \alpha_3 \geq n+3, \beta_1 + \beta_2 + \beta_3 \leq m+2.$$

The natural continuation of this procedure to any $t \leq \min\{m, T\}$ gives either

$$\alpha_t \geq n+1, \beta_t \leq m; \text{ or}$$

$$\alpha_t \leq n, \sum_{i=1}^t \alpha_i \geq n+2, \sum_{i=1}^t \beta_i \leq m+1; \text{ or}$$

$$\vdots$$

$$\sum_{i=2}^t \alpha_i \leq n+t-2, \sum_{i=1}^t \alpha_i \geq n+t, \sum_{i=1}^t \beta_i \leq m+t-1.$$

Suppose $m \leq T$. Then the final line in the preceding display cannot hold at $t=m$ since it requires $\sum_{i=1}^m \beta_i \leq 2m-1$, whereas $\sum_{i=1}^m \beta_i \geq 2m$. In addition, if $m < T$, then continuance to $t \in \{m+1, \dots, T\}$ gives either

$$\alpha_t \geq n+1, \beta_t \leq m; \text{ or}$$

$$\alpha_t \leq n, \sum_{i=1}^t \alpha_i \geq n+2, \sum_{i=1}^t \beta_i \leq m+1; \text{ or}$$

$$\vdots$$

$$\sum_{i=m+3}^t \alpha_i \leq n+m-3, \sum_{i=m+2}^t \alpha_i \geq n+m-1, \sum_{i=m+2}^t \beta_i = 2m-2.$$

The line in this display that has $\sum_{i=k}^t \alpha_i \leq n+k$ and $\sum_{i=k-1}^t \alpha_i \geq n+k+2$ requires either

$$\sum_{i=k-1}^t \beta_i \leq m+k+1 \text{ or } \sum_{i=k-1}^{t-1} \beta_i \geq m+k+1$$

according to A($k+2$) in Lemma 2(a) to prevent \mathcal{C} from being reducible. However, $\sum_{i=k}^t \alpha_i \leq n+k$ implies $\sum_{i=k}^{t-1} \alpha_i \leq n+k-1$, or

$$\sum_{(i-1)-(k-1)}^{(t-1)} \alpha_i \leq n+(k-1),$$

and it follows from the predecessor display for $t - 1$ that

$$\sum_{(t-1)-(k-1)-1}^{(t-1)} \beta_i \leq m + k, \quad \text{i.e.,} \quad \sum_{t-k-1}^{t-1} \beta_i \leq m + k.$$

Hence $\sum_{t-k-1}^{t-1} \beta_i \geq m + k + 1$ is precluded in the line indicated for the t display, which therefore is correct as it stands.

It follows from A1–Am and irreducibility that for each $1 \leq t \leq T$ there is $0 \leq k \leq \min \{t - 1, m - 2\}$ such that

$$\sum_{t-k}^t \alpha_i \geq n + k + 1 \quad \text{and} \quad \sum_{t-k}^t \beta_i \leq m + k.$$

Beginning at T , proceed backwards through \mathcal{C} as follows. Select $k_1 \geq 1$ for which

$$\sum_{k_1}^T \alpha_i \geq n + (T - k_1) + 1 \quad \text{and} \quad \sum_{k_1}^T \beta_i \leq m + (T - k_1).$$

If $k_1 > 1$, select $1 \leq k_2 \leq k_1 - 1$ for which

$$\sum_{k_2}^{k_1-1} \alpha_i \geq n + (k_1 - 1 - k_2) + 1, \quad \sum_{k_2}^{k_1-1} \beta_i \leq m + (k_1 - 1 - k_2),$$

and continue in the obvious way back to the beginning of \mathcal{C} . In each backwards step,

$$m(P\text{-excess}) - n(I\text{-excess}) \geq m(n) - n(m - 1) = n,$$

so $me_P - ne_I \geq n$ for \mathcal{C} . However, such a \mathcal{C} is reducible: replace a P^2 segment by P to get \mathcal{C}' with $me'_P - ne'_I \geq n - m \geq 0$. Therefore our supposition that \mathcal{C} is not reducible is false. \square

Proof of Lemma 4. Assume for definiteness that $\beta_1 \geq m + 1$, and suppose \mathcal{C} is not reducible. For $t \leq \min \{m, T\}$, a procedure like that of the preceding proof, but with Lemma 2(b) instead of 2(a), gives either

$$\begin{aligned} &\beta_t \geq m + 1, \quad \alpha_{t+1} \leq n; \quad \text{or} \\ &\beta_t \leq m, \quad \beta_{t-1} + \beta_t \geq m + 2, \quad \alpha_t + \alpha_{t+1} \leq n + 1; \quad \text{or} \\ &\vdots \\ &\sum_2^t \beta_i \leq m + t - 2, \quad \sum_1^t \beta_i \geq m + t, \quad \sum_2^{t+1} \alpha_i \leq n + t - 1. \end{aligned}$$

At $t = m$, the first inequality in the final line implies $\beta_2 = \dots = \beta_m = 2$. If $m < T$, a similar display with m lines for A1–Am applies to each $t \in \{m + 1, \dots, T\}$. It follows that each $t \in \{1, \dots, T\}$ has $0 \leq k \leq \min \{t - 1, m - 1\}$ such that

$$\sum_{t-k}^t \beta_i \geq m + k + 1 \quad \text{and} \quad \sum_{t-k+1}^{t+1} \alpha_i \leq n + k.$$

Beginning at T , proceed backwards: select $k_1 \geq 1$ for which

$$\sum_{k_1}^T \beta_i \geq m + (T - k_1) + 1 \quad \text{and} \quad \sum_{k_1+1}^{T+1} \alpha_i \leq n + (T - k_1),$$

where $\alpha_{T+1} = \alpha_1$, then select $1 \leq k_2 \leq k_1 - 1$ in a similar manner if $k_1 > 1$, and so forth. In each step, $n(I\text{-excess}) - m(P\text{-excess}) \geq n(m) - m(n - 1) = m$, so $ne_I > me_P$ for \mathcal{C} . But then \mathcal{C} is not forbidden. Hence forbidden \mathcal{C} is reducible when some $\beta_i \geq m + 1$. \square

Proof of Lemma 5. Given the hypotheses of the lemma, we suppose that none of (a), (b) and (c) hold, and proceed to a contradiction.

Suppose first that the α_j do not satisfy (a), and without loss of generality take

$$\sum_{i=1}^{k+2} \alpha_i \cong n + k + 2.$$

Since $\sum_2^{k+2} \alpha_i \cong n + k$ and $\sum_1^{k+1} \beta_i \cong m + k$ by hypothesis, the presumed failure of (c) resulting from application of A($k+2$) in Lemma 2(a) requires $\sum_1^{k+2} \beta_i \cong m + k + 1$. Continuing as in the proof of Lemma 3, for each $t \in \{k+2, \dots, T, T+1, \dots, T+k+1\}$ we get either

$$\begin{aligned} \sum_{i=k}^t \alpha_i \cong n + k, & \quad \sum_{i=k-1}^t \alpha_i \cong n + k + 2, & \quad \sum_{i=k-1}^t \beta_i \cong m + k + 1; \quad \text{or} \\ \sum_{i=k-1}^t \alpha_i \cong n + k + 1, & \quad \sum_{i=k-2}^t \alpha_i \cong n + k + 3, & \quad \sum_{i=k-2}^t \beta_i \cong m + k + 2; \quad \text{or} \\ \vdots & & \\ \sum_{i=k-x}^t \alpha_i \cong n + k + x, & \quad \sum_{i=k-x-1}^t \alpha_i \cong n + k + x + 2, & \quad \sum_{i=k-x-1}^t \beta_i \cong m + k + x + 1, \end{aligned}$$

where $x = \min \{(t-2) - k, (m-3) - k, (T-2) - k\}$. As we proceed through larger values of t , the other possible inequality on the β_i sum that arises from the presumed failure of reduction by Lemma 2(a) is precluded by the inequalities on the α_i sums of that case along with those obtained at $t-1$.

It follows for each $t \in \{k+2, \dots, T+k+1\}$ that there is a $k+1 \cong y \cong \min \{t-1, m-2, T-1\}$ such that

$$\sum_{i=y}^t \alpha_i \cong n + y + 1 \quad \text{and} \quad \sum_{i=y}^t \beta_i \cong m + y.$$

We therefore have $m(\mathcal{P}\text{-excess}) - n(\mathcal{I}\text{-excess}) \cong n$ in the part of \mathcal{C} covered by these two inequalities. Although the backwards procedure used in the proof of Lemma 3 (begin at $T+k+1$, get y for this ending point; take the next t as $T+k+1-y-1$, get y for this t ; \dots) may not come out evenly by ending precisely at $k+2$, we can continue around the picycle an arbitrarily large number of times and conclude that the average difference between me_P and ne_I per revolution is at least n . Consequently, \mathcal{C} must have $me_P - ne_I \cong n$.

However, this would satisfy conclusion (b), so to maintain the supposition that none of (a), (b) and (c) holds, we need to suppose that the β_j do not satisfy (a). But then, given $\beta_i + \dots + \beta_{i+k+1} \cong m + k + 2$ for some i , a similar proof (see also the proof of Lemma 4) leads to the conclusion that $ne_I > me_P$, which contradicts forbiddenness. As in the preceding paragraph, it may be necessary to cycle backwards around \mathcal{C} a large number of times to conclude that $ne_I > me_P$. We omit the details. \square

5. Proof completion. We complete the sufficiency proof of Theorem 1 by showing that an order which satisfies A1 $_{mn}$ through A m_{mn} has no (m, n) -forbidden picycle. By Lemma 1, such an order is in $\mathcal{P}[m, n]$. As in the preceding section, the mn designation is often omitted.

Henceforth, assume that A1–A m hold for order (A, P) . We shall suppose that (A, P) has a forbidden picycle, and proceed to a contradiction.

According to our supposition, (A, P) must have a minimum-length forbidden picycle, designated as \mathcal{C} with index $(\alpha_1, \beta_1, \dots, \alpha_T, \beta_T)$. Since \mathcal{C} is not reducible, Lemma 2 tells us that $\alpha_i \geq 2$ and $\beta_i \geq 2$ for all i .

If $m = 1$ then \mathcal{C} has $e_P \geq ne_I$, and a contradiction to irreducibility follows immediately from the $IP^{n+1} \subseteq P$ part of A1 since $\alpha_i - 1 \geq n$ for some i . This proves Corollary 1.

Assume henceforth that $m \geq 2$. Then $n > m$, and therefore $\alpha_i \geq 3$ for some i . Without loss in generality we shall assume that $\alpha_1 \geq 3$. Consider T versus m .

Suppose first that $T \geq m$. Then Lemmas 3 and 4, and induction with Lemma 5 if $m \geq 3$, give

$$\sum_{j=i}^{i+m-2} \alpha_j \leq n + m - 2 \quad \text{and} \quad \sum_{j=i}^{i+m-2} \beta_j \leq m + m - 2$$

for $i = 1, \dots, T$, where as usual $\alpha_j = \alpha_{j-T}$ and $\beta_j = \beta_{j-T}$ when $j > T$. The inequality on the β_j implies that $\beta_i = 2$ for all i . Then, in view of Lemma 2(b) for Am , irreducibility of \mathcal{C} requires either $\sum_i^{i+m-1} \alpha_j \leq n + m - 1$ or $\sum_i^{i+m-2} \alpha_j \geq n + m - 1$ for each i . Since the latter inequality is false,

$$\sum_{j=i}^{i+m-1} \alpha_j \leq n + m - 1 \quad (i = 1, \dots, T).$$

Therefore $\alpha_1 + \dots + \alpha_T \leq T(n + m - 1)/m$. But then

$$me_P \leq m[T(n + m - 1)/m - T] = T(n - 1) < Tn = ne_I,$$

which contradicts forbiddenness. Hence $T \geq m$ yields a contradiction to our supposition that (A, P) has a forbidden picycle.

Assume henceforth that $T < m$. By Lemmas 3 and 4 if $T = 1$, and by Lemmas 3 and 4 and induction with Lemma 5 if $T \geq 2$, we get

$$(2) \quad \sum_{i=1}^T \alpha_i \leq n + T - 1 \quad \text{and} \quad \sum_{i=1}^T \beta_i \leq m + T - 1.$$

Let $\Delta = me_P - ne_I$. Since \mathcal{C} is irreducible, $m > \Delta \geq 0$. Let \mathcal{C}_0 be the picycle obtained from \mathcal{C} by shortening the initial chain in \mathcal{C} from length α_1 to $\alpha_1 - 1$. \mathcal{C}_0 has index $(\alpha_1 - 1, \beta_1, \dots, \alpha_T, \beta_T)$, and is not forbidden since $m(P\text{-excess of } \mathcal{C}_0) - n(I\text{-excess of } \mathcal{C}_0) = \Delta - m < 0$.

We now form forbidden picycles $\mathcal{C}^{(s)}$ with excesses $e_P^{(s)}$ and $e_I^{(s)}$ for $s = 1, 2, \dots$, by repetitions of \mathcal{C} , mixed with repetitions of \mathcal{C}_0 when $\Delta > 0$. In what follows, $\mathcal{C}\mathcal{C}$ denotes \mathcal{C} followed by a copy of itself; it has index $(\alpha_1, \beta_1, \dots, \alpha_T, \beta_T, \alpha_1, \beta_1, \dots, \alpha_T, \beta_T)$. Likewise, $\mathcal{C}\mathcal{C}_0$ denotes \mathcal{C} followed by \mathcal{C}_0 . It has index $(\alpha_1, \dots, \beta_T, \alpha_1 - 1, \beta_1, \dots, \beta_T)$ and length $2\sum_1^T (\alpha_i + \beta_i) - 1$. Expressions $\mathcal{C}\mathcal{C}\mathcal{C}_0, \mathcal{C}\mathcal{C}\mathcal{C}, \mathcal{C}\mathcal{C}_0\mathcal{C}, \dots$ are defined similarly.

Let $\mathcal{C}^{(1)} = \mathcal{C}$, and for each $s \geq 1$ take

$$\mathcal{C}^{(s+1)} = \begin{cases} \mathcal{C}^{(s)}\mathcal{C} & \text{if } m > me_P^{(s)} - ne_I^{(s)} + \Delta, \\ \mathcal{C}^{(s)}\mathcal{C}_0 & \text{if } me_P^{(s)} - ne_I^{(s)} + \Delta \geq m, \end{cases}$$

so that $m > me_P^{(s)} - ne_I^{(s)} \geq 0$ for all s . The index of $\mathcal{C}^{(s)}$ will be written as

$$(\alpha_1^{(s)}, \beta_1^{(s)}, \dots, \alpha_{sT}^{(s)}, \beta_{sT}^{(s)}),$$

and subscripts on $\alpha^{(s)}$ and $\beta^{(s)}$ will be taken modulo sT when they exceed sT . It should be noted that each $\mathcal{C}^{(s)}$ is forbidden, and no $\mathcal{C}^{(s)}$ is reducible by shortening a chain.

Hence the type of reducibility specified in Lemma 5(b) never applies to $\mathcal{C}^{(s)}$.

We shall consider $s = 2, 3$, then generalize to larger s . Our aim is to show that the “shortest forbidden picycle” supposition for \mathcal{C} along with $1 \leq T < m$ forces the contradiction that m is infinite.

Consider $s = 2$: $\mathcal{C}^{(2)}$ is $\mathcal{C}\mathcal{C}$ or $\mathcal{C}\mathcal{C}_0$. In either case, (2) implies that

$$(3) \quad \sum_{j=i}^{i+T-1} \alpha_j^{(2)} \leq n + T - 1 \quad \text{and} \quad \sum_{j=i}^{i+T-1} \beta_j^{(s)} \leq m + T - 1$$

for $i = 1, \dots, 2T$. We eliminate the case of $T = m - 1$ before considering other cases.

Suppose $T = m - 1$. Then, by (2), $\beta_i = 2$ for all i , so $\beta_i^{(2)} = 2$ for all i . In addition, $\sum_1^T \alpha_i = \sum_1^{m-1} \alpha_i \leq n + m - 2$, and therefore

$$\sum_{j=i}^{i+T-1} \alpha_j^{(2)} \leq n + m - 2 \quad \text{for } i = 1, \dots, 2T.$$

If $\sum_i^{i+T} \alpha_j^{(2)} \leq n + m - 1$ for $i = 1, \dots, 2T$ then, as in the paragraph preceding (2), we get a contradiction to forbiddenness:

$$me_p^{(2)} \leq m[2T(n + m - 1)/m - 2T] = 2T(n - 1) < 2Tn = ne_i^{(2)}.$$

Therefore $\sum_i^{i+T} \alpha_j^{(2)} \geq n + m$ for some i . Then, according to Lemma 2 with $t = m$, $\mathcal{C}^{(2)}$ is reducible to \mathcal{C}' with

$$h(\mathcal{C}') = h(\mathcal{C}^{(2)}) - [m + n + 2(m - 1)] < h(\mathcal{C}),$$

where the inequality follows from the fact that

$$h(\mathcal{C}^{(2)}) - h(\mathcal{C}) \leq h(\mathcal{C}) \leq (n + m - 2) + 2(m - 1).$$

But then \mathcal{C}' is a forbidden picycle that is shorter than \mathcal{C} , in contradiction to our minimality supposition for \mathcal{C} . Therefore $T = m - 1$ is impossible.

Assume henceforth that $T \leq m - 2$. Given (3), we apply Lemma 5 to $\mathcal{C}^{(2)}$ with $2T$ in place of T for this application. Since $T - 1 \leq m - 3$ and $(T - 1) + 2 \leq 2T$, and by virtue of (3), the hypotheses of Lemma 5 for $\mathcal{C}^{(2)}$ hold for $k = T - 1$. Since $m \geq me_p^{(2)} - ne_i^{(2)}$ by construction, conclusion (b) is false. Therefore either (a) or (c) holds for $k = T - 1$. That is, either

$$(4) \quad \sum_{j=i}^{i+T} \alpha_j \leq n + T \quad \text{and} \quad \sum_{j=i}^{i+T} \beta_j \leq m + T \quad \text{for } i = 1, \dots, 2T,$$

or $\mathcal{C}^{(2)}$ is reducible to \mathcal{C}' with $h(\mathcal{C}') = h(\mathcal{C}^{(2)}) - [m + n + 2(t - 1)]$ for some $t \geq k + 2 = T + 1$. Suppose the latter possibility holds. Then

$$h(\mathcal{C}') \leq h(\mathcal{C}^{(2)}) - [m + n + 2T].$$

By (2), $h(\mathcal{C}) \leq (n + T - 1) + (m + T - 1) = m + n + 2T - 2$. Since $h(\mathcal{C}^{(2)}) - h(\mathcal{C})$ is either $h(\mathcal{C})$ or $h(\mathcal{C}) - 1$, it follows that $h(\mathcal{C}^{(2)}) - h(\mathcal{C}) \leq m + n + 2T - 2$, i.e., that

$$h(\mathcal{C}^{(2)}) - [m + n + 2T] \leq h(\mathcal{C}) - 2.$$

Therefore $h(\mathcal{C}') \leq h(\mathcal{C}) - 2$. But this contradicts minimality for \mathcal{C} .

Therefore (4) holds. By repeating the use of Lemma 5 applied to $\mathcal{C}^{(2)}$ for increasing values of $k \geq T - 1$, we conclude that, with $K = \min \{m - 3, 2T - 2\}$,

$$\sum_{j=i}^{i+K+1} \alpha_j^{(2)} \leq n + K + 1 \quad \text{and} \quad \sum_{j=i}^{i+K+1} \beta_j^{(2)} \leq m + K + 1$$

for $i = 1, \dots, 2T$. Suppose here that $K = m - 3$, so that

$$\sum_{j=i}^{i+m-2} \alpha_j^{(2)} \leq n + m - 2 \quad \text{and} \quad \sum_{j=i}^{i+m-2} \beta_j^{(2)} \leq m + m - 2$$

for $i \leq 2T$. Then $\beta_i^{(2)} = 2$ for all i and, as in the paragraph following (3), we obtain a contradiction: either $\sum_i^{i+m-1} \alpha_j^{(2)} \leq n + m - 1$ for all i , which gives a contradiction to forbiddenness, or else $\sum_i^{i+m-1} \alpha_j^{(2)} \geq n + m$ for some i , in which case Lemma 2 with $t = m$ shows that $\mathcal{C}^{(2)}$ is reducible to \mathcal{C}' with $h(\mathcal{C}') < h(\mathcal{C})$. Therefore $K = 2T - 2 < m - 3$, so that $2T \leq m - 2$ with

$$(5) \quad \sum_1^{2T} \alpha_i^{(2)} \leq n + 2T - 1 \quad \text{and} \quad \sum_1^{2T} \beta_i^{(2)} \leq m + 2T - 1.$$

Assume henceforth that $2T \leq m - 2$ along with (5), and consider $\mathcal{C}^{(3)}$. According to the definition of $\mathcal{C}^{(3)}$, each $\sum_i^{i+2T-1} \beta_j^{(3)}$ includes each β_i ($i = 1, \dots, T$) exactly twice so that these sums are equal: (5) gives

$$(6) \quad \sum_i^{i+2T-1} \beta_j^{(3)} \leq m + 2T - 1 \quad \text{for } i = 1, \dots, 3T.$$

In a similar manner, if $\mathcal{C}^{(3)}$ is anything other than $\mathcal{C}\mathcal{C}_0\mathcal{C}$, then the first inequality in (5) implies that

$$(7) \quad \sum_i^{i+2T-1} \alpha_j^{(3)} \leq n + 2T - 1 \quad \text{for } i = 1, \dots, 3T.$$

When this is true, an analysis similar to that of the preceding two paragraphs, applied now to $\mathcal{C}^{(3)}$ with $3T$ in place of T for Lemma 5, leads to the conclusion that $3T \leq m - 2$ with

$$(8) \quad \sum_1^{3T} \alpha_i^{(3)} \leq n + 3T - 1 \quad \text{and} \quad \sum_1^{3T} \beta_i^{(3)} \leq m + 3T - 1.$$

The only way for (7) to fail is to have $\mathcal{C}^{(3)} = \mathcal{C}\mathcal{C}_0\mathcal{C}$ and $\sum_1^{2T} \alpha_i^{(2)} = n + 2T - 1$. We consider this further in the next paragraph.

Suppose $\mathcal{C}^{(3)} = \mathcal{C}\mathcal{C}_0\mathcal{C}$ with $\sum_1^{2T} \alpha_i^{(2)} = n + 2T - 1$. Then instead of (7) we have

$$\sum_i^{i+2T-1} \alpha_j^{(3)} = \begin{cases} n + 2T - 1 & \text{if } i \leq T + 1 \text{ or } i \geq 2T + 2, \\ n + 2T & \text{if } T + 2 \leq i \leq 2T + 1. \end{cases}$$

Since $\alpha_i^{(3)} \geq 2$ for all i , it is true that $\sum_i^{i+2T-2} \alpha_j^{(3)} \leq n + 2T - 2$ for $i = 1, \dots, 3T$. Therefore, taking account of (6) with $\beta_i^{(3)} \geq 2$, the hypotheses of Lemma 5 for $\mathcal{C}^{(3)}$ hold at $k = 2T - 2$. Since (b) of Lemma 5 is false by construction for $\mathcal{C}^{(3)}$, it follows for $k = 2T - 2$ that either (a) holds, i.e. either (6) and (7) hold, or else (c) holds with $\mathcal{C}^{(3)}$ reducible to \mathcal{C}' via Lemma 2 for some $t \in \{k + 2, \dots, \min\{m, 3T\}\}$ with $h(\mathcal{C}') = h(\mathcal{C}^{(3)}) - [m + n + 2(t - 1)]$. Suppose (c) holds. Then it could hold for $t = k + 2 = 2T$ only if $\sum_i^{i+2T-1} \beta_j^{(3)} \geq m + 2T$ for some i [to apply A(2T) in Lemma 2], and this is false by (6). Therefore (c) requires $t \geq 2T + 1$, in which case

$$h(\mathcal{C}') \leq h(\mathcal{C}^{(3)}) - [m + n + 4T] = 3h(\mathcal{C}) - 1 - [m + n + 4T],$$

so that $h(\mathcal{C}') - h(\mathcal{C}) \leq 2h(\mathcal{C}) - 1 - [m + n + 4T]$. In fact, $\mathcal{C}^{(2)} = \mathcal{C}\mathcal{C}_0$ and the presumed $\sum_1^{2T} \alpha_i^{(2)} = n + 2T - 1$ give $2h(\mathcal{C}) = \sum_1^{2T} (\alpha_i^{(2)} + \beta_i^{(2)}) + 1 \leq n + m + 4T - 1$, so $2h(\mathcal{C}) -$

$1 - [m + n + 4T] \leq -2$. Therefore $h(\mathcal{C}') < h(\mathcal{C})$, contrary to minimality for \mathcal{C} , and we conclude that (6) and (7) hold as stated.

Having arrived at (8) with $3T \leq m - 2$, we proceed by induction. Our induction hypothesis is $NT \leq m - 2$ and

$$(9) \quad \sum_1^{NT} \alpha_i^{(N)} \leq n + NT - 1 \quad \text{and} \quad \sum_1^{NT} \beta_i^{(N)} \leq m + NT - 1$$

for $N \geq 3$. We desire to show that $(N + 1)T \leq m - 2$ and that (9) holds with $N + 1$ in place of N . Given (9), the construction of $\mathcal{C}^{(2)}$ with repetitions of the original β_i gives

$$(10) \quad \sum_{j=i}^{i+NT-1} \beta_j^{(N+1)} \leq m + NT - 1 \quad \text{for } i = 1, \dots, (N + 1)T.$$

We also wish to have

$$(11) \quad \sum_{j=i}^{i+NT-1} \alpha_j^{(N+1)} \leq n + NT - 1 \quad \text{for } i = 1, \dots, (N + 1)T.$$

This follows from (9) unless $\mathcal{C}^{(N+1)}$ has an internal \mathcal{C}_0 and ends with \mathcal{C} , as in $\mathcal{C}^{(N+1)} = \mathcal{C} \cdots \mathcal{C}\mathcal{C}_0 \cdots \mathcal{C}$, and the first inequality in (9) is an equality. Then the sum in (11) will equal $n + NT$ when i lies in a \mathcal{C}_0 block after the initial $\alpha_1 - 1$ in that block. Moreover, $h(\mathcal{C}^{(N+1)}) = h(\mathcal{C}^{(N)}) + h(\mathcal{C}) \leq n + m + 2NT - 2$, the latter by (9).

Suppose (11) is violated as described. Then the hypotheses of Lemma 5 for $\mathcal{C}^{(N+1)}$ hold at $k = NT - 2$, and we conclude from the lemma that either (10) and (11) hold or else $\mathcal{C}^{(N+1)}$ is reducible to \mathcal{C}' via Lemma 2 with $h(\mathcal{C}') = h(\mathcal{C}^{(N+1)}) - [m + n + 2(t - 1)]$ for some $t \geq k + 2 = NT$. However, this can hold at $t = NT$ only if $\sum_i^{i+NT-1} \beta_j^{(N+1)} \geq m + NT$ for some i [to apply A(NT) in Lemma 2], and this is false by (10). Therefore a reduction from $\mathcal{C}^{(N+1)}$ to \mathcal{C}' requires $t \geq NT + 1$, in which case

$$\begin{aligned} h(\mathcal{C}') &\leq h(\mathcal{C}^{(N+1)}) - [m + n + 2NT] \\ &= h(\mathcal{C}^{(N)}) + h(\mathcal{C}) - [m + n + 2NT] \\ &\leq h(\mathcal{C}) - 2, \end{aligned}$$

contrary to minimality for \mathcal{C} . Therefore (10) and (11) hold.

Given (10) and (11), we apply Lemma 5 to $\mathcal{C}^{(N+1)}$, beginning at $k = NT - 1$. Since conclusion (b) never holds and since conclusion (c) at any step contradicts the minimality of \mathcal{C} , we require (a) in all cases. If $\min\{m - 3, (N + 1)T - 2\} = m - 3$ then a contradiction obtains as in the paragraph preceding (5). Hence $(N + 1)T - 2 \leq m - 4$, or $(N + 1)T \leq m - 2$, along with

$$\sum_1^{(N+1)T} \alpha_i^{(N+1)} \leq n + (N + 1)T - 1 \quad \text{and} \quad \sum_1^{(N+1)T} \beta_i^{(N+1)} \leq m + (N + 1)T - 1,$$

which is conclusion (a) for $k = (N + 1)T - 2$. This verifies the desired induction conclusions.

It then follows that $m \geq NT$ for all integers N , which is obviously absurd since $T \geq 1$. Therefore (A, P) has no minimum-length forbidden picycle and hence no forbidden picycle.

REFERENCES

[1] P. C. FISHBURN, *Utility Theory for Decision Making*, John Wiley, New York, 1970.
 [2] ———, *Mathematics of Decision Theory*, Mouton, the Hague, 1972.

- [3] ———, *Restricted thresholds for interval orders: a case of nonaxiomatizability by a universal sentence*, J. Math. Psychol., 24 (1981), pp. 276–283.
- [4] R. D. LUCE, *Semiorders and a theory of utility discrimination*, Econometrica, 24 (1956), pp. 178–191.
- [5] B. G. MIRKIN, *Description of some relations on the set of real-line intervals*, J. Math. Psychol., 9 (1972), pp. 243–252.
- [6] I. RABINOVITCH, *The Scott–Suppes theorem on semiorders*, J. Math. Psychol., 15 (1977), pp. 209–212.
- [7] F. S. ROBERTS, *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Addison-Wesley, Reading, MA, 1979.
- [8] D. SCOTT, *Measurement structures and linear inequalities*, J. Math. Psychol., 1 (1964), pp. 233–247.
- [9] D. SCOTT AND P. SUPPES, *Foundational aspects of theories of measurement*, J. Symbolic Logic, 23 (1958), pp. 113–128.
- [10] R. J. TITIEV, *Measurement structures in classes that are not universally axiomatizable*, J. Math. Psychol., 9 (1972), pp. 200–205.

ON THE HARMONIOUS COLORING OF GRAPHS*

J. E. HOPCROFT† AND M. S. KRISHNAMOORTHY‡

Abstract. In this report we define a new coloring of graphs, namely harmonious coloring of graphs, which arises as an extension of harmonious and graceful numbering of graphs. We show that the harmonious coloring problem for general graphs is NP-complete.

Key words. NP-complete, graceful number, harmonious number, harmonious coloring, perfect hash function

1. Introduction. Various coloring problems such as the map coloring, vertex coloring and edge coloring problems have been studied in the literature [6]. The map coloring problem is to color the regions of a planar map with a minimum number of colors, such that no two adjacent regions are colored with the same color. The vertex coloring problem is to color the vertices of a graph with a minimum number of colors, such that no two adjacent vertices are colored the same color. The edge coloring problem is to color the edges of a graph with a minimum number of colors, such that no two adjacent edges are colored the same color. The complexities of these coloring problems have been studied in the past [1], [3].

In this paper, we define a new coloring problem—the *harmonious coloring problem* of a graph. Assign colors to the vertices of a graph. The color of an edge is defined to be the unordered pair of the colors of its end vertices. Then the harmonious coloring problem is to find the minimum number of colors needed to color the vertices of a graph such that all edge colors are distinct. For example, the graph in Fig. 1 is harmoniously colored with four colors.

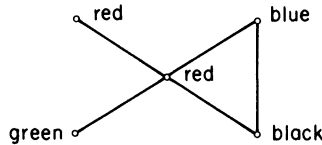


FIG. 1

Unlike vertex coloring, two colors are not sufficient to harmoniously color a tree. It is easy to see that if a graph with n vertices, e edges, can be harmoniously colored with r colors then

$$\frac{r(r+1)}{2} \geq e.$$

For example, the tree in Fig. 2 can be harmoniously colored with five colors, but the addition of any other edge requires an additional color.

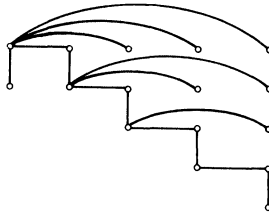


FIG. 2

* Received by the editors April 22, 1982 and in revised form September 17, 1982. This research was supported, in part, by the Office of Naval Research under contract N00014-76-C-0018.

† Department of Computer Science, Cornell University, Ithaca, New York 14853.

‡ Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181.

For a path P_n , it is easy to obtain a minimum harmonious coloring. If we do not allow the edge colors of the form (i, i) , harmonious coloring of the path tree could be obtained from the Eulerian path in a complete graph K_{2l+1} . For example, the path tree P_{11} could be harmoniously colored with five colors as shown in Fig. 3.

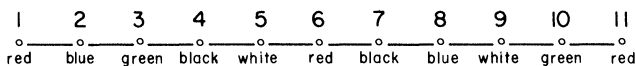


FIG. 3

In general, the harmonious coloring problem may be viewed as an Eulerian path decomposition in graphs.

The harmonious coloring problem is closely related to the harmonious numbering and graceful numbering problems [4], [5]. In the harmonious numbering problem, distinct numbers are assigned to the vertices of a graph and the edge number is computed as the sum of the numbers assigned to its end vertices taken modulo the number of edges of the graph. In graceful numbering, distinct numbers are assigned to the vertices of a graph, and the edge number is computed as the absolute difference of the numbers assigned to the end vertices of the graph. In both problems, the goal is to have all the edge numbers distinct. In general, one may assign distinct numbers to vertices and compute the edge number as a function of the numbers assigned to its end vertices. In harmonious coloring, our function is the unordered pair of its end vertex colors. In order to make the problem interesting, we allow assignment of the same color (number) to several vertices of the graph, but require distinct edge colors and a minimum number of vertex colors.

Another problem closely related to the harmonious coloring problem is that of constructing minimal perfect hash functions [2]. Form a graph whose vertices are the first and last letters of the words with an edge between two vertices if there is a word whose first and last letters are the end vertices. Then the minimal perfect hash function problem is closely related to that of harmonious coloring of graphs. For definitions related to graph theoretic terms see Harary [6] and definitions related to NP-completeness see Aho, Hopcroft and Ullman [1] or Garey and Johnson [3].

In this report, we show that the harmonious coloring problem for graphs is NP-complete by reducing the 3-SAT problem to the harmonious coloring problem.

2. Problem statement and main result. The harmonious coloring problem for graphs may be formally stated as follows:

Instance. Graph $G = (V, E)$, positive integer $k \leq |V|$.

Question. Is G k -harmoniously colorable, i.e., does there exist a function $f: V \rightarrow \{1, 2, \dots, k\}$, such that for every pair of edges $\{u, v\}, \{x, y\} \in E, \{x, y\} \neq \{u, v\}, \{f(u), f(v)\} \neq \{f(x), f(y)\}$.

We prove that this problem is NP-complete, by reducing an already known NP-complete problem, namely the 3-SAT problem, to the harmonious coloring problem.

THEOREM. *The harmonious coloring problem for graphs is NP-complete.*

Proof. It is fairly easy to see that the harmonious coloring problem is in NP, since a nondeterministic algorithm need only guess a mapping f , and check in polynomial time whether it is a harmonious coloring.

We transform 3-SAT to harmonious coloring. Let $U = \{u_1, u_2, \dots, u_n\}$ be the set of variables and $c = \{c_1, c_2, \dots, c_m\}$ be the set of clauses of an instance c of 3-SAT. We construct a graph $G = (V, E)$ and a positive integer $k \leq |V|$ such that G is k -harmoniously colorable if and only if c is satisfiable.

The vertices of G are $V = V_1 \cup V_2 \cup \dots \cup V_8$ where

$$\begin{aligned} V_1 &= \{1, \dots, n+2\}, \\ V_2 &= \{1', \dots, n+2'\}, \\ V_3 &= \{1'', \dots, n+2''\}, \\ V_4 &= \{1''', \dots, n+2'''\}, \\ V_5 &= \{\langle c, i, j \rangle \mid i = 1, \dots, m, j = 1, \dots, 7\}, \\ V_6 &= \{\langle f, i, r \rangle \mid i = 1, \dots, m, r = 1, \dots, 5\}, \\ V_7 &= \{\langle g, i, r \rangle \mid i = 1, \dots, m, r = 1, \dots, 5\}, \\ V_8 &= \{\langle q, i \rangle \mid i = 1, \dots, m\}. \end{aligned}$$

The vertices in V_1 and V_2 correspond to the variables u_i and \bar{u}_i respectively plus two additional vertices. Vertices in V_3 and V_4 have a similar correspondence as those in V_1 and V_2 . V_5 has seven vertices for each clause corresponding to the seven possible assignments to the three variables of the clause that make it true. The symbol c simply is a reminder that the vertices correspond to clauses. Vertices in V_6 and V_7 are forcing vertices. Each vertex in V_8 corresponds to a clause. The edges in graph G are as follows:

1. There is a complete graph on V_1 .
2. There is a complete graph on V_2 .
3. There is a complete graph on $V_5 \cup V_6$.
4. The vertices $n+1, n+2$ are connected to all vertices in V_2, V_5 , and V_6 .
5. The vertices $n+1', n+2'$ are connected to all vertices in V_1, V_5 , and V_6 .
6. There are also edges $\{(i, j) \mid i \in V_1, j \in V_2, i \neq j\}$.

Let G_1 be the subgraph with vertices V_1, V_2, V_5 and V_6 and the edges described in 1, 2, 3, 4, 5, 6. For all edges to have distinct colors each vertex in V_1, V_2, V_5, V_6 must have a distinct color. Additional edges are added to G_1 as described in the next two steps.

7. Let the clause c_i contain literals u_a, u_b, u_c such that $a < b < c$. Intuitively the vertices $\langle c, i, 1 \rangle, \dots, \langle c, i, 7 \rangle$ may be thought of as $c_{F_a F_b T_c}^i, c_{F_a T_b F_c}^i, c_{F_a T_b T_c}^i, c_{T_a F_b F_c}^i, c_{T_a F_b T_c}^i, c_{T_a T_b F_c}^i$, and $c_{T_a T_b T_c}^i$. Then vertex a is connected to all those vertices $\langle c, i, r \rangle$ that have false assignment for u_a (i.e. $c_{F_a **}^i$; the symbol $*$ denotes the “don’t care” condition). Vertex a' is connected to all those vertices $\langle c, i, r \rangle$ that have true assignment for u_a (i.e. $c_{T_a **}^i$). Similarly, vertex b is connected to all those vertices $\langle c, i, r \rangle$ that have false assignment for u_b (i.e. $c_{* F_b *}^i$). Vertex b' is connected to all those vertices $\langle c, i, r \rangle$ that have true assignment for u_b (i.e. $c_{* T_b *}^i$). Vertex c is connected to all those vertices $\langle c, i, r \rangle$ that have false assignment for u_c (i.e. $c_{** F_c}^i$). Vertex c' is connected to all those vertices $\langle c, i, r \rangle$ that have true assignment for u_c (i.e. $c_{** T_c}^i$). (It may be verified that vertex a is connected to $\langle c, i, 1 \rangle, \langle c, i, 2 \rangle, \langle c, i, 3 \rangle$ and vertex a' is connected to $\langle c, i, 4 \rangle, \langle c, i, 5 \rangle, \langle c, i, 6 \rangle$ and $\langle c, i, 7 \rangle$. Vertex b is connected to $\langle c, i, 1 \rangle, \langle c, i, 4 \rangle, \langle c, i, 5 \rangle$ and b' is connected to $\langle c, i, 2 \rangle, \langle c, i, 3 \rangle, \langle c, i, 6 \rangle$ and $\langle c, i, 7 \rangle$. Vertex c is connected to $\langle c, i, 2 \rangle, \langle c, i, 4 \rangle, \langle c, i, 6 \rangle$ and c' is connected to $\langle c, i, 1 \rangle, \langle c, i, 3 \rangle, \langle c, i, 5 \rangle$ and $\langle c, i, 7 \rangle$.)

If u_d or \bar{u}_d does not appear in clause c_j then vertices d and d' are connected to all vertices $\langle c, j, r \rangle$ for $r = 1$ to 7. For example, if the second clause contains $(u_1 + \bar{u}_2 + u_3)$, then $\langle c, 2, 1 \rangle, \dots, \langle c, 2, 7 \rangle$ may be thought of as $c_{F_1 F_2 T_3}^2, c_{F_1 T_2 F_3}^2, \dots, c_{T_1 T_2 T_3}^2$.

- 1 is connected to $c_{F_1 F_2 T_3}^2, c_{F_1 T_2 F_3}^2, c_{F_1 T_2 T_3}^2$.
- 1' is connected to $c_{T_1 F_2 F_3}^2, c_{T_1 F_2 T_3}^2, c_{T_1 T_2 F_3}^2, c_{T_1 T_2 T_3}^2$.

- 2 is connected to $c_{F_1 F_2 T_3}^2, c_{T_1 F_2 F_3}^2, c_{T_1 F_2 T_3}^2$.
- 2' is connected to $c_{F_1 T_2 F_3}^2, c_{F_1 T_2 T_3}^2, c_{T_1 T_2 F_3}^2, c_{T_1 T_2 T_3}^2$.
- 3 is connected to $c_{F_1 T_2 F_3}^2, c_{T_1 F_2 F_3}^2, c_{T_1 T_2 F_3}^2$.
- 3' is connected to $c_{F_1 F_2 T_3}^2, c_{F_1 T_2 T_3}^2, c_{T_1 F_2 T_3}^2, c_{T_1 T_2 T_3}^2$.

The rest of the literals are connected to all of $\langle c, 2, 1 \rangle, \dots, \langle c, 2, 7 \rangle$.

With the edges described in step 7, each vertex j is not connected to $c_{T_*}^i$ for all those clauses i , in which literal u_j appears. Also each vertex in V_5 is not connected to exactly three vertices of G_1 . For example, if the i th clause contains literals u_s, u_t, u_p then $\langle c, i, 2 \rangle$ (i.e. $c_{F_s T_p F_t}^i$) is not connected to s', t, p' .

8. The vertices in $V_6 \langle f, i, r \rangle, r = 1$ to 5 are connected to all vertices p and p' such that neither u_p nor \bar{u}_p appears in the i th clause. With the edges described in 8, each vertex in V_6 is not connected to exactly six vertices. For example, if the i th clause contains literals u_q, u_t, u_p then $\langle f, i, 1 \rangle, \dots, \langle f, i, 5 \rangle$ are not connected to s, s', p, p', t, t' .

This completes the description for graph G_1 (one connected component of G). The other connected component of G is G_2 , which has vertex set V_3, V_4, V_7 , and V_8 and the edges as described below.

- 9. Vertex i'' in V_3 is connected to i''' in V_4 for all $i = 1$, to n .
- 10. Vertex i'' is connected to $\langle q, j \rangle$ if u_i appears in clause c_j .
- 11. Vertex i''' is connected to $\langle q, j \rangle$ if \bar{u}_i appears in clause c_j .
- 12. If u_i or \bar{u}_i appears in clause c_j , then vertices i'' and i''' are connected to vertex $\langle g, j, r \rangle$ for $r = 1$, to 5.

The idea behind these connections is that we should be able to harmoniously color the graph G_2 , with the same colors used to harmoniously color G_1 , if and only if the given expression is satisfiable. If all the literals take a false value in a clause c_j , then we cannot color $\langle q, j \rangle$ with any of the earlier colors; therefore we will not be able to harmoniously color the graph G .

As an example, let us construct a graph (see Fig. 4) corresponding to the following 3-SAT problem,

$$U = \{u_1, u_2, u_3\}, \quad c_1 = (u_1 + u_2 + u_3),$$

$$c = \{c_1, c_2\}, \quad c_2 = (\bar{u}_1 + \bar{u}_2 + \bar{u}_3).$$

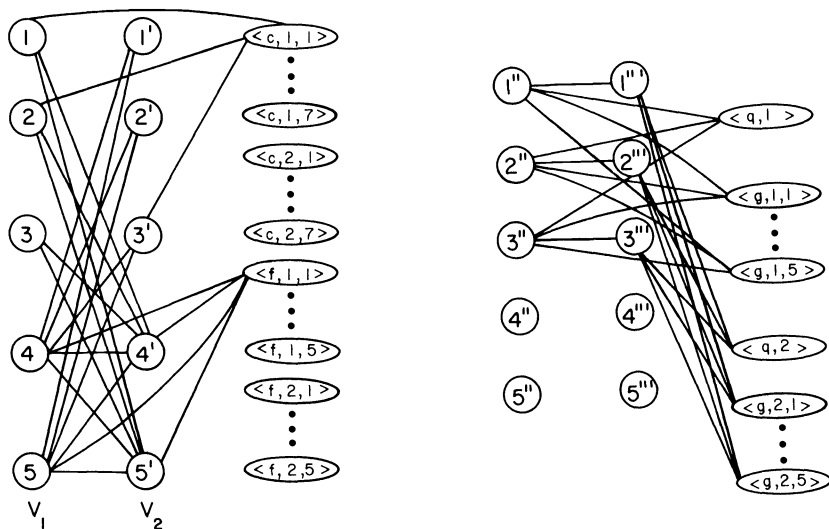


FIG. 4

For a given 3-SAT problem with n literals and m clauses, the constructed graph has $4n + 8 + 18m$ vertices—a polynomial in the size of the 3-SAT problem. In the example above we have 56 vertices in the constructed graph.

Next, we will show that G is colorable with $2n + 4 + 12m$ colors (number of vertices in G_1) if and only if the given 3-SAT problem is satisfiable. Assume the given 3-SAT problem is satisfiable, and fix a satisfying assignment. Color the vertices of V_1 and V_2 with $2n + 4$ distinct colors.

Color the vertices in V_5 as follows: color $\langle c, i, 1 \rangle, \langle c, i, 2 \rangle, \dots, \langle c, i, 7 \rangle$ with $c_{FFT}^i, c_{FTF}^i, c_{FTT}^i, c_{TFF}^i, c_{TFT}^i, c_{TTF}^i$, and c_{TTT}^i respectively. Color the vertices in V_6 with $5m$ distinct colors.

So far, we have used all $2n + 4 + 12m$ colors and we have not violated any of the coloring constraints. Color the vertices in V_7 with same colors used in coloring of V_6 . Color vertices in V_3 with the satisfying assignments of the corresponding variables. (These will also be the colors assigned to the vertices in V_1 or V_2 .) Color the vertices in V_4 with the negation of the corresponding variables. (Again these will be the colors assigned to the vertices in V_1 or V_2 .) Color the vertices in V_8 with c_{x_a, x_b, x_c}^l where x_a, x_b, x_c are the assignment of literals in the l th clause. It can easily be verified that this is a valid harmonious coloring.

In the example, vertices 1, 2, 3, 4, 5 in V_1 are colored with T_1, T_2, T_3, T_4, T_5 and $1', 2', 3', 4', 5'$ in V_2 are colored with F_1, F_2, F_3, F_4 and F_5 . Vertices $\langle c, 1, 1 \rangle, \dots, \langle c, 1, 7 \rangle$ are colored with $c_{FFT}^1, \dots, c_{TTT}^1$. Vertices $\langle c, 2, 1 \rangle, \dots, \langle c, 2, 7 \rangle$ are colored with $c_{FFT}^2, \dots, c_{TTT}^2$. Vertices $\langle f, 1, 1 \rangle, \dots, \langle f, 1, 5 \rangle$ are colored with f_1^1, \dots, f_5^1 , and $\langle f, 2, 1 \rangle, \dots, \langle f, 2, 5 \rangle$ are colored with f_1^2, \dots, f_5^2 . Vertices $1'', 2'', 3'', 4'', 5''$ are colored with T_1, T_2, T_3, T_4, T_5 (i.e. $u_1 = T, u_2 = T, u_3 = F$) and $1''', 2''', 3''', 4''', 5'''$ are colored with F_1, F_2, T_3, F_4, F_5 . Vertices $\langle q, 1 \rangle, \langle q, 2 \rangle$ are colored with c_{TTF}^1 and c_{FFT}^2 respectively.

On the other hand we will prove that if there is a harmonious coloring of G then the 3-SAT problem is satisfiable.

Let the colors assigned to vertices in V_1 and V_2 be T_1, T_2, \dots, T_{n+2} , and F_1, \dots, F_{n+2} and let the colors assigned to vertices in V_5 be of the color type $c_{FFT}^i, \dots, c_{TTT}^i$ and vertices in V_6 be colored with f_1^i, \dots, f_5^i . (We could rearrange the colors if it is done in some other manner.)

CLAIM 1. *Vertices in V_3 (of degree >1) and V_4 cannot be colored with $c_{FFT}^i, \dots, c_{TTT}^i$ or f_1^i, \dots, f_5^i .*

This is so because exactly three vertices are not connected to each of $c_{FFT}^i, \dots, c_{TTT}^i$ and exactly six vertices are not connected to f_1^i, \dots, f_5^i . But vertices in V_3 and V_4 are connected to at least seven vertices and hence the claim.

CLAIM 2. *Vertices in V_7 must be colored with f_1^i, \dots, f_5^i .*

They cannot be colored with T_i or F_i by Claim 1 (none is left). Colors of type $c_{FFT}^i, \dots, c_{TTT}^i$ cannot be given to V_6 because each vertex in V_6 (of degree >0) is connected to exactly six vertices.

This forces us to color the vertices in V_8 with a color of type $c_{FFT}^i, \dots, c_{TTT}^i$. This will force a corresponding coloring assignment in V_3 and V_4 . Also it is easy to see that if i'' in V_3 is colored with $T_i(F_i)$ then the vertex i''' in V_4 has to be colored with $F_i(T_i)$. It can be verified that it is possible to color the node i'' with t_i or F_i if there is a valid harmonious coloring. This will give us a truth value assignment to the variables appearing in each clause.

This proves the theorem.

3. Conclusion. In this paper, we have defined the harmonious coloring problem for graphs and shown that this problem is NP-complete. The harmonious coloring

problem for directed graphs can also be shown to be NP-complete, using a similar construction. The theorem, we have proved is interesting, as the corresponding complexity results are not known for graceful and harmonious numbering of graphs. Finding efficient algorithms for the harmonious coloring problem on restricted classes of graphs merits further study.

Appendix (*added in proof*). After the acceptance of our paper, the authors were notified that David S. Johnson came up with a short proof of our main theorem. His proof is as follows:

It's by a transformation from INDEPENDENT SET. Suppose we want to know whether a graph $G = (V, E)$ has an independent set of size K . The corresponding instance of HARMONIOUS COLORING consists of two connected components G' and G'' . The first is G with three additional vertices u_1, u_2 , and u_3 , each joined to the other two and all vertices in V . Note that any harmonious coloring of G' must use $|V| + 3$ colors, one for each vertex. The second component is a clique on K vertices. The claim is that this two-component graph can be harmoniously colored with $|V| + 3$ colors if and only if G has an independent set of size K .

REFERENCES

- [1] A. AHO, J. HOPCROFT AND J. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.
- [2] R. CICHELLI, *Minimal perfect hash functions made simple*, Comm. ACM, 1 (1980), pp. 17–19. (Also see Technical Correspondences, Comm. ACM, 12 (1980), pp. 728–729; 5 (1981), p. 322.)
- [3] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability—A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.
- [4] S. W. GOLOMB, *How to number a graph*, in Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1972, pp. 23–37.
- [5] R. L. GRAHAM AND N. J. A. SLOANE, *On additive bases and harmonious graphs*, this Journal, 1 (1980), pp. 382–404.
- [6] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

CROSSING NUMBER IS NP-COMPLETE*

M. R. GAREY† AND D. S. JOHNSON†

Abstract. In this paper we consider a problem related to questions of optimal circuit layout: Given a graph or network, how can we embed it in a planar surface so as to minimize the number of edge-crossings? We show that this problem is NP-complete, and hence there is not likely to be any efficient way to design an *optimal* embedding.

A fundamental concept in graph theory is that of the *crossing number* $\nu(G)$ of a graph $G = (V, E)$. This is the least integer K such that G can be embedded in the plane so that there are no more than K pair-wise intersections of curves representing edges (not counting the required intersections at common endpoints). Recent work by Leighton [4] has shown that the crossing number of a graph can be used to obtain a lower bound on the amount of chip area required by that graph in a VLSI (very large scale integration) circuit layout, and the relevance of crossings to older technologies, such as printed circuits, has been discussed by Sinden [5].

There already exist efficient, linear-time algorithms for testing whether a graph has crossing number $\nu(G) = 0$, i.e., for testing whether a graph is planar [3]. In this paper we show that the general CROSSING NUMBER decision problem "Given G and an integer K is $\nu(G) \leq K$?" is NP-complete [1] and hence likely to be intractable. As a consequence, future research into crossing numbers will be justified in focusing on inexact methods that only *estimate* crossing numbers, and the quest for exact values of $\nu(G)$ will have to be restricted to promising special cases.

As defined, CROSSING NUMBER is in NP. One need only guess the K or fewer crossings (and the order in which they occur along edges involved in more than one crossing), create a new "crosspoint" vertex for each, replace each edge involved in one or more crossings by a path that contains all the crosspoint vertices associated with that edge in the appropriate order, and then test the resulting graph for planarity. Note that the above approach also allows us, for any fixed value of K , to test whether $\nu(G) \leq K$ in polynomial time (the degree of the polynomial depending on K).

To prove that CROSSING NUMBER is NP-complete, we must show that a known NP-complete problem can be transformed to it. Our "known" NP-complete problem will be OPTIMAL LINEAR ARRANGEMENT [2]: "Given a graph $G = (V, E)$ and an integer K , is there a one-to-one function $f: V \rightarrow \{1, 2, \dots, |V|\}$ such that

$$\sum_{\{u, v\} \in E} |f(u) - f(v)| \leq K?$$

We transform OPTIMAL LINEAR ARRANGEMENT to CROSSING NUMBER via an intermediate problem, which we shall call BIPARTITE CROSSING NUMBER: "Given a connected bipartite multigraph $G = (V_1, V_2, E)$ and an integer K , can G be embedded in a unit square so that all vertices of V_1 are on the northern boundary, all vertices in V_2 are on the southern boundary, all edges are within the square and there are at most K crossings?"

LEMMA 1. OPTIMAL LINEAR ARRANGEMENT \propto BIPARTITE CROSSING NUMBER.

* Received by the editors March 29, 1982, and in revised form August 20, 1982.

† Bell Laboratories, Murray Hill, New Jersey 07974.

Proof. Suppose we are given an instance $G = (V, E)$, K of OPTIMAL LINEAR ARRANGEMENT, where $V = \{\nu_1, \nu_2, \dots, \nu_n\}$. We may assume without loss of generality that G is connected. The corresponding instance of BIPARTITE CROSSING NUMBER is $G' = (V_1, V_2, E_1 \cup E_2)$, K' , where

$$\begin{aligned} V_1 &= \{u_i : 1 \leq i \leq n\}, \\ V_2 &= \{w_i : 1 \leq i \leq n\}, \\ E_1 &= \{|E|^2 \text{ copies of } \{u_i, w_i\} : 1 \leq i \leq n\}, \\ E_2 &= \{\{u_i, w_j\} : i < j \text{ and } \{\nu_i, \nu_j\} \in E\}, \\ K' &= |E|^2(K - |E|) + (|E|^2 - 1). \end{aligned}$$

Note that both G' and K' are constructible in polynomial time, given G and K . Note also that G' is connected because G is. We must show that the answer for G, K is yes if and only if the answer for G', K' is also yes.

Suppose first that the desired ordering function f exists for G . Then we can construct the following layout of G' . Suppose the corners of the unit square have coordinates $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$. We place each $u_i \in V_1$ at position $(1, f(\nu_i)/n)$ and each $w_i \in V_2$ at position $(0, f(\nu_i)/n)$, $1 \leq i \leq n$. We then embed the multiple edges joining pairs $\{u_i, w_i\}$ so that none cross, as in Fig. 1. Each edge $\{u_i, w_j\} \in E_2$ will then cross $(|f(\nu_j) - f(\nu_i)| - 1) \cdot |E|^2$ edges of E_1 and the total number of crossings of edges in E_2 with edges in E_1 will be at most

$$\sum_{\{u, \nu\} \in E} (|f(u) - f(\nu)| - 1) \cdot |E|^2 \leq (K - |E|) \cdot |E|^2.$$

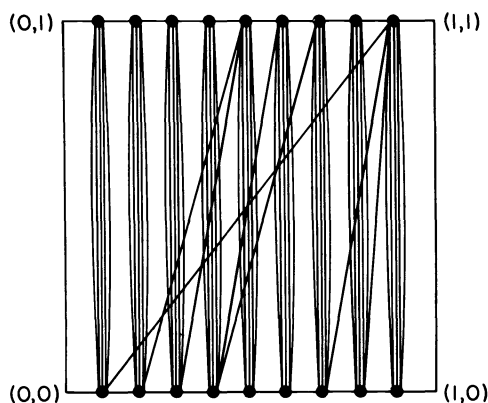


FIG. 1. Embedding for Lemma 1.

Since the total number of crossings between edges in E_2 is less than $(|E|^2 - 1)$, we conclude that the overall number of edge-crossings is at most K' .

Conversely, suppose the desired embedding of G' into the unit square exists. It naturally defines two one-to-one functions $f_1, f_2: V \rightarrow \{1, 2, \dots, |V|\}$ determined by the orderings of the vertices of V_1 and V_2 from left to right along their respective boundaries. These functions must be identical, since if $f_1(\nu_i) < f_1(\nu_j)$ and $f_2(\nu_i) > f_2(\nu_j)$, the embedding would contain at least $|E|^4$ crossings of edges $\{u_i, w_i\}$ with edges $\{u_j, w_j\}$, a contradiction of our bound on the number of crossings in the embedding. Thus the embedding looks like the one pictured in Fig. 1 and each edge $\{u_i, w_j\} \in E_2$ must be

involved in at least $(|f_1(v_i) - f_1(v_j)| - 1) \cdot |E|^2$ crossings. From this we conclude that

$$\sum_{\{u,v\} \in E} (|f_1(u) - f_1(v)| - 1) \cdot |E|^2 \leq K' = (K - |E|) \cdot |E|^2 + (|E|^2 - 1),$$

which implies that

$$\sum_{\{u,v\} \in E} (|f_1(u) - f_1(v)| - 1) \leq K - |E|,$$

and so f_1 will serve as the desired ordering for G . This completes the proof of the lemma. \square

LEMMA 2. BIPARTITE CROSSING NUMBER \propto CROSSING NUMBER.

Proof. We actually give a transformation to the version of CROSSING NUMBER where multigraphs are allowed. The final step to CROSSING NUMBER for graphs with no multiple edges allowed is obtained by simply adding a new degree-two vertex into the middle of each (multiple) edge, which eliminates the multiple edges without affecting the crossing number.

Suppose we are given an instance $G = (V_1, V_2, E), K$ of BIPARTITE CROSSING NUMBER. It is easy to construct the following multigraph $G' = (V', E \cup E_1 \cup E_2 \cup E_3)$ in polynomial time, where

$$\begin{aligned} V' &= V_1 \cup V_2 \cup \{u_0, w_0\}, \\ E_1 &= \{3K + 1 \text{ copies of } \{u_0, u\} : u \in V_1\}, \\ E_2 &= \{3K + 1 \text{ copies of } \{w_0, w\} : w \in V_2\}, \\ E_3 &= \{3K + 1 \text{ copies of } \{u_0, w_0\}\}. \end{aligned}$$

We claim that G has an embedding of the required form into the unit square (with K or fewer crossings) if and only if G' can be embedded in the plane with K or fewer crossings (the same K for both instances).

First, suppose the desired embedding of G into the unit square exists. Fig. 2 shows how the extra vertices and edges of G' can be added to the embedding (by being placed *outside* the unit square) with no increase in crossings.

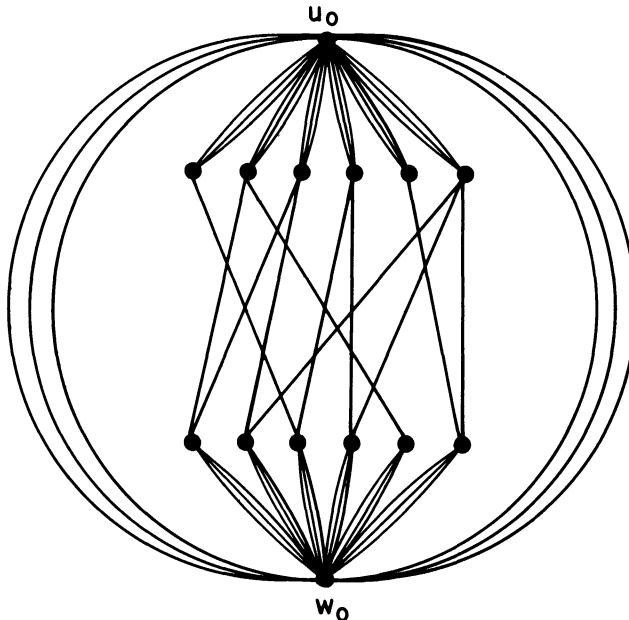


FIG. 2. Embedding for Lemma 2.

We now wish to argue that if the desired embedding of G' exists, there must be one whose form is just like that of Fig. 2. We proceed by a series of "normal form" simplifications.

Normalization 1. We may assume that each pair of edges crosses either 0 or 1 times and edges which share an endpoint do not cross at all. (This is easily proved using the transformation illustrated in Fig. 3, which always decreases the total number of crossings.) Thus each set of $3K + 1$ multiple edges can be viewed as creating an ordered sequence of $3K$ bounded regions.

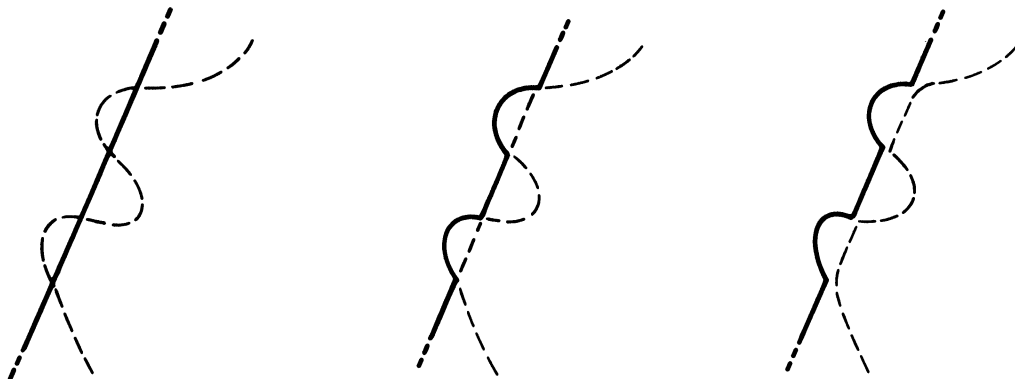


FIG. 3. Removing multiple crossings.

Normalization 2. The edges of E_1 divide the plane into a collection of regions, one of which is unbounded. By a standard transformation, we may assume that w_0 is inside (in the interior of) the unbounded region. Then, since each vertex in V_2 is connected by $3K + 1$ edges to w_0 , all these vertices must be inside the unbounded region too (if any such vertex were in a different region, it would introduce at least $3K + 1$ crossings, which is too many).

Normalization 3. We may assume that no vertex is inside any of the $3K$ regions formed by the edges (u_0, u) , for any fixed $u \in V_1$ and that no edge crosses any of these $3K + 1$ edges. We may also assume that the same properties hold for the $3K$ regions formed by the edges $\{w_0, w\}$, for any fixed $w \in V_2$. We shall prove this for the case of $\{u_0, u\}$; the other case follows analogously.

From Normalization 2 none of the $3K$ regions bounded by edges $\{u_0, u\}$ can contain a vertex from $V_2 \cup \{w_0\}$. Thus an interior vertex, if it exists, must be from V_1 . First let us make two observations about the middle K regions.

(a) No vertex from V_1 can be contained in any of the central K regions: such a vertex would have an edge to some vertex in V_2 since G is connected and that edge would have at least $K + 1$ crossings.

(b) No edge can cross any of the K middle regions: such an edge would have to cross *all* K regions if it crossed any, since by Normalization 1 it cannot double back, and by (a) its end-points must be at least K regions (and hence $K + 1$ boundary edges) apart.

Given (a) and (b), it follows that we can transform the embedding, by moving all edges joining u_0 and u into the interior of a single one of the middle regions, and no new crossing will be created. As a result, all vertices other than u_0 and u are left on the outside, and no edge will cross any of the boundaries.

Note that at this point we have obtained an embedding which is topologically equivalent to one like that in Fig. 2, except possibly for the edges in the original set E and the $3K + 1$ edges joining u_0 to w_0 .

Normalization 4. We may assume that all of the vertices in $V_1 \cup V_2$ and all of the edges in E are contained inside the same one of the $3K$ bounded regions formed by the edges joining u_0 to w_0 .

Let us begin our proof of this claim by numbering the bounded regions in order, R_1 through R_{3K} , with R_0 being the unbounded region. Suppose there is a vertex ν inside region R_I . Then there can be no vertex ν' in regions $R_{I+K+1(\bmod 3K+1)}$ through $R_{I+2K(\bmod 3K+1)}$. This is because there was a path in our original graph from ν to ν' , and this path would have to cross at least $K+1$ of the edges $\{u_0, w_0\}$ if ν' were in one of the prescribed regions. Consequently, as in claim (b) of the proof of Normalization 3, there can be no edges passing through any of these K regions. Thus, as in Normalization 3 we can move all the edges joining u_0 to w_0 into just one of these empty regions, without creating any new crossings. This leaves V_1 , V_2 and all of E in the single unbounded region. Now a simple transformation sends our embedding to one in which all of G is contained within the same *bounded region*.

Finalization. At this point we are done with the proof of Lemma 2, for the embedding created by our four normalizations is now topologically equivalent to one in the form of Fig. 2 and hence induces the desired embedding of G into the unit square. \square

The main theorem of this paper (and its title) follow as an immediate consequence of Lemmas 1 and 2.

Acknowledgment. The authors thank Gary Miller and Tom Leighton for suggestions that improved the presentation of this result.

REFERENCES

- [1] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.
- [2] M. R. GAREY, D. S. JOHNSON AND L. J. STOCKMEYER, *Some simplified NP-complete graph problems*, *Theor. Comput. Sci.*, 1 (1976), pp. 237–267.
- [3] J. E. HOPCROFT AND R. E. TARJAN, *Efficient planarity testing*, *J. Assoc. Comput. Mach.*, 21 (1974), pp. 549–568.
- [4] F. T. LEIGHTON, *New lower bound techniques for VLSI*, in Proc. 22nd Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Long Beach, CA, 1981, pp. 1–12.
- [5] F. W. SINDEN, *Topology of thin film circuits*, *Bell Syst. Tech. J.* XLV (1966), pp. 1639–1666.

WEIGHT ENUMERATORS OF NORMALIZED CODES II. THE HERMITIAN CASE*

STEPHEN M. GAGOLA, JR.†

Abstract. Let F be a finite field whose order is a square. A linear code C in F^n is self-dual if C coincides with its vector space dual with respect to the natural "Hermitian" form on F^n . If C contains the all-ones vector, then C is said to be normalized. The complete weight enumerator of a normalized self-dual code over F is invariant under the action of a linear group G which is explicitly determined. The character of this linear group is then used to calculate the Molien series.

Other conditions may be imposed on C which lead to its weight enumerator being invariant under the action of a larger linear group containing G . However, there are only finitely many finite linear groups containing G with the property that the only scalar matrices appearing are those already contained in G . In fact, if the characteristic of F is odd and if G^0 is the unimodular subgroup of G , then the finite unimodular subgroups containing G^0 are contained in a unique maximal such linear group.

1. Introduction. In [3] the author determined, for every field F , a finite linear group of complex $|F| \times |F|$ matrices which leaves invariant the weight enumerator of any normalized self-dual code over F . Duality, as defined in that paper, was with respect to the natural "dot-product" on F^n . However, if F is a quadratic extension of some subfield, then a natural "Hermitian" form exists on F^n . The present work represents the natural extension of the results in [3] to this "Hermitian case."

The terminology and notation used in [3, § 2], with some slight variations, will be assumed. General references for group theory are [5] and [8], while those for representation theory are [1] and [9], although [3, § 3] summarizes most of the necessary group theoretic results needed here.

The notation "Theorem I.4.3" will be used to refer to Theorem 4.3 of [3]. As in that paper, a code (linear subspace of F^n) is *normalized* if it contains the all-ones vector. The definition of duality as used here is slightly different, however.

Throughout this paper F denotes a finite field which is a quadratic extension of the field F_0 . Let $\bar{}$ denote the unique automorphism of F with fixed field F_0 , and write $|F| = q^2$ so that $|F_0| = q$. The symbol $\langle \cdot, \cdot \rangle$ will be used to denote the "standard Hermitian form on F^n ", that is,

$$\langle v, w \rangle = \sum v_i \bar{w}_i$$

for all v and w in F^n . For any code C contained in F^n , the *dual* of C is the subspace defined by

$$C^* = \{v \in F^n \mid \langle v, c \rangle = 0 \text{ for all } c \in C\}.$$

A code C is *self-dual* if $C = C^*$. Notice that $C^* = \bar{C}^\perp$ in the notation of [3].

If the characteristic of F is p , let $GF(p)$ denote the prime subfield of F and $\text{tr}: F \rightarrow GF(p)$ the usual trace map. As in [3], let $\mu_0: GF(p) \rightarrow \mathbb{C}^\times$ be the "standard character" given by $\mu_0(j) = \exp(2\pi i j/p)$ and set $\lambda = \mu_0 \circ \text{tr}$. For $a \in F$ define $\lambda_a: F \rightarrow \mathbb{C}^\times$ by $\lambda_a(x) = \lambda(ax)$ so that $\{\lambda_a \mid a \in F\}$ is the full set of irreducible characters of the group $(F, +)$. Notice that $\lambda_1 = \lambda$ and that λ_0 is the principal character of F .

It is convenient to record here the version of the MacWilliams identity which will be used:

* Received by the editors December 30, 1981, and in revised form July 16, 1982.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843, and Kent State University, Kent, Ohio 44242.

THEOREM 1.1 (MacWilliams). *Let C be a code in F^n , and for each $a \in F$ let M_a be the $F \times F$ matrix whose (r, s) entry is $(1/q)\lambda_a(r\bar{s})$. Then the complete weight enumerators of C and C^* are related by the equations*

$$W_{C^*} = W_C \cdot M_a$$

which hold for all $a \in F^\times$. In particular, if C is self-dual, then W_C is invariant under the action of the matrices M_a for $a \in F^\times$.

Notice that the definition of M_a differs from the definition in [3] to accommodate for the Hermitian form $\langle \cdot, \cdot \rangle$. The proof of Theorem 1.1 will be omitted.

For each $a \in F$ define $F \times F$ matrices N_a, D_a and E_a as follows:

$$(N_a)_{r,s} = \delta_{ra,ss}, \quad (D_a)_{r,s} = \delta_{r,s}\lambda_a(r\bar{r}), \quad (E_a)_{r,s} = \delta_{r,s}\lambda_a(r).$$

The matrices N_a and E_a coincide with their earlier definitions in [3], but D_a is slightly different, again to accommodate the Hermitian form.

The relation $\sum c_i \bar{c}_i = 0$ holds for any vector c of a self-dual code C , and this leads easily to $W_C \cdot D_a = W_C$ for all $a \in F$. Moreover, since C is also a subspace of F^n , we have $W_C \cdot N_a = W_C$ for all $a \in F^\times$. Hence, W_C is an invariant of the linear group

$$G_0 = \langle M_a, N_a, D_b \mid a, b \in F, a \neq 0 \rangle.$$

If in addition C contains the all-ones vector, then $\sum c_i = 0$ holds for all c in C . This leads easily to $W_C \cdot E_a = W_C$ for every $a \in F$, and W_C is invariant under the larger group:

$$G = \langle M_a, N_a, D_b, E_b \mid a, b \in F, a \neq 0 \rangle.$$

LEMMA 1.2. *For every $a, b, c, d \in F$ with $a \neq 0$ and $b \neq 0$, we have*

$$\begin{aligned} M_a M_b &= N_{-a\bar{b}^{-1}}, & M_a^{-1} N_b M_a &= N_{\bar{b}^{-1}}, \\ N_a N_b &= N_{ab}, & N_a^{-1} D_c N_a &= D_{c(a\bar{a})^{-1}}, \\ D_c D_d &= D_{c+d}, & N_a^{-1} E_c N_a &= E_{ca^{-1}}, \\ E_c E_d &= E_{c+d}. \end{aligned}$$

Moreover, $D_c = I$ if and only if $c + \bar{c} = 0$.

Proof. These relations are similar to those appearing in Lemma I.2.3, and the proof for all but the last assertion is omitted.

The equation $D_c = I$ is equivalent to $\lambda_c(r\bar{r}) = 1$ for every $r \in F$. Since the norm map $F \rightarrow F_0$ is surjective, this is equivalent to $\lambda_c(F_0) = \{1\}$. Hence $D_c = I$ if and only if $\{1\} = \lambda_c(F_0) = \lambda(cF_0)$, or $cF_0 \subseteq \ker \lambda$. Now let $K = \{x \in F \mid x + \bar{x} = 0\}$ be the kernel of the trace map from F to F_0 . Clearly $K = \ker(\text{tr}) \subseteq \ker \lambda$. Now $\ker \lambda < F$ and since $\dim_{F_0} F = 2$, $\ker \lambda$ can contain at most one nontrivial F_0 -subspace of F . Hence $D_c = I$ if and only if $cF_0 = \{0\}$ or $cF_0 = K$ (that is, if and only if $c \in K$), as desired. \square

The groups G_0 and G are determined abstractly in §§ 2 and 3, and the Molien series (in unsimplified form) for the given matrix representations of them are worked out in §§ 4 and 5.

It is possible to impose conditions on the weight enumerator of a normalized code C , which implies that it is invariant under the action of a larger matrix group containing G . For example, if C is setwise invariant under the action of the Galois group of F over some subfield, then this larger matrix group is obtained by forming the semidirect product of G with the Galois group.

The case of $GF(4)$ illustrates this nicely. If C is a self-dual code (in the non-Hermitian sense) over $GF(4)$ in which all weights are even, then C is also self-dual

in the Hermitian sense, and C is stabilized under the action of $\text{Gal}(GF(4)/GF(2))$ (see [10, § 2]). If C is also normalized, then the complete weight enumerator of C is invariant under the internal semidirect product $G \rtimes \langle A \rangle$ where $\langle A \rangle \cong \text{Gal}(GF(4)/GF(2))$ is cyclic of order 2 and is generated by the permutation matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

which corresponds to the permutation of the elements of $GF(4)$ by the nontrivial automorphism. In § 3, G is shown to be an extension of an extraspecial group of order $2 \cdot 4^2$ by $U(2, GF(4))$ which has order 18. Hence, the order of $G \rtimes \langle A \rangle$ is $2|G| = 1152$. This group appears in the proof of [10, Thm. 17], where it is denoted by G and is defined as the matrix group generated by the matrices

$$AM_\alpha, D_\alpha, M_1E_1M_1, N_\alpha$$

where $GF(4) = \{0, 1, \alpha, \alpha^2 = \alpha + 1\}$.

The full ring of polynomial invariants is a free ring with 4 generators $R_0 = \mathbb{C}[f_2, f_6, f_8, f_{12}]$ (using the notation of [10]). Clearly, this is contained in the ring of invariants for G , say R . In fact, by consideration of the Molien series for G given in the appendix, if a homogeneous polynomial Δ of degree 12 can be found which is invariant under G , but not under $G \rtimes \langle A \rangle$, then

$$R = R_0 + R_0 \cdot \Delta.$$

Such a polynomial is easily found, and

$$\Delta = (X_0^2 - X_1^2)(X_0^2 - X_\alpha^2)(X_0^2 - X_{\alpha+1}^2)(X_1^2 - X_\alpha^2)(X_1^2 - X_{\alpha+1}^2)(X_\alpha^2 - X_{\alpha+1}^2)$$

is one such. Clearly, Δ is invariant under any diagonal matrix whose entries are ± 1 , as well as any permutation matrix corresponding to an even permutation. Hence $\Delta \cdot H = \Delta$ where H is D_a, E_a , or N_a (for $a \neq 0$ in the last case). Moreover, $(X_0 + X_1), (X_0 - X_1), \dots$ (the linear factors of Δ) are permuted by M_1 without sign changes, so Δ is invariant under M_1 , as well as $M_a = N_{-a}M_1^{-1}$. Clearly, $\Delta \cdot A = -\Delta \neq \Delta$, so Δ is invariant under G , but not $G \rtimes \langle A \rangle$.

The general question of calculating rings of invariants for arbitrary fields is not considered here, however.

Finite linear groups containing G in the odd characteristic case are discussed in § 6. For a justification for considering only *finite* subgroups of $GL(|F|, \mathbb{C})$, see the remarks following [3, Thm. 2.1].

The Molien series for G_0 and G are given in simplified form for $|F| = 4, 9$ and 16 in the appendix.

The author would like to take this opportunity to correct an error which occurs on [3, p. 365] after Lemma 6.5. There it is asserted that $f_+(x)$ is the first polynomial stated in the lemma if $p \equiv 1 \pmod 4$, while it is the second polynomial if $p \equiv 3 \pmod 4$. In these congruences, p should be replaced by p^k . The author is indebted to the current referee for finding this error.

2. Construction of G (odd characteristic). Throughout this section, the characteristic of F is an odd prime p . Since F has cardinality q^2 , we may write $F = GF(q^2)$. Recall that $\bar{\cdot}$ denotes the unique automorphism of F having order 2. (Thus $\bar{\alpha} = \alpha^q$ for all $\alpha \in F$.)

Let V denote the F -space $\{(a, b) | a, b \in F\}$, and define the group E as follows. As a set, E is $V \times GF(p)$. If (u, m) and (v, n) are in E , then the product $(u, m)(v, n)$ is defined to be $(u + v, m + n + \text{tr}(u_1\bar{v}_2 - u_2\bar{v}_1))$. Here, $u = (u_1, u_2)$, $v = (v_1, v_2)$, and as defined in the previous section, $\text{tr}: F \rightarrow GF(p)$ is the trace map. As is readily checked, E is a group under this operation, and is extraspecial of exponent p and order q^4p . Occasionally it is convenient to identify E as $F \times F \times GF(p)$.

Let $U(2, F)$ denote the full unitary group in dimension 2 over F defined with respect to the (skew-Hermitian) form whose matrix is given by $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Thus:

$$U(2, F) = \left\{ g \in GF(2, F) \mid g \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \bar{g}^T = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right\}.$$

If $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is replaced by $c \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ where $c \neq 0$ satisfies $c + \bar{c} = 0$, then $c \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is a Hermitian matrix and the definition of $U(2, F)$ is unaffected. This justifies using the unitary group notation. Properties of the unitary group $U(2, F)$ may be found in [8, Chap. II, § 10].

Define subgroups H and P of $U(2, F)$ by setting

$$H = \left\{ \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \mid a \in F^\times \right\}, \quad P = \left\{ \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \mid a \in F_0 \right\}.$$

It is easy to check that if $g \in GL(2, F)$ has all its entries in F_0 , then the condition $g \in U(2, F)$ is equivalent to $\det g = 1$. Hence $SL(2, F_0) \leq U(2, F)$, and since $SL(2, F_0)$ is isomorphic to $SU(2, F)$ ([8, p. 194]), we have $SL(2, F_0) = SU(2, F)$.

If $g \in U(2, F)$, then $\det g$ is in the kernel of the norm map $F^\times \rightarrow F_0^\times$. By Hilbert's Theorem 90 we may write $\det g = a\bar{a}^{-1}$ for some $a \in F^\times$. Now

$$h = \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \in H$$

and $h^{-1}g = s \in SU(2, F) = SL(2, F_0)$. Hence $g = hs$, and this proves the factorization $U(2, F) = H \cdot SL(2, F_0)$. The index of $SL(2, F_0)$ in $U(2, F)$ is $q + 1$, and so the order of $U(2, F)$ is $(q + 1)^2q(q - 1)$.

Define an action of $U(2, F)$ on E as follows. For $g \in U(2, F)$ and $e = (v, n) \in E$, set $e^g = (vg, n)$. It is easy to verify that this is indeed an action. Moreover, since the multiplication within E may be written as

$$(u, m)(v, n) = \left(u + v, m + n + \text{tr} \left(u \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \bar{v}^T \right) \right),$$

the action is by automorphisms. Clearly, the induced action of $U(2, F)$ on $E/Z(E)$ may be identified with the action of $U(2, F)$ on V . Let \tilde{G} denote the semidirect product $U(2, F) \ltimes E$ with respect to the action defined above. Thus, as a set \tilde{G} is $U(2, F) \times E$ with multiplication defined by

$$(g, e)(h, f) = (gh, e^h f).$$

As usual, $U(2, F)$ and E will be identified as subgroups of \tilde{G} .

Various subgroups of E (and hence of \tilde{G}) are defined as follows:

$$\begin{aligned} A_0 &= \{0\} \times F \times \{0\} \leq E, & A &= A_0 \times \mathbb{Z} \leq E, \\ \mathbb{Z} &= \{0\} \times \{0\} \times GF(p) \leq E, & T &= F \times \{0\} \times \{0\} \leq E. \end{aligned}$$

Notice that H normalizes P and that the group PH normalizes A_0 so that PHA_0 is also a group. Since \mathbb{Z} is centralized by this last group, $PHA = PHA_0\mathbb{Z} = PHA_0 \times \mathbb{Z}$

is also a group. Note that T is a transversal for PHA in PHE (as well as for A in E).

Recall that μ_0 is the “standard character” of $GF(p) \simeq \mathbb{Z}$, and so μ_0 will be regarded as a character of \mathbb{Z} . Define the representation ρ_1 of PHE by

$$\rho_1 = (1_{PHA_0} \# \mu_0)^{PHE}$$

where the transversal T is used in the construction of the induced representation. For a discussion of induced representations, see [3, § 3]. Since T is a group naturally isomorphic to F , $\rho_1(g)$ will be viewed as an $F \times F$ matrix for each $g \in PHE$.

THEOREM 2.1. *The representations $\rho_1, \rho_1|_{PE}$ and $\rho_1|_E$ are all faithful and irreducible. Moreover,*

(1) *If $e = (a, b, m) \in E$, then the (r, s) entry of $\rho_1(e)$ is $\delta_{r+a, s} \mu_0(m) \lambda_{2b}(\bar{r} + \bar{a}/2) = \delta_{r+a, s} \mu_0(m) \lambda_{2\bar{b}}(r + a/2)$. In particular, for $e = (0, b, 0) \in A_0$ we have $\rho_1(e) = E_{2\bar{b}}$.*

(2) *If $x = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \in P$ (so that $a \in F_0$), then $\rho_1(x) = D_a$.*

(3) *If $h = \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \in H$, then $\rho_1(h) = N_a$.*

The proof of this last result parallels closely the proof of the corresponding result in [3], and will be omitted.

We have already observed that $U(2, F)$ contains $SL(2, F_0)$ as a normal subgroup whose quotient is cyclic of order $q + 1$. Thus the group $U(2, F)$ has a unique sign character with kernel containing $SL(2, F_0)$. Since $U(2, F)$ is a homomorphic image of $\tilde{G} = U(2, F) \times E$, \tilde{G} has a unique sign character with kernel containing $SL(2, F_0) \times E$. Denote this character by δ . Since H supplements $SL(2, F_0)$ and H is cyclic ($H \simeq F^\times$), we have $\delta(h) = 1$ if and only if $h \in H$ is a square. Hence δ corresponds to the “Legendre symbol of F^\times .” It follows easily from the definition of N_a that

$$\delta \left(\begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \right) = \det N_a.$$

THEOREM 2.2. *The representation ρ_1 is uniquely extendible to a representation of \tilde{G} . Moreover, the determinant of this extension is δ .*

Proof. As in the proof of Theorem I.4.3, the representation $\rho_1|_E$ is extendible to \tilde{G} . Let ρ be an extension of $\rho_1|_E$ to \tilde{G} . Since $\rho|_{PHE}$ and ρ_1 are both extensions of $\rho_1|_E$, we have $\rho|_{PHE} = \lambda \rho_1$ for some linear character λ of PHE with $E \leq \ker \lambda$. We first prove that λ is extendible to \tilde{G} .

The commutator subgroup of PHE is PE , so $PE \leq \ker \lambda$, and hence $\lambda^{q^2-1} = 1$. Moreover,

$$(\det \rho)_{PHE} = \det (\rho|_{PHE}) = \lambda^{q^2} \det \rho_1 = \lambda \det \rho_1.$$

Now $(\det \rho_1)(x) = \delta(x)$ for $x \in PH$ is easily checked (using Theorem 2.1), so $\det \rho_1 = \delta|_{PHE}$. Set $\mu = \delta \cdot \det \rho$. Then $\mu_{PHE} = \delta_{PHE} \cdot \lambda \cdot \det \rho_1 = \lambda$, proving that μ is an extension of λ to \tilde{G} .

Hence $(\bar{\mu}\rho)_{PHE} = \bar{\lambda}\rho|_{PHE} = \bar{\lambda}\lambda\rho_1 = \rho_1$ so that $\bar{\mu}\rho$ is an extension of ρ_1 to \tilde{G} . Replace ρ by $\bar{\mu}\rho$ if necessary so as to assume ρ is an extension of ρ_1 . If $\tilde{\rho}$ is another representation extending ρ_1 , then $\tilde{\rho} = \theta\rho$ for some linear character θ of \tilde{G} . Since $\tilde{\rho}|_{PHE} = \rho|_{PHE} = \rho_1$, we have $\theta_{PHE} = 1_{PHE}$ so that $PHE \leq \ker \theta$. But $\tilde{G}' \leq \ker \theta$ and $\tilde{G} = \tilde{G}' \cdot H$ so $\theta = 1_{\tilde{G}}$. Hence $\tilde{\rho} = \rho$ is the unique extension of ρ_1 to \tilde{G} .

Finally, $\det \rho_1 = \delta_{PHE}$ so $\det \rho|_{PHE} = \delta_{PHE}$. Since H supplements \tilde{G}' in \tilde{G} , every linear character of \tilde{G} is uniquely determined by its restriction to H . The equation $\det \rho = \delta$ now follows, completing the proof of Theorem 2.2. \square

THEOREM 2.3. *Let ρ be the unique extension of ρ_1 to \tilde{G} . Then*

$$\rho\left(\begin{matrix} 0 & a/2 \\ -2/\bar{a} & 0 \end{matrix}\right) = -M_a \quad \text{for all } a \in F^\times.$$

Proof. By a direct computation (similar to the one appearing in the proof of Theorem I.4.4) we have that

$$M_1^{-1}\rho(e)M_1 = \rho\left(\left(\begin{matrix} 0 & 1/2 \\ -2 & 0 \end{matrix}\right)^{-1} e\left(\begin{matrix} 0 & 1/2 \\ -2 & 0 \end{matrix}\right)\right)$$

holds for all $e \in E$. Since $\rho|_E$ is irreducible, Schur's lemma implies that $\rho\left(\begin{smallmatrix} 0 & 1/2 \\ -2 & 0 \end{smallmatrix}\right)$ is a scalar multiple of M_1 . Write $\rho\left(\begin{smallmatrix} 0 & 1/2 \\ -2 & 0 \end{smallmatrix}\right) = cM_1$. Squaring this equation yields $\rho\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) = c^2M_1^2 = c^2N_{-1}$. But $\rho\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) = N_{-1}$ by Theorem 2.1 (3), so $c^2 = 1$ and $c = \pm 1$. Notice that $\left(\begin{smallmatrix} 0 & 1/2 \\ -2 & 0 \end{smallmatrix}\right) \in SL(2, F_0) = U(2, F)$ so that $cM_1 = \rho\left(\begin{smallmatrix} 0 & 1/2 \\ -2 & 0 \end{smallmatrix}\right)$ has determinant 1. Hence $1 = \det(cM_1) = c^{q^2} \det M_1 = c \det M_1$, and therefore $c = \det M_1$.

Now $M_1^4 = I$, so the eigenvalues of M_1 are 1, -1 , i and $-i$, occurring with multiplicities α, β, γ and δ , say. Hence $c = \det M_1 = (-1)^\beta i^{\gamma-\delta}$. To evaluate c , then, it suffices to evaluate β, γ and δ . Now the (r, r) entry of M_1 is $(1/q)\lambda_1(r\bar{r}) = \mu_0(\text{tr}(r\bar{r}))/q$ and hence

$$\text{trace}(M_1) = (1/q) + (1/q) \sum_{r \in F^\times} \mu_0(\text{tr}(r\bar{r})).$$

Let tr^0 denote the trace map from F_0 to $GF(p)$. Then

$$\begin{aligned} \text{trace}(M_1) &= (1/q) + ((q+1)/q) \sum_{s \in F_0^\times} \mu_0(\text{tr}^0(s)) \\ &= (1/q) + ((q+1)/q) \sum_{s \in F_0} \mu_0(\text{tr}^0(s)) - ((q+1)/q) \\ &= -1. \end{aligned}$$

Hence $\alpha - \beta + \gamma i - \delta i = -1$, and this leads to $\alpha - \beta = -1$ and $\gamma = \delta$. Moreover, $M_1^2 = N_{-1}$ has trace 1, so $\alpha + \beta - \gamma - \delta = 1$. Clearly $\alpha + \beta + \gamma + \delta = q^2$. The unique solution to this system is $(\alpha, \beta, \gamma, \delta) = ((q^2 - 1)/4, (q^2 + 3)/4, (q^2 - 1)/4, (q^2 - 1)/4)$. Since q is odd, so is $\beta = (q^2 + 3)/4$, and hence $c = (-1)^\beta i^{\gamma-\delta} = -1$.

Finally, for $a \neq 0$,

$$\rho\left(\begin{matrix} 0 & a/2 \\ -2/\bar{a} & 0 \end{matrix}\right) = \rho\left(\begin{matrix} a & 0 \\ 0 & \bar{a}^{-1} \end{matrix}\right)\rho\left(\begin{matrix} 0 & 1/2 \\ -2 & 0 \end{matrix}\right) = -N_a M_1 = -M_a,$$

as desired. \square

The last result implies that M_a does not belong to $\rho(\tilde{G})$ for any $a \in F^\times$. Otherwise, since $-M_a \in \rho(\tilde{G})$ we would have $-I \in \rho(\tilde{G})$. However, since $U(2, F)$ acts faithfully on E , the center of \tilde{G} is contained in E (and so must coincide with $\mathbb{Z}(E)$). Since \tilde{G} is isomorphic to $\rho(\tilde{G})$, the statement $-I \in \rho(\tilde{G})$ is equivalent to $-I \in \rho(\mathbb{Z}(E))$, which is clearly false since $|\mathbb{Z}(E)|$ is odd.

Since the matrices M_a leave invariant the complete weight enumerator of a self-dual code, it is convenient to extend the group \tilde{G} and the representation ρ as follows. Let $\langle -1 \rangle$ be the cyclic group of order 2 generated by -1 , and set $G = \langle -1 \rangle \times \tilde{G}$. The representation ρ of \tilde{G} extends naturally to G by defining $\rho(s, g) = s\rho(g)$ for $s \in \langle -1 \rangle$ and $g \in \tilde{G}$. It is readily checked that ρ is faithful on G and that $M_a \in \rho(G)$.

It is also convenient to define G_0 as $\langle -1 \rangle \times U(2, F)$. Hence G_0 is a subgroup of G and $M_a \in \rho(G_0)$.

The next theorem establishes that the matrix groups defined in § 1 are matrix representations of the abstract groups determined in this section.

THEOREM 2.4. $\rho(G_0) = \langle M_a, N_a, D_b | a, b \in F, a \neq 0 \rangle$ and

$$\rho(G) = \langle M_a, N_a, D_b, E_b | a, b \in F, a \neq 0 \rangle.$$

Proof. Let $X = \langle M_a, N_a, D_b | a, b \in F, a \neq 0 \rangle$. Combining all of the previous theorems we have

$$M_a = -\rho \begin{pmatrix} 0 & a/2 \\ -2/\bar{a} & 0 \end{pmatrix} = \rho \left(-1, \begin{pmatrix} 0 & a/2 \\ -2/\bar{a} & 0 \end{pmatrix} \right),$$

$$N_a = \rho \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix},$$

$$D_b = \rho \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \quad (b \in F_0).$$

Moreover, since the characteristic is not two, we have $F = F_0 \dot{+} \ker(\tau)$ where $\tau : F \rightarrow F_0$ is the trace map. If $b \in F$, write $b = r + s$ where $r \in F_0$ and $s \in \ker \tau$. Then $D_b = D_r$, and this proves X is contained in $\rho(G_0)$. Now

$$U(2, F) = PH \cup PH \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} P,$$

so that the matrices

$$\rho \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \rho \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix}, \quad \rho \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$$

for $a \in F^\times$ and $b \in F_0$ generate $\rho(U(2, F)) \cong \rho(G_0)$. Hence $\rho(U(2, F)) \cong X \cong \rho(G_0)$. Now

$$-M_1 = \rho \begin{pmatrix} 0 & 1/2 \\ -2 & 0 \end{pmatrix} \in \rho(U(2, F)) \cong X,$$

and also $M_1 \in X$, so $-I \in X$. We already observed that $-I$ does not belong to $\rho(\tilde{G})$ and hence does not belong to $\rho(U(2, F))$. This forces $\rho(U(2, F)) < X$ and hence $X = \rho(G_0)$.

Now let $Y = \langle M_a, N_a, D_b, E_b | a, b \in F, a \neq 0 \rangle$. By Theorem 2.1, $E_b = \rho(0, \bar{b}/2, 0) \in \rho(\tilde{G}) \cong \rho(G)$, so $Y \cong \rho(G)$ follows from the first part of the proof. Clearly $X \cong Y$, so $\rho(G_0) \cong Y$.

Since the action of $U(2, F)$ on $E/\mathbb{Z}(E) \cong V$ is irreducible, and $E' = \mathbb{Z}(E) = \mathbb{Z}$, the only subgroups of $\rho(G)$ containing $\rho(G_0) = X$ are $\rho(G_0)$, $\rho(G_0\mathbb{Z})$ and $\rho(G)$. For $b \neq 0$, $E_b \notin \rho(G_0\mathbb{Z})$, and it follows that $Y = \rho(G)$, as required. \square

It will now be convenient to identify the character of ρ restricted to various subgroups of G_0 and G .

Fix $\nu \in F$ to be any nonzero element in the kernel of the trace map $F \rightarrow F_0$. Thus $\nu + \bar{\nu} = 0$. If $r \in F$ has norm 1 ($r\bar{r} = 1$), then define

$$M(r) = \begin{pmatrix} (r+1)/2 & -(r-1)/2\nu \\ -\nu(r-1)/2 & (r+1)/2 \end{pmatrix}.$$

It is readily checked that $M(r) \in U(2, F)$ and that $M(rs) = M(r)M(s)$. By direct calculation $\det M(r) = r$, and so if C is the group $\{M(r) | r\bar{r} = 1\}$, then C is a complement

for $SU(2, F)$ in $U(2, F)$. Clearly, C is isomorphic to $N = \{r \in F^\times | r\bar{r} = 1\}$, the kernel of the norm map $F^\times \rightarrow F_0^\times$. Define

$$D = \left\{ \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \mid s\bar{s} = 1 \right\}$$

so that D is a subgroup of $U = U(2, F)$ isomorphic to N as well. Since $C \cap D = \{1\}$ and D is the center of U , CD is isomorphic to $C \times D$. Let

$$\tilde{C} = \left\{ M(r) \begin{pmatrix} r^{-1} & 0 \\ 0 & r^{-1} \end{pmatrix} \mid r \in N \right\}.$$

Then \tilde{C} is another complement for $SU(2, F)$ in U and $\tilde{C} \cap C = \{1\}$. Obviously $CD = C\tilde{C}$.

For any group L , let reg_L denote the regular character of L . If $L_0 \trianglelefteq L$ then reg_{L/L_0} may be viewed as a character of L with kernel containing L_0 . We have already used the notation 1_L to denote the principal character.

THEOREM 2.5. *Let ψ denote the character of ρ . Then*

- (a) $\psi_H = 1_H + \text{reg}_H$,
- (b) $\psi_{PD} = 1_{PD} + \text{reg}_{PD} - \text{reg}_{PD/P}$,
- (c) $\psi_{CD} = \psi_{C\tilde{C}} = \text{reg}_{C\tilde{C}} - \text{reg}_{C\tilde{C}/C} - \text{reg}_{C\tilde{C}/\tilde{C}} + 1_{C\tilde{C}}$.

Proof. Identify $(1_{PHA_0} \# \mu_0)^{PHE}$ as the character of ρ_1 (rather than ρ_1 itself), we have $\psi_{PHE} = (1_{PHA_0} \# \mu_0)^{PHE}$. The group PHE acts doubly transitively on the cosets of PHA , so by the Mackey decomposition we have

$$\psi_{PHA} = (1_{PHA_0} \# \mu_0)^{PHE} |_{PHA} = 1_{PHA_0} \# \mu_0 + ((1_{PHA_0} \# \mu_0)^x |_{PA})^{PHA}.$$

Here x is any element of $PHE - PHA$ and we have used $PA = PHA \cap (PHA)^x$. Now $PA \trianglelefteq PHE$, so

$$(1_{PHA_0} \# \mu_0)^x |_{PA} = ((1_{PHA_0} \# \mu_0)_{PA})^x = (1_{PA_0} \# \mu_0)^x = \mu \# \mu_0$$

for some linear character μ of PA_0 . The particular form of μ is not important (μ depends on x), although μ satisfies $\mu_P \neq 1_P$. Therefore

$$\psi_{PHA} = 1_{PHA_0} \# \mu_0 + (\mu \# \mu_0)^{PHA} = (1_{PHA_0} + \mu^{PHA_0}) \# \mu_0$$

so that

$$\psi_{PHA_0} = 1_{PHA_0} + \mu^{PHA_0}.$$

As $(PA_0)(PH) = PHA_0$ and $PA_0 \cap PH = P$, we have $\mu^{PHA_0} |_{PH} = (\mu_P)^{PH}$, so

$$\psi_{PH} = 1_{PH} + (\mu_P)^{PH}.$$

Now (a) follows by restricting this equation to H (as $(\mu_P)^{PH} |_H = (1_1)^H = \text{reg}_H$). Restricting to PD yields

$$\psi_{PD} = 1_{PD} + (\mu_P)^{PH} |_{PD} = 1_{PD} + \sum_y ((\mu_P)^y |_{P^y \cap PD})^{PD},$$

where the sum extends over (double) coset representatives for (P, PD) in PH . However, $P^y \cap PD = P$, and the double cosets for (P, PD) in PH are the cosets of PD . Moreover,

$PD \simeq P \times D$ is the inertia group for $\mu_P \neq 1_P$ in PH , and PH/PD acts transitively on the nonprincipal characters of P . Therefore

$$\begin{aligned} \psi_{PD} &= 1_{PD} + \sum_y ((\mu_P)^y)^{PD} = 1_{PD} + \sum_y (\mu_P)^y \neq \text{reg}_D \\ &= 1_{PD} + \left(\sum_y (\mu_P)^y \right) \neq \text{reg}_D \\ &= 1_{PD} + (\text{reg}_P - 1_P) \neq \text{reg}_D \\ &= 1_{PD} + \text{reg}_{PD} - \text{reg}_{PD/P}, \end{aligned}$$

proving (b).

To prove (c), notice that

$$x = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \in C$$

if and only if $s = 1$, and $x \in \tilde{C}$ if and only if $rs = 1$. Thus (c) is equivalent to the assertion that

$$\psi(x) = \begin{cases} 1 & \text{if } s \neq 1 \text{ and } rs \neq 1, \\ -q & \text{if } s = 1 \text{ or } rs = 1, \text{ but not both,} \\ q^2 & \text{if } r = s = 1. \end{cases}$$

This will be shown directly by calculating $\psi(x)$. For $r = 1$ we have

$$x = \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \in H.$$

From part (a), then, $\psi(x) = 1$ if $s \neq 1$ while $\psi(x) = q^2$ for $s = 1$. Therefore, we may assume $r \neq 1$, and hence $M(r) \notin PH$. From the Bruhat decomposition of $U = PH \cup HP\omega P$, we have

$$M(r) = \begin{pmatrix} b & 0 \\ 0 & \bar{b}^{-1} \end{pmatrix} \begin{pmatrix} 1 & c_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & c_2 \\ 0 & 1 \end{pmatrix}$$

for some $b \in F^\times$ and $c_1, c_2 \in F_0$. Solving for b, c_1 and c_2 yields

$$b = \frac{2}{\nu} \frac{r}{r-1}, \quad c_1 = \frac{\nu}{4}(\bar{r}-r), \quad c_2 = -\frac{1}{\nu} \frac{r+1}{r-1},$$

where we have used $\bar{\nu} = -\nu$ and $r\bar{r} = 1$. Now $\psi(x)$ is the trace of $\rho(x)$ where (using Theorems 2.1 and 2.3)

$$\begin{aligned} \rho(x) &= \rho \begin{pmatrix} b & 0 \\ 0 & \bar{b}^{-1} \end{pmatrix} \rho \begin{pmatrix} 1 & c_1 \\ 0 & 1 \end{pmatrix} \rho \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \rho \begin{pmatrix} 1 & c_2 \\ 0 & 1 \end{pmatrix} \rho \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \\ &= N_b D_{c_1} (-M_2) D_{c_2} N_s. \end{aligned}$$

Hence

$$\begin{aligned} \psi(x) &= \sum_{t,u,v \in F} \delta_{tb, u} \lambda_{c_1}(u\bar{u}) \left(-\frac{1}{q} \lambda_2(u\bar{v}) \right) \lambda_{c_2}(v\bar{v}) \delta_{vs,t} \\ &= (-1/q) \sum_{v \in F} \lambda_{c_1}(vsb\bar{v}\bar{s}\bar{b}) \lambda_2(vsb\bar{v}) \lambda_{c_2}(v\bar{v}) \\ &= (-1/q) \sum_{v \in F} \lambda(c_1 b \bar{b} + 2sb + c_2) v \bar{v} \end{aligned}$$

where we have used $s\bar{s} = 1$. Let $k = c_1b\bar{b} + 2s\bar{b} + c_2$. Then

$$\begin{aligned} \psi(x) &= (-1/q) \sum_{v \in F} \lambda_k(v\bar{v}) \\ &= (-1/q) - (q+1)/q \sum_{v \in F_0} \lambda_k(v) \\ &= -\frac{1}{q} + \frac{q+1}{q} - \frac{q+1}{q} \sum_{v \in F_0} \lambda_k(v) \\ &= 1 - (q+1) \cdot (\lambda_k|_{F_0}, 1_{F_0}) \end{aligned}$$

where $(\lambda_k|_{F_0}, 1_{F_0})$ denotes an inner product of characters. This last inner product is 1 if $\lambda_k|_{F_0} = 1_{F_0}$ and is zero otherwise.

Now $\lambda_k|_{F_0} = 1_{F_0}$ if and only if kF_0 is in the kernel of λ . Obviously, kF_0 is an F_0 -subspace of F . let $K = \{t \in F | t + \bar{t} = 0\}$, so that K is an F_0 -subspace of F and is the kernel of the trace map $F \rightarrow F_0$. Clearly $K \subseteq \ker \lambda$. Since $\dim_{F_0} F = 2$ and $\ker \lambda < F$, $\ker \lambda$ can contain at most one F_0 -subspace of F . This shows that $\lambda_k|_{F_0} = 1_{F_0}$ if and only if $kF_0 \subseteq K$ (or $k + \bar{k} = 0$).

When $k + \bar{k}$ is written in terms of ν, r and s and the expression is simplified (using $\bar{\nu} = -\nu, r\bar{r} = 1$ and $s\bar{s} = 1$), we have

$$k + \bar{k} = \frac{4\bar{s}r}{\nu(r-1)}(s^2 - (1 + \bar{r})s + \bar{r}).$$

Thus, $k + \bar{k} = 0$ if and only if $s = 1$ or $s = \bar{r}$ (i.e. $rs = 1$). Hence for $s = 1$ or $rs = 1$, $(\lambda_k|_{F_0}, 1_{F_0}) = 1$ and $\psi(x) = 1 - (q+1) \cdot 1 = -q$, while for $s \neq 1$ and $rs \neq 1$, $(\lambda_k|_{F_0}, 1_{F_0}) = 0$ and $\psi(x) = 1 - 0 = 1$, as desired. \square

Recall that $\nu \in F^\times$ is a fixed element in the kernel of the trace map $F \rightarrow F_0$. Define a subgroup E_0 of E by setting $E_0 = \{(s\nu, s, n) | s \in F, n \in GF(p)\}$. It is readily checked that E_0 is a subgroup of E , and that E itself is the central product of E_0 with $C_E(E_0)$ over $Z(E)$ (equivalently, E_0 is extraspecial). Moreover, since $(s\nu, s)$ is an eigenvector with eigenvalue 1 for each matrix in the group C , it follows that C centralizes E_0 in the group G . Hence $CE_0 \cong C \times E_0$.

The character ψ_E is irreducible and is the unique character of E whose restriction to $Z(E)$ is a multiple of μ_0 (i.e. ψ is fully ramified over $Z(E)$). We already noted that E is the central product of E_0 with $C_E(E_0)$, and it follows from this that $\psi|_{E_0}$ is a multiple of a fixed irreducible character ψ_0 of E_0 that is also fully ramified over $Z(E) = Z(E_0)$. By consideration of degrees, we have $\psi|_{E_0} = q\psi_0$. For a discussion of characters of extraspecial p -groups, see [1, Chap. 31].

THEOREM 2.6. $\psi_{C \times E_0} = (\text{reg}_C - 1_C) \neq \psi_0$.

Proof. Since the irreducible characters of $C \times E_0$ have the form $\alpha \neq \beta$ where α and β are irreducible characters of C and E_0 respectively, we may write

$$\psi_{C \times E_0} = \sum \alpha_i \neq \beta_i.$$

Restrict this equation to E_0 and use $\psi_{E_0} = q\psi_0$ to conclude $q\psi_0 = \sum \alpha_i(1)\beta_i$. Therefore $\beta_i = \psi_0$ for all i and $\psi_{C \times E_0} = (\sum \alpha_i) \neq \psi_0$. Restrict this second equation to C and use Theorem 2.5(c) to conclude

$$q(\text{reg}_C - 1) = (\sum \alpha_i) \cdot q,$$

so

$$\sum \alpha_i = \text{reg}_C - 1$$

and

$$\psi_{C \times E_0} = (\text{reg}_C - 1) \neq \psi_0,$$

as desired. \square

3. Construction of G (characteristic 2). Throughout this section F will denote the field $GF(q^2)$ where q is a power of 2. As usual, $\bar{\cdot}$ denotes the unique automorphism of F having order 2 ($\bar{\alpha} = \alpha^q$ for all $\alpha \in F$). The fixed field of $\bar{\cdot}$ is $F_0 = GF(q)$.

Let $V = F \times F$, and define the group E as follows. As a set, E is $V \times GF(2)$ (sometimes viewed as $F \times F \times GF(2)$), and the multiplication is defined by

$$(a, b, m) \cdot (c, d, n) = (a + c, b + d, m + n + \text{tr}(a\bar{d}))$$

where $\text{tr} : F \rightarrow GF(2)$ is the trace map.

Define $U(2, F)$ to be the unitary group with respect to the Hermitian form whose matrix is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, that is

$$U(2, F) = \left\{ g \in GF(2, F) \mid g \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \bar{g}^T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}.$$

As usual, let $SU(2, F) = \{g \in U(2, F) \mid \det g = 1\}$. As in § 2, let

$$H = \left\{ \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \mid a \in F^\times \right\} \quad \text{and} \quad P = \left\{ \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \mid a \in F_0 \right\}.$$

The equation $SU(2, F) = SL(2, F_0)$ and the factorization $U(2, F) = H \cdot SL(2, F_0)$ hold in this section for the same reasons that they did in the last section.

Define

$$H_0 = H \cap SL(2, F_0) \quad \text{and} \quad \omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The group H (and hence H_0) normalizes P so that PH and PH_0 are groups. From the well-known Bruhat decomposition of $SL(2, F_0) = PH_0 \cup PH_0 \omega P$, we have $U(2, F) = PH \cup PH \omega P$.

Define matrices B and C by setting

$$B = \begin{pmatrix} 0 & 0 \\ \nu & 0 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

where ν is any fixed element of F satisfying $\nu + \bar{\nu} = 1$ (ν exists as the trace map $F \rightarrow F_0$ is surjective). Notice that for $v = (a, b)$ and $w = (c, d)$ we have $\text{tr}(a\bar{d}) = \text{tr}(nC\bar{w}^T)$, so that the multiplication in E may be written as

$$(v, m)(w, n) = (v + w, m + n + \text{tr}(vC\bar{w}^T)).$$

For each $g \in U(2, F)$ it is convenient to define the (Hermitian) quadratic form ϕ_g by setting $\phi_g(v) = v(B + gB\bar{g}^T)\bar{v}^T$ for each $v \in V$. Now, for $g \in U(2, F)$ and $e = (v, n) \in E$ define $e^g \in E$ by the equation:

$$e^g = (vg, n + \text{tr}(\phi_g(v))).$$

LEMMA 3.1. *The function $(e, g) \mapsto e^g$, as defined above, is an action of $U(2, F)$ on E by automorphisms.*

Proof. The functional equation (the “1-cocycle condition”):

$$\phi_{gh}(v) = \phi_g(v) + \phi_h(vg)$$

is easily checked (using characteristic 2) and implies that $(e, g) \mapsto e^g$ is an action. The action by automorphisms condition amounts to checking that

$$\text{tr}(vC\bar{w}^T) + \text{tr}(\phi_g(v+w)) = \text{tr}(\phi_g(v)) + \text{tr}(\phi_g(w)) + \text{tr}(vgC\bar{g}^T\bar{w}^T).$$

When $\phi_g(v+w)$ is expanded, the term $w(B+gB\bar{g}^T)\bar{v}^T$ may be replaced by the term $v(\bar{B}^T+g\bar{B}^T\bar{g}^T)\bar{w}^T$, as these field elements are algebraically conjugate and so have the same trace. When this is done, like terms are cancelled, and all terms are written on one side, the equation becomes:

$$\text{tr}(v(B+\bar{B}^T+C+g(B+\bar{B}^T+C)\bar{g}^T)\bar{w}^T) = 0.$$

From the definition of B and C , the matrix $B+\bar{B}^T+C$ is a scalar multiple of $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which is the matrix of the form over which $U(2, F)$ is defined. Since $g \in U(2, F)$ we have

$$g(B+\bar{B}^T+C)\bar{g}^T = B+\bar{B}^T+C,$$

and the previous equation is verified, since the characteristic is two. \square

It is worth pointing out that B and C may be replaced by any other pair of matrices which satisfy $B+\bar{B}^T+C = s\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ where s is any scalar in F which lies outside of F_0 . The last condition is needed to guarantee that $C+\bar{C}^T = (s+\bar{s})\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \neq 0$ so that E is a non-abelian group. (The commutator $[(v, m), (w, n)]$ simplifies to $(0, \text{tr}(v(C+\bar{C}^T)\bar{w}^T))$.)

COROLLARY 3.2. *Let $(v, m) = (v_1, v_2, m) \in E$. Then each of the matrices*

$$h = \begin{pmatrix} c & 0 \\ 0 & \bar{c}^{-1} \end{pmatrix}, \quad \omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix},$$

where $a \in F_0$ and $c \in F^\times$, belongs to $U(2, F)$ and

- (1) $(v, m)^h = (vh, m)$,
- (2) $(v, m)^\omega = (v\omega, m + \text{tr}(v_1\bar{v}_2))$,
- (3) $(v, m)^y = (vy, m + \text{tr}(\alpha v_1\bar{v}_1))$ where $\alpha \in F$ is any field element satisfying $\alpha + \bar{\alpha} = a$.

Proof. We have already seen that each of the matrices h, ω and y belongs to $U(2, F)$. Moreover, $hB\bar{h}^T = B$ so that $\phi_h(v) = 0$ for all v , and the first equation follows. Also, $\omega B\bar{\omega}^T = B^T$ so that

$$\phi_\omega(v) = v(B+B^T)\bar{v}^T = v\begin{pmatrix} 0 & \nu \\ \nu & 0 \end{pmatrix}\bar{v}^T = \nu\bar{v}_1v_2 + \nu v_1\bar{v}_2.$$

Since $\overline{\nu\bar{v}_1v_2} = \bar{\nu}v_1\bar{v}_2 = \nu v_1\bar{v}_2 + v_1\bar{v}_2$, we have:

$$\text{tr}(\phi_\omega(v)) = \text{tr}(\nu\bar{v}_1v_2 + \nu v_1\bar{v}_2) = \text{tr}(\overline{\nu\bar{v}_1v_2} + \nu v_1\bar{v}_2) = \text{tr}(v_1\bar{v}_2),$$

and the second equation follows. To verify the third equation, let $\alpha \in F$ satisfy $\alpha + \bar{\alpha} = a$. Then $B+yB\bar{y}^T = \begin{pmatrix} \alpha\nu & 0 \\ 0 & 0 \end{pmatrix}$ so $\phi_y(v) = \alpha\nu v_1\bar{v}_1$. Hence

$$\text{tr}(\phi_y(v)) = \text{tr}(\alpha\nu v_1\bar{v}_1) = \text{tr}(\nu\alpha v_1\bar{v}_1 + \nu\bar{\alpha}v_1\bar{v}_1).$$

Now $\nu\bar{\alpha}v_1\bar{v}_1$ and $\overline{\nu\bar{\alpha}v_1\bar{v}_1} = (\nu+1)\alpha v_1\bar{v}_1$ have the same trace, and this implies that

$$\text{tr}(\phi_y(v)) = \text{tr}(\alpha\nu v_1\bar{v}_1),$$

verifying the last equation. \square

By Lemma 3.1, the semidirect product $G = U(2, F) \ltimes E$ is defined, and as usual, $U(2, F)$ and E are viewed as subgroups of G . Recall that multiplication within G is defined by setting

$$(g, e) \cdot (h, e') = (gh, e^h e').$$

Define the subgroups of E (and hence of G) by setting:

$$\begin{aligned} A_0 &= \{0\} \times F \times \{0\}, & Z &= \{0\} \times \{0\} \times GF(2), \\ T &= F \times \{0\} \times \{0\}, & A &= A_0 Z \quad (= A_0 \dot{\times} Z). \end{aligned}$$

The group H normalizes each of these four subgroups of E , and PH normalizes A_0 and Z . Hence $PHA = PHA_0 \dot{\times} Z$ is a group. Notice that T is transversal for A in E , as well as for PHA in PHE .

Recall that μ_0 denotes the standard character of $GF(p) = \mathbb{Z}$, and hence we may regard μ_0 as a character of \mathbb{Z} . Define the representation ρ_1 of PHE by setting $\rho_1 = (1_{PHA_0} \# \mu_0)^{PHE}$, with the understanding that the transversal T is used in the construction of the induced representation. Since T naturally corresponds with F , $\rho_1(g)$ will be viewed as an $F \times F$ matrix for $g \in PHE$.

THEOREM 3.3. *The representations $\rho_1, \rho_1|_{PE}$ and $\rho_1|_E$ are all faithful and irreducible. Moreover,*

(1) *If $e = (a, b, m) \in E$, then the (r, s) entry of $\rho_1(e)$ is $\delta_{r+a, s} \mu_0(m) \lambda_{\bar{b}}(r)$. In particular, for $e = (0, b, 0) \in A_0$, we have $\rho_1(e) = E_{\bar{b}}$.*

(2) *If $y = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \in P$, then $a \in F_0$ may be written in the form $a = \alpha + \bar{\alpha}$ for some $\alpha \in F$, and $\rho_1(y) = D_\alpha$.*

(3) *If $h = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \in H$, then $\rho_1(h) = N_a$.*

Proof. The proof of the first assertion, as well as the formulas in (1) and (3), is similar to that of Theorem I.5.1 and will be omitted. Only the assertion in (2) will be proved here. Let $y \in P$ so that $y = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$ for some $a \in F_0$. Then $a = \alpha + \bar{\alpha}$ for some $\alpha \in F$, as the trace map $F \rightarrow F_0$ is surjective. If $r \in F$ then $PHA(r, 0, 0)y = PHAy(r, 0, 0)^y = PHA(r, ar, \text{tr}(\alpha r \bar{r}))$. Hence, the (r, s) entry of $\rho_1(y)$ is zero unless $s = r$. The (r, r) entry is:

$$\begin{aligned} (1_{PHA_0} \# \mu_0)((r, 0, 0)y(r, 0, 0)) &= (1_{PHA_0} \# \mu_0)(y(0, ar, \text{tr}(\alpha r \bar{r}))) \\ &= \mu_0(\text{tr}(\alpha r \bar{r})) = \lambda_\alpha(r \bar{r}). \end{aligned}$$

By definition of D_α , then, $\rho_1(y) = D_\alpha$. \square

THEOREM 3.4. *The representation ρ_1 of PHE is uniquely extendible to a representation ρ of G . Moreover, ρ is unimodular if $q^2 > 4$, while if $q^2 = 4$, the index in G of $\ker(\det \rho)$ is 2.*

Proof. The uniqueness of ρ will be established first. Assume that ρ and $\tilde{\rho}$ are extensions of ρ_1 to G . Then $\tilde{\rho} = \kappa \rho$ for some linear character κ of G . Now $G' \subseteq \ker \kappa$, and since $G = G' \cdot PH$, the character κ is uniquely determined by its restriction to PH . However, $\tilde{\rho}|_{PH} = \rho|_{PH}$ (since both equal $\rho_1|_{PH}$), and hence $\kappa_{PH} = 1_{PH}$. This proves $\kappa = 1_G$ and $\tilde{\rho} = \rho$.

To prove the existence of ρ , notice first that, as in the proof of Theorem I.4.3, the representation $\rho_1|_E$ is extendible to G . Let $\tilde{\rho}$ be an extension of $\rho_1|_E$ to G . Since $\tilde{\rho}|_{PHE}$ and ρ_1 are both extensions of $\rho_1|_E$ to PHE , we have $\tilde{\rho}|_{PHE} = \theta \rho_1$ for some linear character θ of PHE . Hence $(PHE)' \subseteq \ker \theta$.

Assume first that $q^2 = 4$. Then $(PHE)' = E$ and PH is a complement for G' in G . Thus, $PHE/E \cong G'/G'$ so that θ has a unique extension, say λ , to G . Then $\rho = \lambda \tilde{\rho}$ is an extension of ρ_1 to G .

Suppose now $q^2 > 4$. We still have $\tilde{\rho}|_{PHE} = \theta\rho_1$, but this time $(PHE)' = PE$, so that $PE \subseteq \ker \theta$. Hence $\theta^{q^2-1} = 1_{PHE}$. Now let $\delta = \det \tilde{\rho}$ and $\rho = \tilde{\delta}\tilde{\rho}$. Since $G = G'H$ we have $\delta^{q^2-1} = 1_G$. Hence $\det \rho = \det(\tilde{\delta}\tilde{\rho}) = \tilde{\delta}^{q^2} \det \tilde{\rho} = \delta^{1-q^2} = 1_G$. Therefore

$$\begin{aligned} 1_{PHE} &= (\det \rho)_{PHE} = \det(\rho|_{PHE}) = \det(\tilde{\delta}_{PHE}\tilde{\rho}_{PHE}) = \det(\tilde{\delta}_{PHE}\theta\rho_1) \\ &= (\tilde{\delta}_{PHE}\theta)^{q^2} \det \rho_1 = \tilde{\delta}_{PHE}\theta \det \rho_1. \end{aligned}$$

The equation $\det \rho_1 = 1_{PHE}$ follows from Theorem 3.3 (where the hypothesis $q^2 > 4$ is also used in the second part of that theorem) so $1_{PHE} = \tilde{\delta}_{PHE}\theta$ or $\delta_{PHE} = \theta$. Hence, $\rho|_{PHE} = \tilde{\delta}_{PHE}\theta\rho_1 = \rho_1$, so that ρ extends ρ_1 .

Finally, since $G = PHG'$, we have $|G : \ker(\det \rho)| = |PHE : \ker(\det \rho_1)|$. We already observed that ρ_1 is unimodular when $q^2 > 4$, and hence so is ρ in this case. When $q^2 = 4$, Theorem 3.3 (1) and (3) imply that $\ker(\det \rho_1) \supseteq HE$. However, by the second part of that theorem, $\det \rho_1(y) = -1$ where $y = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is the generator of P . Hence, this index is 2 and the proof of Theorem 3.4 is complete. \square

THEOREM 3.5. *Let ρ be the unique extension of ρ_1 to G guaranteed by Theorem 3.4. Then*

$$\rho \begin{pmatrix} 0 & a \\ \bar{a}^{-1} & 0 \end{pmatrix} = -M_a \quad \text{for all } a \in F^\times.$$

Proof. By Lemma 1.2, $M_1^{-2} = N_{-1} = N_1 = I$, so $M_1^{-1} = M_1$. Using this and Theorem 3.3 (1), the (r, s) entry of $M_1^{-1}\rho(a, b, m)M_1$ is

$$\begin{aligned} &(1/q^2) \sum_{t, u \in F} \lambda_1(r, \bar{t}) \delta_{t+a, u} \mu_0(m) \lambda_{\bar{b}}(t) \lambda_1(u\bar{s}) \\ &= (1/q^2) \sum_{t \in T} \lambda_1((\bar{r} + \bar{b} + \bar{s})t + a\bar{s}) \mu_0(m) \end{aligned}$$

where we have used the identities $\lambda_1(\bar{x}) = \lambda_1(x)$ and $\lambda_1(x+y) = \lambda_1(x)\lambda_1(y)$. If $\bar{r} + \bar{b} + \bar{s} \neq 0$ then the sum reduces to zero. Hence, the (r, s) entry reduces to

$$\begin{aligned} \delta_{r+b, s} \mu_0(m) \lambda_1(a\bar{s}) &= \delta_{r+b, s} \mu_0(m) \lambda_a(b) \lambda_{\bar{a}}(r) \\ &= \delta_{r+b, s} \mu_0(m + \text{tr}(a\bar{b})) \lambda_{\bar{a}}(r). \end{aligned}$$

By Theorem 3.3 again, this is the (r, s) entry of $\rho_1(b, a, m + \text{tr}(a\bar{b}))$. Hence

$$\begin{aligned} M_1^{-1}\rho(a, b, m)M_1 &= \rho(b, a, m + \text{tr}(a\bar{b})) = \rho((a, b, m)^\omega) \\ &= \rho(\omega)^{-1}\rho(a, b, m)\rho(\omega) \end{aligned}$$

where $\omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Therefore, $\rho(\omega)M_1^{-1}$ centralizes $\rho(E)$ and hence is a scalar matrix by Schur's lemma. This proves that $\rho \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = cM_1$ for some scalar c . Now $M_aM_1 = N_{-a} = N_a$, so $M_a = N_aM_1$. By Theorem 3.3 (3), we have

$$\rho \begin{pmatrix} 0 & a \\ \bar{a}^{-1} & 0 \end{pmatrix} = \rho \left(\begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right) = N_a \cdot cM_1 = cM_a.$$

It remains to prove $c = -1$. Notice at this point, since $\rho \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and M_1 both have order 2, we know $c^2 = 1$ so $c = \pm 1$.

The (r, r) entry of M_1 is $(1/q)\lambda_1(r\bar{r}) = (1/q)\mu_0(\text{tr}(r\bar{r}))$. As $r\bar{r} \in F_0$, $\text{tr}(r\bar{r}) = 0$ and $\mu_0(0) = 1$, so each diagonal element of M_1 is $1/q$. Therefore, $\text{trace}(M_1) = q^2/q = q$ and $c = (1/q) \cdot (\text{trace } \rho \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix})$. Now $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is conjugate to $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ in $SL(2, GF(2)) \leq SL(2, F_0) \leq U(2, F)$ so $c = (1/q) \cdot (\text{trace } \rho \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}) = (1/q)(\text{trace } D_\nu)$ where $\nu + \bar{\nu} = 1$.

Hence

$$c = (1/q) \sum_{r \in F} \lambda_\nu(r\bar{r}) = (1/q) \sum_{r \in F} \mu_0(\text{tr}(\nu r\bar{r})).$$

Let $\text{tr}^0: F_0 \rightarrow GF(2)$ be the trace map. Then $\text{tr}(\nu r\bar{r}) = \text{tr}^0(\nu r\bar{r} + \bar{\nu} r r) = \text{tr}^0(r\bar{r})$. Hence.

$$\begin{aligned} c &= (1/q) \left(1 + (q+1) \sum_{s \in F_0^\times} \mu_0(\text{tr}^0(s)) \right) \\ &= (1/q) \left((q+1) \sum_{s \in F_0} \mu_0(\text{tr}^0(s)) - q \right) \\ &= (1/q)(0 - q) = -1. \end{aligned}$$

This completes the proof of Theorem 3.5. \square

Define $G_0 = U(2, F) \cdot \mathbb{Z} \leq G$. The next result identifies the linear groups defined in § 1 as representations of the groups G_0 and G .

THEOREM 3.6.

$$\rho(U(2, F) \cdot \mathbb{Z}) = \langle M_a, N_a, D_b \mid a, b \in F, a \neq 0 \rangle$$

and

$$\rho(G) = \langle M_a, N_a, D_b, E_b \mid a, b \in F, a \neq 0 \rangle.$$

Proof. Let $X = \langle M_a, N_a, D_b \mid a, b \in F, a \neq 0 \rangle$ and let X_1 be the group generated by X and $-I$. Clearly, $-I = \rho(0, 0, 1) \in \rho(\mathbb{Z}) \leq \rho(U(2, F)\mathbb{Z})$. By Theorems 3.3 and 3.5, the generators of the group X belong to $\rho(U(2, F)\mathbb{Z})$, and hence $X_1 \leq \rho(U(2, F)\mathbb{Z})$. Now the decomposition

$$U(2, F) = PH \cup PH\omega P \quad \text{where } \omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

proves the reverse inclusion (using Theorem 3.3 again). The first assertion of the theorem will follow from the equation $X = X_1$, i.e. $-I \in X$.

Clearly the index of X in X_1 is at most 2, and hence $X'_1 \leq X$. Now $X'_1 = \rho(U(2, F)\mathbb{Z})' = \rho(U(2, F)')$. If $q^2 > 4$ then $U(2, F)' = SU(2, F)$, so

$$-M_1 = \rho \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \rho(SU(2, F)) = \rho(U(2, F)') \leq X.$$

Since $M_1 \in X$, this leads to $-I \in X$. Suppose then $q^2 = 4$. Recall that $\nu \in F$ was chosen so that $\nu + \bar{\nu} = 1$. By direct calculation, $-I = (D_\nu M_1)^3 \in X$. In either case, then, $X = X_1$ and the first assertion of the theorem follows.

Now let $Y = \langle M_a, N_a, D_b, E_b \mid a, b \in F, a \neq 0 \rangle$. Clearly $X \leq Y$. As $E_b = \rho(0, \bar{b}, 0) \in \rho(G)$, we have $Y \leq \langle X, E_b \mid b \in F \rangle \leq \rho(G)$. But $\rho(U(2, F)\mathbb{Z}) = X \leq Y$, and since $U(2, F)$ acts irreducibly on $E/\mathbb{Z}(E)$, it follows that $U(2, F)\mathbb{Z}$ is a maximal subgroup of G . Hence X is a maximal subgroup of $\rho(G)$. For $b \neq 0$, $E_b = \rho(0, \bar{b}, 0) \in \rho(E) - \rho(\mathbb{Z}(E))$, so $E_b \in Y - X$, and this proves $Y = \rho(G)$. \square

As in the preceding section, the restriction of the character of ρ to various subgroups of G_0 and G will be identified.

Let C denote the subgroup of $U(2, F)$ consisting of those matrices which fix $(\nu, 1) \in V$. Recall that $\nu \neq 0$ and $\nu + \bar{\nu} = 1$. Any matrix M fixing $(\nu, 1)$ has the form

$$\begin{pmatrix} a & b \\ a\nu + \nu & b\nu + 1 \end{pmatrix}.$$

If, in addition, $M \in U(2, F)$, then $a\bar{b} = \bar{a}b$ and $\bar{a} + \bar{b}\nu = 1$, so that $a + b\bar{\nu} = 1$. Setting $r = \det M$, we have $a + b\nu = r$, and this leads easily to $b = 1 + r$ and $a = 1 + (1+r)\bar{\nu}$. In particular, any matrix M in C is uniquely determined by its determinant r , and is given by

$$M(r) = \begin{pmatrix} 1 + (1+r)\bar{\nu} & 1+r \\ \nu\bar{\nu}(1+r) & 1 + (1+r)\nu \end{pmatrix}.$$

The determinant of any matrix in $U(2, F)$ belongs to N , so that $C = \{M(r) | r \in N\}$. Since $M(r_1)M(r_2) \in C$ has determinant r_1r_2 , the preceding paragraph implies $M(r_1r_2) = M(r_1)M(r_2)$. Of course, this can be checked directly. Hence $C \cong N$ and since $C \cap SU(2, F) = \{1\}$, the group $U(2, F)$ is an internal semidirect product of C and $SU(2, F)$.

Let $\tilde{C} = y^{-1}Cy$ where

$$y = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in U(2, F).$$

Since $(\nu, 1)y = (\bar{\nu}, 1)$, \tilde{C} may be defined equivalently as the subgroup of $U(2, F)$ fixing the vector $(\bar{\nu}, 1)$. For $r \in N$,

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} M(r) \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = M(\bar{r}) \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix} \in CD,$$

where D is the set of scalar matrices belonging to $U(2, F)$. Clearly, $CD = \tilde{C}C$ is isomorphic to $N \times N$.

THEOREM 3.7. *Let ψ denote the character of ρ . Then*

- (a) $\psi_H = 1_H + \text{reg}_H$,
- (b) $\psi_{PD} = 1_{PD} + \text{reg}_{PD} - \text{reg}_{PD/P}$,
- (c) $\psi_{CD} = \text{reg}_{C\tilde{C}} - \text{reg}_{C\tilde{C}/C} - \text{reg}_{C\tilde{C}/\tilde{C}} + 1_{C\tilde{C}}$.

Proof. The proof of (a) and (b) is identical to the corresponding parts of the proof of Theorem 2.5, and will be omitted. As in part (c) of Theorem 2.5, the assertion that

$$\psi_{CD} = \text{reg}_{CD} - \text{reg}_{C\tilde{C}/C} = \text{reg}_{C\tilde{C}/\tilde{C}} - 1_{C\tilde{C}}$$

amounts to showing that

$$\psi(x) = \begin{cases} 1 & \text{if } s \neq 1 \text{ and } rs \neq 1, \\ -q & \text{if } s = 1 \text{ or } rs = 1, \text{ but not both,} \\ q^2 & \text{if } r = s = 1 \end{cases}$$

where $x = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \in CD$.

When $r = 1$, the formula for $\psi(x)$ follows from (a) as $x \in D \cong H$. For $r \neq 1$, x has the form

$$\begin{pmatrix} b & 0 \\ 0 & \bar{b}^{-1} \end{pmatrix} \begin{pmatrix} 1 & c_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & c_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix},$$

where

$$\begin{aligned} b &= r/\nu\bar{\nu}(r+1), \\ c_1 &= (1 + (1+r)\bar{\nu}) \cdot \nu\bar{\nu}(1+\bar{r}), \\ c_2 &= (1 + (1+r)\nu)/\nu\bar{\nu}(1+r). \end{aligned}$$

Let $\gamma_1 = \nu c_1$ and $\gamma_2 = \nu c_2$ so that $\gamma_1 + \bar{\gamma}_1 = c_1$ and $\gamma_2 + \bar{\gamma}_2 = c_2$. By a calculation similar to the one appearing in the proof of Theorem 2.5 (c), we have

$$\psi(x) = 1 - (q+1)(\lambda_k|_{F_0}, 1_{F_0})$$

where $k = \gamma_1 b\bar{b} + sb + \gamma_2$. The inner product $(\lambda_k|_{F_0}, 1_{F_0})$ is equal to 1 if kF_0 is in the kernel of λ , and is zero otherwise. Since F_0 is the only nontrivial F_0 subspace of F in the kernel of λ , we have $\psi(x) = 1$ if $k \notin F_0$ and $\psi(x) = -q$ if $k \in F_0$. Now $k \in F_0$ if and only if $k + \bar{k} = 0$. When $k + \bar{k}$ is expressed in terms of r and s , the result is

$$k + \bar{k} = \frac{\bar{s}r}{\nu\bar{\nu}(r+1)}(s^2 + (1 + \bar{r})s + \bar{r}).$$

Thus, $k + \bar{k} = 0$ if and only if $s \in \{1, \bar{r}\}$. The formula for $\psi(x)$ follows, and with it, the proof of Theorem 3.7 is complete. \square

Define a subgroup E_0 of E by setting $E_0 = \{(s\nu, s, n) | s \in F, n \in GF(2)\}$. The group E_0 is an extraspecial subgroup of E . In fact, $Z(E_0) = Z(E)$, $C_E(E_0) = \{(s\bar{\nu}, s, n) | s \in F, n \in GF(2)\}$, and E is the central product of E_0 with $C_E(E_0)$. Since $(s\nu, s) \in V$ is an eigenvector with eigenvalue 1 for any matrix in C , the group $E_0/Z(E_0)$ is centralized by C . Clearly $Z(E_0)$ is centralized by C since $|Z(E_0)| = 2$, and hence $[E_0, C, C] = 1$. As $|C|$ is odd this implies $[E_0, C] = 1$, and so C centralizes E_0 . Of course, this last result may be verified by direct calculation. Hence $CE_0 = C \times E_0$.

As in the discussion preceding Theorem 2.6, the character $\psi|_{E_0} = q\psi_0$ where ψ_0 is a fixed irreducible character of E_0 . The proof of the next result is similar to that of Theorem 2.6, and so will be omitted.

THEOREM 3.8. $\psi_{C \times E_0} = (\text{reg}_C - 1_C) \# \psi_0$.

4. Molien series (odd characteristic). The Molien series of a representation σ of a finite group H is the rational function

$$\Phi_{H,\sigma}(X) = (1/|H|) \sum_{g \in H} 1/\det(I - X\sigma(g)).$$

The coefficient of X^d in this expansion is the dimension of the space of homogeneous polynomials of degree d which are invariant under the action of the matrix group $\sigma(H)$. A proof of this fact may be found in [11].

Each polynomial $\det(I - X\sigma(g))$ is an invariant of the conjugacy class of g , and it is convenient to group together terms from the same class. The sum may then be calculated once the conjugacy classes (and their sizes) are known, and for each class the polynomial $\det(I - X\sigma(g))$ is determined. The construction will be carried out here for the groups G_0 and G with the representation ρ defined in § 2.

Recall that $\langle -1 \rangle$ (the cyclic group of order 2) occurs as a direct factor of both G_0 and G , namely

$$G = \langle -1 \rangle \times \tilde{G} \quad \text{and} \quad G_0 = \langle -1 \rangle \times U,$$

where $U = U(2, F)$. Moreover, -1 is represented by $-I$ under ρ . Omitting the dependence of the Molien series on ρ , we may write:

$$\Phi_G(X) = (\Phi_{\tilde{G}}(X) + \Phi_{\tilde{G}}(-X))/2 \quad \text{and} \quad \Phi_{G_0}(X) = (\Phi_U(X) + \Phi_U(-X))/2.$$

These equations show that Φ_G and Φ_{G_0} are easily determined from $\Phi_{\tilde{G}}$ and Φ_U . The goal for this section is to establish a workable notation for the classes of \tilde{G} and U , and for each class representative g , to determine $\det(I - X\rho(g))$. The results are tabulated in Tables 4.1 and 4.2, and the Molien series are given in Theorems 4.6 and 4.7. Notice that since $\tilde{G} = U \rtimes E$, the group U occurs as a factor group of \tilde{G} , and distinct conjugacy classes of U are contained in distinct conjugacy classes of \tilde{G} (although their sizes are not generally the same).

TABLE 4.1
Valid for all q .

Class representative g	Number of classes	Size of class	$\det(I - X\rho(g))$	Remarks
$\begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}, s \in N$	$q+1$	1	$(1-X)(1-X^l)^{(q^2-1)/l}$	$o(s) = l (q+1)$
$\begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, s \in N$	$q+1$	$(q+1)(q-1)$	$(1-X)(1-X^{pl})^{q(q+1)/pl} / (1-X^l)^{(q+1)/l}$	$o(s) = l (q+1)$
$\begin{pmatrix} r & 0 \\ 0 & r^{-1} \end{pmatrix}, r \in \Gamma$	$(q+1)(q-2)/2$	$q(q+1)$	$(1-X)(1-X^m)^{(q^2-1)/m}$	$o(r) = m (q^2-1)$ and $m \nmid (q+1)$
$M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}, (r, s) \in \Delta$	$q(q+1)/2$	$q(q-1)$	$\frac{(1-X^{abk})^{(q+1)/2/abk} (1-X)}{(1-X^{ak})^{(q+1)/ak} (1-X^{bk})^{(q+1)/bk}}$	$a = (g) \cap C ,$ $b = (g) \cap \check{C} ,$ $abk = \text{l.c.m.} \{o(rs), o(s)\}$ $= o(g) (q+1)$

TABLE 4.2

Class representative g	Number of classes	Size of class	$\det(I - X\rho(g))$ for $\zeta = 1$	Remarks
$\zeta, \zeta \in \mathbb{Z}$	p	1	$(1 - X)^{q^2}$	$o(d) = l(q + 1)$
$d\zeta, d \in D^\#, \zeta \in \mathbb{Z}$	pq	q^4	$(1 - X)(1 - X^l)^{(q^2-1)/l}$	$o(r) = l(q^2 - 1), l \nmid (q + 1)$
$\begin{pmatrix} r & 0 \\ 0 & r^{-1} \end{pmatrix} \zeta, r \in \Gamma, \zeta \in \mathbb{Z}$	$p(q + 1)(q - 2)/2$	$q^5(q + 1)$	$(1 - X)(1 - X^l)^{(q^2-1)/l}$	$o(r) = a (q + 1)$ $a = g \cap C,$ $b = g \cap \bar{C},$ $abk = \text{l.c.m. } \{o(rs), o(s)\}$ $= o(g) (q + 1)$
$M(r)\zeta, r \in N^\#, \zeta \in \mathbb{Z}$	pq	$q^3(q - 1)$	$((1 - X^a)^{(q+1)/a} / (1 - X))^q$	
$M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \zeta, (r, s) \in \Delta_2, \zeta \in \mathbb{Z}$	$pq(q - 1)/2$	$q^5(q - 1)$	$\frac{(1 - X^{abk})^{(q+1)/2} / abk (1 - X)}{(1 - X^{ak})^{(q+1)/ak} (1 - X^{bk})^{(q+1)/bk}}$	
$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \zeta, \zeta \in \mathbb{Z}$	p	$q^2(q + 1)(q - 1)$	$(1 - X)(1 - X^p)^{q(q+1)/p} / (1 - X)^{q+1}$	
$\begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \zeta, r \in N^\#, \zeta \in \mathbb{Z}$	pq	$q^4(q + 1)(q - 1)$	$(1 - X)(1 - X^{p^l})^{(q+1)q/p^l} / (1 - X^l)^{(q+1)/l}$	$o(r) = l (q + 1)$
$M(r)(sv, s, 0), r \in N^\#, s \in \Lambda$	$q(q - 1)$	$pq^2(q + 1)(q - 1)$	$(1 - X^{p^l})^{q(q+1)/p^l} / (1 - X^p)^{q/p}$	$o(r) = l (q + 1)$
$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} (r, r, 0), r \in \Lambda$	1 $(q - 1)$	$p(q + 1)^2(q - 1)$ $p(q + 1)q(q - 1)$	$(1 - X^p)^{q^2/p}$ $(1 - X^p)^{q^2/p}$	
$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} (s, 0, 0), s \in \Lambda$	$(q - 1)$	$p(q + 1)^2q^2(q - 1)$	$p > 3: (1 - X^p)^{q^2/p}$	
			$p = 3: \begin{cases} (1 - X^3)^{q^2/3} \\ (1 - \omega X^3)^{q^2/3} \\ (1 - \omega^2 X^3)^{q^2/3} \end{cases}$	$\text{tr}(s\bar{s}) = 0$ $\text{tr}(s\bar{s}) = -1$ $\text{tr}(s\bar{s}) = 1$

Even though the notation of § 2 is carried over to this section, it is worthwhile to observe that the results of the next four lemmas remain valid in characteristic 2, with only minor changes. In fact, the statements of Lemmas 4.1 and 4.2 require no change at all.

LEMMA 4.1. *If $x \in H - D$, then $C_U(x) = H$. If $x \in CD - D$, then $C_U(x) = CD$.*

Proof. In either case, x is not a scalar matrix. For $x \in H - D$, $C(x)$ must consist of diagonal matrices, so $C_U(x) = H$. For $x \in CD - D$, x stabilizes two orthogonal one-dimensional subspaces of V , say Fv and Fw (in fact, we may take $v = (v, 1)$ and $w = (\bar{v}, 1)$). Since x does not act scalarly on V , $C(x)$ must stabilize these subspaces, and $C_U(x)$ acts as a one-dimensional unitary group on each. Now $|U(1, F)| = q + 1$, so $|C_U(x)| \leq (q + 1)^2$. Clearly, $CD \subseteq C_U(x)$, and the result follows. \square

LEMMA 4.2. (a) *Let*

$$x = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \quad \text{and} \quad x' = M(r') \begin{pmatrix} s' & 0 \\ 0 & s' \end{pmatrix}$$

be elements of $CD - C$. Then x is conjugate to x' if and only if $x = x'$ or $r' = r^{-1}$ and $s' = rs$.

(b) *Let*

$$h = \begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix} \quad \text{and} \quad h' = \begin{pmatrix} a' & 0 \\ 0 & \bar{a}'^{-1} \end{pmatrix}$$

be elements of $H - D$. Then h is conjugate to h' if and only if $h = h'$ or $a' = \bar{a}^{-1}$.

Proof. Assume x is conjugate to x' in U . Then the eigenvalues of x and x' must coincide, and so $\{rs, s\} = \{r's', s'\}$. Hence $x = x'$ or else $r' = r^{-1}$ and $s' = rs$. Conversely choose $c \in F^\times$ so that $c\bar{c} = -1$. Then $y = c \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ belongs to U and conjugates $M(r)$ to $M(r^{-1}) \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}$. Hence, y conjugates $M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$ to $M(r^{-1}) \begin{pmatrix} rs & 0 \\ 0 & rs \end{pmatrix}$, completing the proof of (a).

To prove (b), notice that $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ belongs to U and conjugates $\begin{pmatrix} a & 0 \\ 0 & \bar{a}^{-1} \end{pmatrix}$ to $\begin{pmatrix} \bar{a}^{-1} & 0 \\ 0 & a \end{pmatrix}$. Conversely, if h is conjugate to h' , then equating eigenvalues we have $\{a, \bar{a}^{-1}\} = \{a', \bar{a}'^{-1}\}$, and (b) follows. \square

Notice that in the situation of Lemma 4.2, any element x satisfying (a) has order dividing $q + 1$, while no element h satisfying (b) has order dividing $q + 1$. Thus x is never conjugate to h . This same conclusion can be reached using Lemma 4.1, as the elements have centralizers of different orders.

It is convenient to define indexing sets for these two types of conjugacy classes. Recall that $N = \{a \in F \mid a\bar{a} = 1\}$ is the kernel of the norm map. Let \sim denote the equivalence relation on $F^\times - N$ defined by $a \sim b$ if and only if $a = b$ or $a = \bar{b}^{-1}$, and let Γ be a set of representatives for the \sim equivalence classes. Similarly, let \approx be the equivalence relation on $N^\# \times N$ defined by $(r, s) \approx (r', s')$ if and only if $(r, s) = (r', s')$ or $(r', s') = (r^{-1}, rs)$, and let Δ be a set of representatives for the \approx equivalence classes.

We assume that the representatives in Δ are chosen so that if $(r, s) \in \Delta$ and $g = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$, then $|\langle g \rangle \cap C| \cong |\langle g \rangle \cap \tilde{C}|$. This choice is possible, as g is conjugate to $g' = M(r) \begin{pmatrix} r^{-1}s & 0 \\ 0 & r^{-1}s \end{pmatrix}$ where $|\langle g' \rangle \cap C| = |\langle g \rangle \cap \tilde{C}|$ and $|\langle g' \rangle \cap \tilde{C}| = |\langle g \rangle \cap C|$. As a special case, since $(a, a^{-1}) \approx (a, 1)$, we have $(a, a^{-1}) \notin \Delta$ for all $a \in N^\#$, and so $N^\# \times \{1\} \subseteq \Delta$. This choice will be important when classes of \tilde{G} are described.

The preceding lemma showed that the elements

$$\begin{pmatrix} r & 0 \\ 0 & \bar{r}^{-1} \end{pmatrix} \quad \text{for } r \in \Gamma$$

lie in distinct conjugacy classes of U , as do the elements

$$M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \text{ for } (r, s) \in \Delta.$$

Moreover, it was already observed that these two sets of classes are disjoint.

Clearly, the elements $\begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$ for $s \in N$ are in distinct classes of U (since they are central in U) and the p -singular elements $\begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ for $s \in N$ are also in distinct classes. By direct computation the centralizer of $\begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ in U is PD , which has index $(q + 1)(q - 1)$ in U .

At this point we have verified the second and third columns of Table 4.1. As the dot product of these two columns is $(q + 1)^2 q (q - 1) = |U|$, all classes of U are accounted for.

It is convenient to observe that Lemma 4.1 remains valid in characteristic 2 (even though the subgroup C is given by a different definition). Moreover, Lemma 4.2 also remains valid. The only necessary change to make in the proof is to substitute $y = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for the conjugating element $c \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$. No restriction on the characteristic of F was needed to define the relations \sim and \approx (and the corresponding sets Γ and Δ). With this in mind, the entries in the second and third columns remain valid for q a power of 2.

We have already observed that distinct conjugacy classes of U lie in distinct conjugacy classes of \tilde{G} . The following lemma extends this to the subgroup UZ , and determines the second and third columns of Table 4.2 for the first seven lines of that table.

LEMMA 4.3. *If $u \in U$ and $\zeta \in Z$ with $\zeta \neq 1$, then u is not conjugate to $u\zeta$ in \tilde{G} . Moreover, $C_{\tilde{G}}(u) = C_{\tilde{G}}(u\zeta)$, so that the sizes of the conjugacy classes containing u and $u\zeta$ are identical. Table 4.3 determines centralizers in \tilde{G} of specific elements of U .*

TABLE 4.3

u	$u \in D^\#$	$u \in H - D$	$u = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$ $rs, s \in N^\#$	$u = M(r)$ $r \in N^\#$	$u \in P^\#$	$u \in P^\# D^\#$
$C_{\tilde{G}}(u)$	UZ	HZ	CDZ	CDE_0	DPA	PDZ

Proof. Suppose u is conjugate to $u\zeta$ in \tilde{G} where $u \in U$ and $\zeta \in Z^\#$. Any element which conjugates u to $u\zeta$ must centralize $u \pmod E$, and so has the form ce where $c \in C_U(u)$ and $e \in E$. As $C_U(u)$ centralizes u , we may assume the conjugating element is e . Thus $e^{-1}ue = u\zeta$, which is equivalent to $ueu^{-1} = e\zeta$. Write $e = (v, m)$ and $\zeta = (0, n)$. Then $(vu^{-1}, m) = (v, m + n)$, so that $n = 0$. However, $\zeta \neq 1$ so $n \neq 0$, and this contradiction establishes the first part of the lemma. Clearly, since Z is the center of \tilde{G} , $C_{\tilde{G}}(u) = C_{\tilde{G}}(u\zeta)$. The centralizers appearing in the table are easily worked out, and the proof is omitted. \square

Recall that Δ is a set of representatives for the equivalence classes under the relation \approx defined on $N^\# \times N$. From the original definition of Δ we have $N^\# \times \{1\} \subseteq \Delta$. Let $\Delta_1 = N^\# \times \{1\}$ and $\Delta_2 = \Delta - \Delta_1$. By the preceding lemma, $\Delta_1 \times Z$ and $\Delta_2 \times Z$ are natural index sets for the conjugacy classes of \tilde{G} represented by elements contained in $CDZ - DZ$.

Let Λ be a set of coset representatives for N in F^\times . Thus $|\Lambda| = |F^\times : N| = q - 1$. The set Λ is a useful index set for several of the conjugacy classes of \tilde{G} not covered by Lemma 4.3, as the next lemma shows.

For any group element g , let $g_{p'}$ and g_p denote the p' - and the p -part of g respectively. That is, $g = g_p g_{p'} = g_{p'} g_p$ where $g_{p'}$ has order prime to p , and the order of g_p is a power of p . Since PE is a Sylow p -subgroup of \tilde{G} , every element of \tilde{G} is conjugate in \tilde{G} to an element g for which $g_p \in PE$.

LEMMA 4.4. *Let $g \in \tilde{G}$ and assume that g_p , the p -part of g , is contained in PE .*

- (a) *If $g_p \in \mathbb{Z}$, then g is conjugate to an element of $U\mathbb{Z}$.*
- (b) *If $g_p \in E - \mathbb{Z}$, then exactly one of the following cases occurs:*
 - (i) $g_{p'} = 1$ and g is conjugate to $(0, 1, 0)$;
 - (ii) $g_{p'} = 1$ and g is conjugate to $(s\nu, s, 0)$ for some unique s in Λ ;
 - (iii) $g_{p'} \neq 1$ and g is conjugate to $M(r)(s\nu, s, 0)$ for some unique $r \in N^\#$ and $s \in \Lambda$.
- (c) *If $g_p \in PE - E$, then either g is conjugate to an element of $U\mathbb{Z}$, or else g is conjugate to $(\begin{smallmatrix} 1 & \\ 0 & 1 \end{smallmatrix})(s, 0, 0)$ for some unique s in Λ .*

Proof. (a) Assume $g_p \in \mathbb{Z}$. Then the element $g\mathbb{Z}$ of the group \tilde{G}/\mathbb{Z} is p -regular. Let $L = \langle g \rangle E$ so that $|L : E|$ is the order of $g_{p'}$ and equals $|U \cap L|$. By the Schur-Zassenhaus theorem, $g_{p'} \in L = (U \cap L) \cdot E$ must be conjugate in L to an element of $U \cap L$ (see [8, § 1.18, pp. 126ff]). Hence g is conjugate to an element of $U\mathbb{Z}$.

(b) Assume $g_p \in E - \mathbb{Z}$. By part (a), $g_{p'}$ is conjugate to an element of U , and hence is conjugate to an element of H or CD . Since $E - \mathbb{Z}$ is a normal subset of \tilde{G} , we may assume $g_{p'} \in H \cup CD$. Now no nonidentity element of H centralizes any element of $E - \mathbb{Z}$, so $g_{p'} \in CD$.

Suppose first $g_{p'} = 1$. Then $g = g_p \in E - \mathbb{Z}$, so g has the form (v, m) for some $v \in V = F \times F$ and $m \in GF(p)$. Since (v, m) is conjugate to $(v, 0)$ in E , we may assume $m = 0$. The action of U by conjugation on $V \times \{0\} \subseteq E$ may be identified with the action of U on V given by matrix multiplication. Recall that ν is a fixed element of $F^\#$ satisfying $\nu + \bar{\nu} = 0$. The "norm map" $\eta : V \rightarrow F_0$ given by $\eta(v) = \nu v (\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}) \bar{v}^T$ is invariant under the action of U . In fact, U acts transitively on the sets $\eta^{-1}(a)$ for $a \in F_0^\#$, and is transitive on $\eta^{-1}(0) - \{0\}$. Clearly $(0, 1) \in \eta^{-1}(0)$, and since $\eta(s\nu, s) = 2\nu^2 s \bar{s}$, the elements $(s\nu, s)$ for $s \in \Lambda$ lie in the distinct $\eta^{-1}(a)$ for $a \in F_0^\times$. Thus, g is conjugate to $(0, 1, 0)$ or else to $(s\nu, s, 0)$ for some unique $s \in \Lambda$.

Assume now $g_{p'} \neq 1$ and $g_{p'} \in CD$. Write $g_p = (v, m)$ where $v \in V$. Since $g_{p'}$ centralizes g_p , v is an eigenvector of $g_{p'}$ with eigenvalue 1. Since no nonidentity element of D has 1 as an eigenvalue, we have $g_{p'} \in CD - D$. Conjugating within U we may assume $g_{p'} = M(r) (\begin{smallmatrix} s & 0 \\ 0 & s \end{smallmatrix})$ where $(r, s) \in \Delta$. Since $1 \in \{rs, s\}$ and $(s^{-1}, s) \notin \Delta$ by definition of Δ , we must have $s = 1$ and $g = M(r) \in C$. As already observed, v is an eigenvector of $g_{p'}$ with eigenvalue 1, so $v = (s\nu, s)$ for some $s \in F$. Thus $g = M(r)(s\nu, s, m)$ for some $m \in GF(p)$. Conjugating g by $(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}) \in D$ (where $a\bar{a} = 1$) yields $M(r)(sav, sa, m)$, and so we may assume $s \in \Lambda$. Finally $(s\nu, s, m)$ belongs to the nonabelian group E_0 , and is conjugate within E_0 to $(s\nu, s, 0)$. Since E_0 centralizes $M(r)$, we may assume $m = 0$. Thus, g is conjugate to $M(r)(s\nu, s, 0)$ where $r \in N^\#$ and $s \in \Lambda$. The uniqueness of r and s follow from Lemmas 4.2 and 4.4(b)(ii) as $g_{p'} = M(r)$ and $g_p = (s\nu, s, 0)$.

(c) Assume finally $g_p \in PE - E$. In the group \tilde{G}/E , $g_{p'}E$ centralizes g_pE so $g_{p'}E \in DE/E \times PE/E$ and hence $g_{p'} \in DE$. As E normalizes the set $PE - E$, we may conjugate g by an appropriate element of E so as to assume $g_{p'} \in D$. Since $D \leq \mathbb{Z}(U)$, we may conjugate by an element of U to yield $g_p = (\begin{smallmatrix} 1 & \\ 0 & 1 \end{smallmatrix})e$ for some $e \in E$ and still preserve $g_{p'} \in D$.

Suppose first that $g_{p'} \neq 1$. Clearly, $g_{p'}$ centralizes $g_p = (\begin{smallmatrix} 1 & \\ 0 & 1 \end{smallmatrix})e$, and since $(\begin{smallmatrix} 1 & \\ 0 & 1 \end{smallmatrix}) \in U$, $g_{p'}$ centralizes $(\begin{smallmatrix} 1 & \\ 0 & 1 \end{smallmatrix})$. Hence $g_{p'}$ centralizes e . But the only elements of E centralized by any nonidentity element of D are contained in $\mathbb{Z}(E)$. Thus $g \in D \cdot P \cdot \mathbb{Z} \subseteq U\mathbb{Z}$ and there is nothing more to prove.

Assume finally that $g_p = 1$. Then $g = \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(v_1, v_2, m)$ for some $v_1, v_2 \in F$ and $m \in GF(p)$. If $v_1 = 0$, conjugate g by $(v_2, 0, 0)$ to yield $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(0, 0, m - \text{tr}(v_2\bar{v}_2)) \in UZ$. Assume then $v_1 \neq 0$. Since the trace map is surjective, $c \in F$ may be chosen so that

$$\text{tr}(\bar{v}_1c) = (-m + \text{tr}(v_2\bar{v}_2) - \text{tr}(v_1\bar{v}_2))/2.$$

Conjugating $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(v_1, v_2, m)$ by $(v_2, c, 0)$ yields $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(v_1, 0, 0)$ and then conjugating by $\begin{pmatrix} a & \\ 0 & a \end{pmatrix} \in D$ yields $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(av_1, 0, 0)$. Choose $a \in N$ so that $av_1 \in \Lambda$. Suppose now $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(s, 0, 0)$ is conjugate to $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(s', 0, 0)$ for some $s, s' \in \Lambda$. The conjugating element centralizes $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix} \pmod E$ and so belongs to DPE . Let $x = \begin{pmatrix} a & \\ 0 & a \end{pmatrix} \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} e$ be the conjugating element, where $e \in E$. Then $x^{-1} \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(s, 0, 0)x = \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(s', 0, 0)$ implies $as = s'$ and since $a \in N$, s must equal s' by the definition of Λ . \square

Lemma 4.4 completes the list of conjugacy classes of \tilde{G} . The centralizers of the various elements described in that lemma are straightforward to calculate. If $g = (0, 1, 0)$ then $C_{\tilde{G}}(g) = PC_E(g)$, while if $g = (s\nu, s, 0)$ for $s \in \Lambda$ then $C_{\tilde{G}}(g) = CC_E(s\nu, s, 0)$ where $|E: C_E(s\nu, s, 0)| = p$ since $s \neq 0$. For $g = M(r)(s\nu, s, 0)$, we have $C_{\tilde{G}}(g) = CC_{E_0}(s\nu, s, 0)$ where $|E_0: C_{E_0}(s\nu, s, 0)| = p$. Finally, for $g = \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(s, 0, 0)$ we have

$$C_{\tilde{G}}(g) = \left\{ \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}(ks, b, m) \mid k \in F_0, b \in F, m \in GF(p) \text{ and } \text{tr}(2s\bar{b} + ks\bar{s} - k^2s\bar{s}) = 0 \right\}.$$

Notice that for each given $k \in F_0$ there are $|F|/p = q^2/2$ choices for b for which $\begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}(ks, b, m) \in C_{\tilde{G}}(g)$, and so $|C_{\tilde{G}}(g)| = q \cdot q^2/p \cdot p = q^3$.

This verifies the second and third columns of Table 4.2. As a check, the dot product of these two columns equals $p(q+1)^2q^5(q-1) = |\tilde{G}|$.

The polynomials $\det(I - X\rho(g))$ will now be determined for each class representative g in the groups U and \tilde{G} . In the case of \tilde{G} , if $\zeta \in Z$ and $\rho(\zeta) = \varepsilon I$, then $\det(I - X\rho(g\zeta)) = \det(I - (\varepsilon X)\rho(g))$. In particular, $\det(I - X\rho(g))$ need only be calculated for class representatives of \tilde{G} given in Table 4.2 for $\zeta = 1$.

For any group G and any character χ of G , let $f_\chi(g, X)$ denote the polynomial $\det(I - X\sigma(g))$ where σ affords χ . The next lemma provides an effective method for computing $f_\chi(g, X)$.

LEMMA 4.5. *Let G be a group and L a subgroup of G . Assume $g \in L$.*

- (a) *If χ is a character of G such that χ_L decomposes into $\alpha + \beta$, say, then $f_\chi(g, X) = f_\alpha(g, X)f_\beta(g, X)$.*
- (b) *If γ and ψ are characters of L such that γ is a constituent of ψ , say $\psi = \gamma + \delta$, then*

$$f_\gamma(g, X) = f_\psi(g, X)/f_\delta(g, X).$$

- (c) *If θ is the regular character of $\langle g \rangle$, then $f_\theta(g, X) = 1 - X^l$ where l is the order of g .*

- (d) *If θ is the regular character of L , then $f_\theta(g, X) = (1 - X^l)^m$ where l is the order of g and $m = |L: \langle g \rangle|$.*

Proof. Let $n = \chi(1)$ and let σ be a representation affording χ . Then (a) follows readily from the observation that $f_\chi(g, X)$ is $(-1)^n X^n$ times the characteristic polynomial of $\sigma(g)$ evaluated at X^{-1} , and (b) follows from (a). If θ is the regular character of $\langle g \rangle$, then all of the l distinct l th roots of unity, say $1 = \varepsilon^0, \varepsilon, \varepsilon^2, \dots, \varepsilon^{l-1}$ where $l = o(g)$, occur as eigenvalues of $\sigma(g)$. Then $f_\theta(g, X) = \prod_{i=0}^{l-1} (1 - \varepsilon^i X) = 1 - X^l$, and (c) follows. If K is a subgroup of L then $\text{reg}_L|_K = |L: K| \text{reg}_K$. Part (d) follows from this and (c), using a repeated application of (a). \square

The last lemma, together with Theorems 2.5 and 2.6, is sufficient to calculate $\det(I - X\rho(g))$ for all the class representatives of U in Table 4.1 and for most of the

class representatives of \tilde{G} in Table 4.2 (for $\zeta = 1$). The most difficult case covered here is that of $g = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$. Let $a = |\langle g \rangle \cap C|$, $b = |\langle g \rangle \cap \tilde{C}|$ and $abk = o(g)$ (notice that a and b are necessarily relatively prime). We have $\text{reg}_{C\tilde{C}|\langle g \rangle} = (q+1)((q+1)/abk)$ $\text{reg}_{\langle g \rangle}$, and of course $1_{C\tilde{C}|\langle g \rangle} = 1_{\langle g \rangle}$. Also $\text{reg}_{C\tilde{C}/C|\langle g \rangle}$ may be identified with $\text{reg}_{C\tilde{C}/C|\langle g \rangle/C} = (q+1)/bk \text{reg}_{\langle g \rangle/C}$, as the index of $\langle g \rangle C$ in $C\tilde{C}$ is

$$|C\tilde{C} : \langle g \rangle C| = \frac{|C\tilde{C}| \cdot |\langle g \rangle \cap C|}{|\langle g \rangle| |C|} = \frac{(q+1)}{bk}.$$

Similarly, $\text{reg}_{C\tilde{C}/\tilde{C}|\langle g \rangle}$ may be identified with $(q+1)/ak \text{reg}_{\langle g \rangle \tilde{C}/\tilde{C}}$. This, together with Theorem 2.5 (c), verifies the polynomial $\det(I - X\rho(g))$ in Table 4.2 for $g = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$.

Again we observe that the argument given above which calculates $\det(I - X\rho(g))$ for $g = M(r) \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$ is valid in characteristic 2. Thus, all of Table 4.1 remains valid for q a power of 2.

The only case in Table 4.2 not covered by Lemma 4.5 and the theorems of § 2 is that of $g = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} (s, 0, 0)$. In this case, g is conjugate to $g\zeta^i$ for all i , where ζ is a generator of \mathbb{Z} . Hence, if ε is a primitive p th root of unity, then multiplication by ε induces a permutation on the set of eigenvalues of $\rho(g)$ (preserving multiplicity). When $p > 3$, g has order p and so $\det(I - X\rho(g)) = (1 - X^p)^{q^2/p}$. For $p = 3$, g need not have order p . In fact,

$$\left(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} (s, 0, 0) \right)^3 = (0, 0, -\text{tr}(s\bar{s})) \in \mathbb{Z}.$$

Let ω be the cube root of 1 defined by $\rho(0, 0, 1) = \omega I$. Then, the third power of any eigenvalue of $\rho(g)$ is $\omega^{-\text{tr}(s\bar{s})}$. Each of the three cube roots of $\omega^{-\text{tr}(s\bar{s})}$ occurs with the same multiplicity, namely, $q^2/p = q^2/3$, as an eigenvalue of $\rho(g)$. If δ is any one of these, then the other two are $\delta\omega$ and $\delta\omega^2$, and

$$\begin{aligned} \det(I - X\rho(g)) &= [(1 - \delta X)(1 - \delta\omega X)(1 - \delta\omega^2 X)]^{q^2/3} \\ &= (1 - \delta^3 X^3)^{q^2/3} \\ &= (1 - \omega^{-\text{tr}(s\bar{s})} X^3)^{q^2/3}. \end{aligned}$$

This verifies the last entry in Table 4.2.

The Molien series for U (for all characteristics) and \tilde{G} (for characteristic $\neq 2$) follow readily from Tables 4.1 and 4.2.

THEOREM 4.6. *The Molien series $\Phi_U(X)$ with respect to ρ is given by*

$$\begin{aligned} (q+1)^2 q(q-1) \Phi_U(X) &= \sum_{l|q+1} \frac{\phi(l)}{(1-X)(1-X^l)^{(q^2-1)/l}} \\ &+ \sum_{l|q+1} \frac{(q^2-1)\phi(l)(1-X^l)^{(q+1)/l}}{(1-X)(1-X^{pl})^{q(q+1)/pl}} \\ &+ \sum_{\substack{m|q^2-1 \\ m \nmid q+1}} \frac{q(q+1)\phi(m)/2}{(1-X)(1-X^m)^{(q^2-1)/m}} \\ &+ \sum_{\substack{(a,b,k) \\ a > b}} \frac{q(q-1)\phi(k)\phi(abk) \cdot (1-X^{ak})^{(q+1)/ak} (1-X^{bk})^{(q+1)/bk}}{(1-X^{abk})^{(q+1)^2/abk} (1-X)} \\ &+ \sum_{k|q+1} \frac{q(q-1)(\phi(k)^2 - \phi(k))/2}{(1-X)(1-X^k)^{(q^2-1)/k}}. \end{aligned}$$

The fourth sum extends over all triples of positive integers (a, b, k) where $a > b$, $\text{g.c.d.}(a, b) = 1$ and $abk | (q + 1)$.

Proof. The expression $(q + 1)^2 q (q - 1) \Phi_U(X)$ equals $\sum_{g \in U} 1/\det(I - X\rho(g))$ by definition, and is obtained in the theorem by grouping together equal terms as given by Table 4.1. The first three sums correspond directly to the first three rows of that table. The classes corresponding to the last row in Table 4.1 contribute

$$\sum_{(r,s) \in \Delta} \frac{q(q-1)(1-X^{ak})^{(q+1)/ak}(1-X^{bk})^{(q+1)/bk}}{(1-X^{abk})^{(q+1)^2/abk}(1-X)}$$

where a, b and k depend on (r, s) and have the same meaning as given in the table. Notice that by choice of Δ , $a \geq b$.

It remains to prove that the fourth summation appearing in the theorem corresponds to terms in the above sum for which $a > b$, and that the last summation corresponds to $a = b$.

Let $g = M(r) \binom{s}{0} \in C\check{C}$ and let (a, b, k) be the parameters associated with g . Thus $a = |A|$ and $b = |B|$ where $A = \langle g \rangle \cap C$ and $B = \langle g \rangle \cap \check{C}$. Moreover, $A \times B = AB$ is a subgroup of $\langle g \rangle$, and so is cyclic. Thus, $\text{g.c.d.}(a, b) = 1$. Notice that $\langle g \rangle / AB$ is cyclic of order k and intersects CAB/AB and $\check{C}AB/AB$ trivially. If $(r, s) \in \Delta$, then $a \geq b$, and if $a = b$, then $a = b = 1$.

It suffices to prove that the number of pairs $(r, s) \in \Delta$ corresponding to a fixed triple (a, b, k) is $\phi(k)\phi(abk)$ when $a > b$ and is $(\phi(k)^2 - \phi(k))/2$ when $a = b = 1$.

Toward this end, fix (a, b, k) with $\text{g.c.d.}(a, b) = 1$, $a \geq b$ and $abk | q + 1$. Let A and B denote the unique subgroups of orders a and b , respectively, of C and \check{C} . There are exactly $\phi(k)$ cyclic subgroups X/AB of $C\check{C}/AB$ of order k which intersect trivially with CAB/AB and $\check{C}AB/AB$, and each such X is necessarily cyclic. For any fixed X , there are $\phi(|X|) = \phi(abk)$ generators for X . When $a \neq b$, then any generator of X automatically belongs to $C\check{C} - D$, and when $a > b$, all of these generators correspond to elements of Δ . This accounts for the factor of $\phi(k)\phi(abk)$ appearing in the fourth summation of the theorem.

When $a = b = 1$, there are still $\phi(k)^2$ elements g of $\check{C}C$ corresponding to the triple $(1, 1, k)$. However, $\phi(k)$ of these belong to D . The remaining $\phi(k)^2 - \phi(k)$ in $\check{C}C - D$ fall in conjugate pairs, and so exactly half of these correspond to elements of Δ . This accounts for the factor of $(\phi(k)^2 - \phi(k))/2$ in the last summation, and the entire theorem is now proved. \square

Notice that the first and fifth summations appearing in Theorem 4.6 may be combined as a single sum.

A consequence of the proof of the last result is the combinatorial identity $\sum \phi(k)\phi(abk) = m^2$ where the sum extends over all triples of positive integers (a, b, k) satisfying $abk | m$ and $\text{g.c.d.}(a, b) = 1$. This is stated as [Amer. Math. Monthly, 88 (1981), p. 537, problem E2896].

THEOREM 4.7. *The Molien series $\Phi_{\check{C}}(X)$ with respect to ρ is given by*

$$pq^5(q+1)^2(q-1)\Phi_{\check{C}}(X) = \sum_{i=0}^{p-1} A_q(\epsilon^i X) + B_q(X) + C_q(X)$$

where ϵ is a primitive p th root of 1. Here

$$A_q(X) = 1/(1-X)^{q^2} + \sum_{\substack{l|q+1 \\ l \neq 1}} \phi(l)q^4/(1-X)(1-X^l)^{(q^2-1)/l} \\ + \sum_{\substack{l|q^2-1 \\ l \neq q+1}} \phi(l)q^5(q+1)/2(1-X)(1-X^l)^{(q^2-1)/l}$$

$$\begin{aligned}
 & + \sum_{\substack{a|q+1 \\ a \neq 1}} \phi(a)q^3(q-1)(1-X)^a/(1-X^a)^{q(q+1)/a} \\
 & + \sum_{k|q+1} ((\phi(k)^2 - \phi(k))/2)q^5(q-1)/(1-X)(1-X^k)^{(q^2-1)/k} \\
 & + \sum_{\substack{abk|q+1 \\ \text{g.c.d.}(a,b)=1 \\ a > b, bk > 1}} \frac{\phi(k)\phi(abk)q^5(q-1)(1-X^{ak})^{(q+1)/ak}(1-X^{bk})^{(q+1)/bk}}{(1-X)(1-X^{abk})^{(q+1)^2/abk}} \\
 & + q^2(q^2-1)(1-X)^q/(1-X^p)^{q(q+1)/p} \\
 & + \sum_{\substack{l|q+1 \\ l \neq 1}} \phi(l)q^4(q^2-1)(1-X^l)^{(q+1)/l}/(1-X)(1-X^{pl})^{q(q+1)/pl}; \\
 B_q(X) = & \sum_{\substack{l|q+1 \\ l \neq 1}} pq^3(q+1)(q-1)^2\phi(l)(1-X^p)^{q/p}/(1-X^{pl})^{q(q+1)/pl} \\
 & + p(q^4-1)/(1-X^p)^{q^2/p},
 \end{aligned}$$

and for $p > 3$,

$$C_q(X) = pq^2(q^2-1)^2/(1-X^p)^{q^2/p},$$

while for $p = 3$,

$$\begin{aligned}
 C_q(X) = & 3(q+1)^2q^2(q-1)\{(q/3-1)/(1-X^3)^{q^2/3} \\
 & + q/3(1-\omega X^3)^{q^2/3} + q/3(1-\omega^2 X^3)^{q^2/3}\}.
 \end{aligned}$$

Proof. The terms in $A_q(X)$ correspond to the class representatives appearing in the first seven rows of Table 4.2 for which $\zeta = 1$. The next three rows correspond to $B_q(X)$ and the last row to $C_q(X)$. The only terms which need explaining are the fourth and fifth summations appearing in $A_q(X)$. These terms originate from the class representatives in the fifth row of Table 4.2 corresponding to $\zeta = 1$. The natural index set for the classes is $\Delta_2 = \Delta - \Delta_1 = \Delta - N^\# \times \{1\}$. We have already seen in the proof of Theorem 4.6 that for a given triple (a, b, k) with $\text{g.c.d.}(a, b) = 1$, $abk|q+1$ and $a > b$, there are $\phi(k)\phi(abk)$ associated elements of Δ , while for $a = b = 1$ there are $(\phi(k)^2 - \phi(k))/2$. The set Δ_1 corresponds to triples of the form $(a, 1, 1)$ where $a > 1$. Thus, Δ_2 may be decomposed into $\Delta'_2 \cup \Delta''_2$, where Δ'_2 corresponds to all triples of the form $(1, 1, k)$ where $k > 1$, and Δ''_2 corresponds to all triples of the form (a, b, k) where $a > b$ and $bk > 1$. These correspond to the fourth and fifth summations of the theorem. \square

5. Molien series (characteristic 2). This section is concerned with the Molien series for G and G_0 (denoted $\Phi_G(X)$ and $\Phi_{G_0}(X)$) where it is assumed that F has characteristic 2. Recall that from § 3 we have $G = U \rtimes E$ and $G_0 = U\mathbb{Z} = U \dot{\times} \mathbb{Z}$ where $U = U(2, F)$ and $\mathbb{Z} = \mathbb{Z}(E)$. Since \mathbb{Z} is represented by $\pm I$ under ρ , we have

$$\Phi_{G_0}(X) = (\Phi_U(X) + \Phi_U(-X))/2.$$

Actually, the expression for $\Phi_U(X)$ has already been given by Theorem 4.6 when q is a power of 2, and so it remains only to calculate $\Phi_G(X)$.

The main difficulty with G in characteristic 2 stems from the definition of the action of U on E :

$$(v, m)^g = (vg, m + \text{tr}(\phi_g(v))).$$

For odd q , there is no term corresponding to $\phi_g(v)$, and in that case, the subset $V \times \{0\} \subseteq E$ was stabilized by U . This is no longer true in characteristic 2. Nevertheless, Lemmas 4.3 and 4.4 remain valid, although the proofs have to be slightly modified. Hence, the columns in Table 4.2 corresponding to the number and sizes of the conjugacy classes remain valid, as do the descriptions of the polynomials $\det(I - X\rho(g))$ for the first seven rows of that table. These rows correspond to elements of G which are conjugate to elements of UZ .

Let Γ and Δ have the same meanings that they had in the odd characteristic case. Since $F^\times = F_0^\times \times N$ in characteristic 2, the subgroup F_0^\times is a natural set of coset representatives for N in F^\times , and we may as well take $\Lambda = F_0^\times$. This is not necessary, but is obviously convenient.

LEMMA 5.1. *Let $u \in U$ and $1 \neq \zeta \in \mathbb{Z}$ so that $\zeta = (0, 0, 1)$. Then u is not conjugate to $u\zeta$ in G . Moreover, $\mathbb{C}_G(u) = \mathbb{C}_G(u\zeta)$ and Table 4.3 appearing in Lemma 4.3 remains valid for characteristic 2.*

Proof. We shall only prove that u is not conjugate to $u\zeta$ for any $u \in U$. Assume then that u is conjugate to $u\zeta$ for some u in U and, as in the proof of Lemma 4.3, we may assume the conjugating element is $e \in E$. Now $e^{-1}ue = u\zeta$ implies $ueu^{-1} = e\zeta$, and hence $u^2eu^{-2} = ue\zeta u^{-1} = e\zeta^2 = e$, so u^2 centralizes e . Clearly, u itself does not centralize e and hence $u \notin \langle u^2 \rangle$, proving that u has even order. Write $e = (v, m)$ so that

$$(vu^{-1}, m + \text{tr } \phi_{u^{-1}}(v)) = (v, m + 1).$$

Since $v \neq 0$, the element u^{-1} (and hence u) must have 1 for an eigenvalue. Therefore, u is conjugate either to an element of C or else to $y = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}$. Since $|C|$ has odd order and u has even order, u must be conjugate in U to y . Choose then $g \in U$ so that $g^{-1}ug = y$. Then y conjugates $g^{-1}eg$ to $g^{-1}eg\zeta$, and so we may as well assume $u = y$ and $g = 1$. By Theorem 3.2(3) we have $(v, m)^y = (vy, m + \text{tr } (\nu v_1 \bar{v}_1))$ where $v = (v_1, v_2)$ and $\nu + \bar{\nu} = 1$. However, $vy = v$ implies $v_1 = 0$, so $\text{tr } (\nu v_1 \bar{v}_1) = 0$, which contradicts $(v, m)^y = (v, m + 1)$. \square

LEMMA 5.2. *The conclusions of Lemma 4.4 are valid when $p = 2$.*

Proof. The conclusion in (a) did not require $p \neq 2$ and so remains valid.

For (b), assume $y \in G$ satisfies $g_2 \in E - \mathbb{Z}$. As in the earlier lemma, we may assume $g_{2'} \in CD$. If $g_{2'} = 1$ then $g = g_2 \in E - \mathbb{Z}$. The appropriate norm map $\eta: V \rightarrow F_0$ for characteristic 2 is given by $\eta(v) = v \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \bar{v}^T$. As in Lemma 4.4 (where a different η was used), η is invariant under U , and the orbits of U on $V^\#$ are $\eta^{-1}(a)$ for $a \in F_0^\times$ and $\eta^{-1}(0) - \{0\}$. The elements $(s\nu, s)$ for $s \in F_0^\times = \Lambda$ and $(0, 1)$ represent these orbits, and so g is conjugate to $(s\nu, s, m)$ for some unique $s \in F_0^\times$, or to $(0, 1, m)$, where $m \in GF(2)$. Since $(v, 0)$ is conjugate to $(v, 1)$ in E for any $v \in V^\#$, we may assume $m = 0$. This completes the case where $g_{2'} = 1$. If $g_{2'} \neq 1$, the corresponding proof given in Lemma 4.4 requires no changes.

Assume that $g_2 \in PE - E$. As in the earlier proof, we may assume $g_2 = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} (v_1, v_2, m)$. If $g_{2'} \neq 1$ there is no change in the argument, so assume $g_{2'} = 1$. If $v_1 = 0$, then when g is conjugated by $(v_2, 0, 0)$, the result is $\begin{pmatrix} 1 & \\ & 1 \end{pmatrix} (0, 0, m + \text{tr } (\nu v_2 \bar{v}_2)) \in UZ$. Hence assume $v_1 \neq 0$. Then $c \in F$ exists which satisfies

$$\text{tr } (\bar{v}_1 c) = m + \text{tr } (\nu v_2 \bar{v}_2),$$

and for this value of c we have

$$(v_2, c, 0)^{-1} \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} (v_1, v_2, m) (v_2, c, 0) = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} (v_1, 0, 0).$$

The remainder of the argument in Lemma 4.4 now applies, as $\phi_h(v) = 0$ for all $h \in D$. \square

The centralizers in G of the various conjugacy class representatives described by the last two lemmas are easy to work out and are omitted. There is no difference from the odd characteristic case.

At this point we know that the number and sizes of the conjugacy classes of G are the same as in Table 4.2 (columns 2 and 3), and that the polynomials $\det(I - X\rho(g))$ are correctly given for all elements g conjugate to elements of UZ (the first seven rows).

To complete the determination of $\Phi_G(X)$, the polynomials $\det(I - X\rho(g))$ are calculated for the elements $M(r)(s\nu, s, 0)$, $(0, 1, 0)$, $(s\nu, s, 0)$ and $\begin{pmatrix} 1 & \\ & 1 \end{pmatrix}(s, 0, 0)$ where $r \in N^\#$ and $s \in F_0^\times$.

Suppose first $g = M(r)(s\nu, s, 0) \in CE_0$. By Theorem 3.8, $\psi|_{CE_0} = (\text{reg}_C - 1_C) \# \psi_0$ where ψ_0 is the unique faithful irreducible character of E_0 . Now $g_2 = (s\nu, s, 0) \notin Z(E_0)$, and so $\psi_0(g_2^{\pm 1}) = 0$. If g_2 has order 2 then $\psi_0|_{\langle g_2 \rangle}$ is $(q/2)(\text{reg}_{\langle g_2 \rangle})$, while if g_2 has order 4, $\psi_0|_{\langle g_2 \rangle}$ is $q/2(\text{reg}_{\langle g_2 \rangle} - \text{reg}_{\langle g_2 \rangle/\mathbb{Z}})$. Let l denote the order of g_2 . In the first case $\psi|_{\langle g \rangle}$ is $(q(q+1)/2l) \text{reg}_{\langle g \rangle} - (q/2) \text{reg}_{\langle g \rangle/\langle g_2 \rangle}$, while in the second, $\psi|_{\langle g \rangle}$ is $(q(q+1)/2l) \text{reg}_{\langle g \rangle} - (q/2) \text{reg}_{\langle g \rangle/\langle g_2 \rangle} - (q(q+1)/2l) \text{reg}_{\langle g \rangle/\mathbb{Z}} + (q/2) \text{reg}_{\langle g \rangle/\langle g_2 \rangle\mathbb{Z}}$. In the first case $\det(I - X\rho(g))$ is $(1 - X^{2l})^{q(q+1)/2l} / (1 - X^2)^{q/2}$, while in the second case, the polynomial is

$$\frac{(1 - X^{4l})^{q(q+1)/2l} (1 - X^2)^{q/2}}{(1 - X^4)^{q/2} (1 - X^{2l})^{q(q+1)/2l}} = \frac{(1 + X^{2l})^{q(q+1)/2l}}{(1 + X^2)^{q/2}}$$

Since $(g_2)^2 = (s\nu, s, 0)^2 = (0, 0, \text{tr}(\nu s \bar{s}))$, the element g_2 can only have order 2 or 4, so that these are the only two possibilities. Notice that g has order $2l$ when $\text{tr}(\nu s \bar{s}) = 0$ and has order $4l$ when $\text{tr}(\nu s \bar{s}) = 1$. For $s \in \Lambda = F_0^\times$ we have $\text{tr}(\nu s \bar{s}) = \text{tr}^0(s \bar{s}) = \text{tr}^0(s^2) = \text{tr}^0(s)$, where $\text{tr}^0: F_0 \rightarrow GF(2)$ is the trace map. It follows that there are exactly $q(q/2 - 1)$ elements g of the form $M(r)(s\nu, s, 0)$ having order $2l$, where $r \in N^\#$ and $s \in \Lambda = F_0^\times$, and $q \cdot q/2$ elements of this form having order $4l$.

Suppose next that $g = (0, 1, 0)$. Then $g^2 = 1$ and since $\psi(g) = 0$, $\psi|_{\langle g \rangle}$ is $q^2/2$ times the regular character of $\langle g \rangle$. Hence $\det(I - X\rho(g)) = (1 - X^2)^{q^2/2}$.

If $g = (s\nu, s, 0)$ where $s \in F_0^\times$, then the calculation done earlier for $M(r)(s\nu, s, 0)$ is valid for $r = 1$ and so will not be repeated here. Assume then that $g = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}(s, 0, 0)$ where $s \in F_0^\times$. Using Corollary 3.2(c), we have

$$g^2 = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}(s, 0, 0) \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}(s, 0, 0) = (s, s, \text{tr}(\nu s \bar{s}))(s, 0, 0) = (0, s, \text{tr}(\nu s \bar{s})).$$

Hence $g^4 = 1$ and $g^2 \neq 1$ so g has order 4. Now $g^{-1} = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}(s, s, \text{tr}(\nu s \bar{s}))$, and by direct calculation $\begin{pmatrix} 1 & \\ & 1 \end{pmatrix}g \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} = g^{-1}$, so g is conjugate to g^{-1} . Suppose $\rho(g)$ has eigenvalue 1, -1 , i and $-i$ with multiplicities a , b , c and d . Since g is conjugate to g^{-1} , we know $c = d$. Moreover, the matrix $\rho(g^2)$ has trace $\psi(g^2) = \psi(0, s, \text{tr}(\nu s \bar{s})) = 0$ so $\rho(g^2)$ has $q^2/2$ eigenvalues equal to 1, and the remaining $q^2/2$ eigenvalues equal to -1 . Hence $a + b = q^2/2$ and $2c = c + d = q^2/2$. Clearly, the trace of $\rho(g)$ is $a - b + ci - di = a - b$. However, $\rho(g) = \rho \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \rho(s, 0, 0)$, and the trace of $\rho(g)$ is easily calculated by using Theorem 3.3. The result is $a - b = 0$, and so the numbers 1, -1 , i and $-i$ appear with the equal multiplicity $q^2/4$ as eigenvalues of $\rho(g)$. This leads easily to the formula $\det(I - X\rho(g)) = (1 - X^4)^{q^2/4}$.

The results above are tabulated in Table 5.1 opposite.

TABLE 5.1

Class representative g	Number of classes	Size of class	$\det(I - X\rho(g))$	Remarks
$M(r)(sv, s, 0), r \in N^\times, s \in F_0^\times$	$\begin{cases} q(q/2-1) \\ q(q/2) \end{cases}$	$2q^3(q+1)(q-1)$	$\begin{cases} (1-X^{2l})^{q(q+1)/2l} / (1-X^{2l})^{q/2} \\ (1+X^{2l})^{q(q+1)/2l} / (1+X^{2l})^{q/2} \end{cases}$	$\begin{cases} o(g) = 2/2(q+1), \text{tr}^0(s) = 0 \\ o(g) = 4/4(q+1), \text{tr}^0(s) = 1 \end{cases}$
$(0, 1, 0)$	1	$2(q+1)^2(q-1)$	$(1-X^{2l})^{q^2/2}$	$\begin{cases} \text{tr}^0(s) = 0 \\ \text{tr}^0(s) = 1 \end{cases}$
$(sv, s, 0), s \in \Lambda = F_0^\times$	$\begin{cases} q/2-1 \\ q/2 \end{cases}$	$2(q+1)q(q-1)$	$\begin{cases} (1-X^{2l})^{q^2/2} \\ (1+X^{2l})^{q^2/2} \end{cases}$	
$\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}(s, 0, 0), s \in \Lambda = F_0^\times$	$(q-1)$	$2(q+1)^2q^2(q-1)$	$(1-X^{4l})^{q^2/4}$	

Using the first seven rows of Table 4.2 and all of Table 5.1, the Molien series for G may now be written down for the characteristic 2 case. The result is contained in the following:

THEOREM 5.3. *The Molien series $\Phi_G(X)$ with respect to ρ in the characteristic 2 case is given by*

$$\begin{aligned} 2q^5(q+1)^2(q-1)\Phi_G(X) &= A_q(X) + A_q(-X) \\ &+ \sum_{\substack{l|q+1 \\ l \neq 1}} \frac{\phi(l) \cdot (q/2 - 1) \cdot 2q^3(q+1)(q-1)(1-X^2)^{q/2}}{(1-X^{2l})^{q(q+1)/2l}} \\ &+ \sum_{\substack{l|q+1 \\ l \neq 1}} \frac{\phi(l) \cdot (q/2) \cdot 2q^3(q+1)(q-1)(1+X^2)^{q/2}}{(1+X^{2l})^{q(q+1)/2l}} \\ &+ 2(q+1)^2(q-1)/(1-X^2)^{q^2/2} \\ &+ (q/2 - 1) \cdot 2(q+1)q(q-1)/(1-X)^{q^2/2} \\ &+ (q/2) \cdot 2(q+1)q(q-1)/(1+X^2)^{q^2/2} \\ &+ 2(q+1)^2q^2(q-1)^2/(1-X^4)^{q^2/4} \end{aligned}$$

where $A_q(X)$ is defined as in Theorem 4.7.

6. Finite extensions of the unimodular subgroup of $\rho(G)$. Let G and ρ denote the group and representation constructed in § 2. Thus

$$G = \langle -1 \rangle \times \tilde{G}$$

where

$$\tilde{G} = U(2, F) \ltimes E$$

and $|F| = q^2$ is a power of the odd prime p . Let δ denote the unique sign character of \tilde{G} ($|\tilde{G} : \ker \delta| = 2$). By Theorem 2.2 we have $\det \rho|_{\tilde{G}} = \delta$. Since $\rho(-1) = -I$, the unimodular subgroup of $\rho(G)$ is

$$\ker \delta \cup (-1) \cdot (\tilde{G} - \ker \delta).$$

Let G^0 denote this subgroup of G . Hence $G^0 = \ker(\det \rho)$ and $|G : G^0| = 2$.

As an abstract group, G^0 is isomorphic to \tilde{G} . (In fact, $\rho(g) \mapsto \delta(g)\rho(g)$ is an isomorphism from $\rho(\tilde{G})$ to $\rho(G^0)$.) In particular, E is complemented in G^0 by a group U^0 which is isomorphic to $U(2, F)$, and the action of U^0 on E is the same as that of $U(2, F)$.

As outlined in [3, § 8], the full automorphism group of E which centralizes $\mathbb{Z}(E)$ is a split extension of the symplectic group $Sp(4n, p)$ by the inner automorphisms of E , where $q^2 = p^{2n}$. The semidirect product $G_1 = Sp(4n, p) \ltimes E$ may be formed, and we may assume $G_1 \cong U^0 E = G^0$. The representation $\rho|_{G^0}$ extends to a unimodular representation of G_1 which we continue to denote by ρ . As in [3], let $G_2 = G_1 \mathbb{Y} \mathbb{Z}_{q^2}$ where \mathbb{Z}_{q^2} is a cyclic group of order q^2 , and the subgroup of order p in \mathbb{Z}_{q^2} is amalgamated with the center of G_1 in the central product. Finally, ρ may be extended to a unimodular representation of G_2 by requiring \mathbb{Z}_{q^2} to be represented by scalars.

It will turn out (Theorem 6.12 below) that the linear group $\rho(G_2)$ is the unique maximal finite subgroup of $SL(q^2, \mathbb{C})$ containing $\rho(G^0)$. For convenience, throughout the remainder of this section the groups G^0 and G_2 will be identified with the linear group $\rho(G^0)$ and $\rho(G_2)$.

The proof of the next lemma is similar to that of Lemmas I.8.2 and I.8.3 and will be omitted.

LEMMA 6.1. *If \mathbb{Z} denotes the group of scalar matrices in $GL(q^2, \mathbb{C})$, then*

$$\begin{aligned} \mathbb{C}_{GL(q^2, \mathbb{C})}(E/\mathbb{Z}(E)) &= E\mathbb{Z}, \\ \mathbb{C}_{SL(q^2, \mathbb{C})}(E/\mathbb{Z}(E)) &= E\mathbb{Z}_{q^2}, \\ \mathbb{N}_{SL(q^2, \mathbb{C})}(E) &= G_2. \end{aligned}$$

LEMMA 6.2. *In the action of $U(2, F)$ on $V = F \times F$, the subgroups of V stabilized by $S = SL(2, F_0)$ are $\{0\}$, V and the $q + 1$ subgroups of order q^2 of the form cV_0 where $c \in F^\times$ and $V_0 = F_0 \times F_0 \cong V$. These $q + 1$ subgroups are transitively permuted by $U(2, F)$.*

Proof. Clearly $\{0\}$, V and cV_0 are stabilized by S . Notice that S is transitive on $V_0^\#$, and hence each subgroup cV_0 for $c \neq 0$ is irreducible under S . Thus for $c, d \in F^\times$ and $cV_0 \neq dV_0$, we have $V = cV_0 + dV_0$, and V has S -composition length 2. Let W be any S -subgroup of V where $0 < W < V$. Then W is necessarily irreducible under S . Let $(\alpha, \beta) \in W^\#$. As $\begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \in S$, we have $(\alpha, \alpha + \beta) \in W$, and this leads easily to $(0, \beta) \in W$ and $(\alpha, 0) \in W$. If $\alpha \neq 0$, then $(\alpha, 0) \in W \cap \alpha V_0$, so $W = \alpha V_0$. Otherwise, if $\alpha = 0$, then $\beta \neq 0$, leading to $W = \beta V_0$. Hence $W = cV_0$ for some $c \in F^\times$. Clearly, if $c, d \in F^\times$, then $cV_0 = dV_0$ if and only if $cF_0^\times = dF_0^\times$, and there are exactly $|F^\times : F_0^\times| = q + 1$ subgroups of this form.

The subgroup

$$H = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \mid a \in F^\times \right\} \cong U(2, F)$$

normalizes S , and so H permutes the sets $\{cV_0 \mid c \in F^\times\}$. As H acts transitively on the vectors $\{(a, 0) \mid a \in F^\times\}$, it follows that H transitively permutes the subgroups cV_0 for $c \in F^\times$. This proves that $U(2, F)$ is transitive, and the proof of Lemma 6.2 is complete. \square

COROLLARY 6.3. *There are exactly $q + 1$ proper subgroups of E strictly containing $\mathbb{Z}(E)$ which are invariant under the action of $SL(2, F_0)$. Moreover, these are all contained in a single orbit under the action of U^0 , and each of these is nonabelian.*

Proof. The subgroup V_0 of V corresponds to

$$E_0 = \{(a, b, m) \mid a, b \in F_0, m \in GF(p)\} \cong E.$$

The commutator of (a, b, m) with (a', b', m') is

$$(0, 2 \operatorname{tr}(ab' - ba')) = (0, 4 \operatorname{tr}^0(ab' - ba'))$$

where $\operatorname{tr}^0 : F_0 \rightarrow GF(p)$ is the trace map. Since this last expression is not identically zero, E_0 is nonabelian. The rest of the corollary follows from Lemma 6.2. \square

Notice that the conclusion of Corollary 6.3 also holds for the group $E\mathbb{Z}_{q^2}$ with $\mathbb{Z}(E)$ replaced by $\mathbb{Z}(E\mathbb{Z}_{q^2}) = \mathbb{Z}_{q^2}$.

The proofs of the next two lemmas are similar to that of Lemma I.8.4 and Theorem I.8.5 with S replaced by U^0 , and will be omitted.

LEMMA 6.4. *Let X be a finite subgroup of $SL(q^2, \mathbb{C})$ containing G^0 , and assume that $E \cong O_p(X)$. Then $O_p(X) \cong E\mathbb{Z}_{q^2}$ and $E \cong X$. In particular, $X \cong G_2$.*

LEMMA 6.5. *If X is a finite subgroup of $SL(q^2, \mathbb{C})$ containing G^0 and if the Fitting subgroup $\mathbb{F}(X)$ is not contained in $\mathbb{Z}(X)$, then $X \cong G_2$.*

As in [3], it is natural to consider the following situation.

Hypothesis 6.6. X is a finite subgroup of $SL(q^2, \mathbb{C})$ containing G^0 , and $\mathbb{F}(X) \cong \mathbb{Z}(X)$.

LEMMA 6.7. *Assume Hypothesis 6.6 holds for X . Then the generalized Fitting subgroup $\mathbb{F}^*(X)$ has the form $\mathbb{Z}(X)Y$ where Y is quasisimple, and $E \leq Y$.*

Proof. By definition, $\mathbb{F}^*(X) = \mathbb{F}(X)\mathbb{E}(X)$ where $\mathbb{E}(X)$ is the join of all the quasisimple subnormal subgroups of X , say $\mathbb{E}(X) = Y_1 Y_2 \cdots Y_{t'}$. Hypothesis 6.6 implies $\mathbb{F}(X) = \mathbb{Z}(X)$, and so it remains to prove $t = 1$ and $E \leq Y_1$.

Assume that $S = SL(2, F_0)$ does not normalize some Y_i , say Y_1 , and choose notation so that $\{Y_1, Y_2, \dots, Y_{t'}\}$ is the S -orbit containing Y_1 . As $Y_1 Y_2 \cdots Y_{t'}$ is a subnormal subgroup of X , an irreducible constituent of ρ restricted to $Y_1 Y_2 \cdots Y_{t'}$ has degree dividing q^2 and is a tensor product of t' representations. This implies $p^{t'} \leq q^2$. Since t' is the index of a proper subgroup of $SL(2, F_0)$, we have by a theorem of Galois [8, Satz 8.28, p. 241] either $t' \geq q$ or $t' = 6$ and $q = 9, p = 3$. Either case leads to a contradiction. Hence, S normalizes each Y_i , or equivalently, S is in the kernel of the action of X on $\{Y_1, Y_2, \dots, Y_{t'}\}$. Therefore, $E = [E, S]$ is also in this kernel, and E normalizes each Y_i .

Define $R_i = Y_i \mathbb{Z}_{q^2} \cap E \mathbb{Z}_{q^2}$. Thus each R_i is an S -invariant subgroup of $E \mathbb{Z}_{q^2}$ containing \mathbb{Z}_{q^2} .

Suppose there exists i such that $R_i = \mathbb{Z}_{q^2}$. Let $\bar{E} = E \mathbb{Z}_{q^2} / \mathbb{Z}_{q^2}$ and $\bar{Y}_i = Y_i \mathbb{Z}_{q^2} / \mathbb{Z}_{q^2}$, and let \bar{K}_i be the subgroup of \bar{Y}_i containing \mathbb{Z}_{q^2} which satisfies

$$\bar{K}_i = K_i / \mathbb{Z}_{q^2} = \mathbb{N}_{\bar{Y}_i}(\bar{E}).$$

Then $[K_i, \bar{E}] = [\mathbb{N}_{\bar{Y}_i}(\bar{E}), \bar{E}] \leq \bar{Y}_i \cap \bar{E} = (Y_i \mathbb{Z}_{q^2} \cap E \mathbb{Z}_{q^2}) / \mathbb{Z}_{q^2} = \bar{1}$. Hence, \bar{K}_i centralizes \bar{E} so that $K_i \leq \mathbb{C}_{SL(q^2, \mathbb{C})}(E \mathbb{Z}_{q^2} / \mathbb{Z}_{q^2}) = \mathbb{C}_{SL(q^2, \mathbb{C})}(E / \mathbb{Z}(E)) = E \mathbb{Z}_{q^2}$, where the last equality follows from Lemma 6.1. As $K_i \leq Y_i \mathbb{Z}_{q^2}$, we have $K_i \leq Y_i \mathbb{Z}_{q^2} \cap E \mathbb{Z}_{q^2} = R_i = \mathbb{Z}_{q^2}$, proving $\bar{K}_i = \bar{1}$. Hence, $\mathbb{N}_{\bar{Y}_i}(\bar{E}) = \bar{1}$, and this implies that $p \nmid |\bar{Y}_i|$. Thus, p does not divide the order of the Schur multiplier of \bar{Y}_i , and so $\bar{Y}_i = Y_i \mathbb{Z}_{q^2} / \mathbb{Z}_{q^2} = (Y_i \times \mathbb{Z}_{q^2}) / \mathbb{Z}_{q^2} \cong Y_i$. The linear group SEY_i is irreducible (as E is irreducible) and $Y_i \cong SEY_i$. Therefore, all irreducible constituents of $\rho|_{Y_i}$ have degree dividing q^2 . If σ is one of these then the degree of σ is a power of p . However, this degree must also divide $|Y_i|$ which is prime to p , and hence σ has degree 1. As $Y_i' = Y_i$, σ is the principal character, and this means that Y_i is in the kernel of ρ , which is a contradiction as X is a faithful linear group.

The preceding paragraph shows that $R_i = Y \mathbb{Z}_{q^2} \cap E \mathbb{Z}_{q^2} > \mathbb{Z}_{q^2}$ for all i . Suppose $\mathbb{Z}_{q^2} < R_i < E \mathbb{Z}_{q^2}$ for some i . Then $\mathbb{Z}_{q^2} < \mathbb{C}_{E \mathbb{Z}_{q^2}}(R_i) < E \mathbb{Z}_{q^2}$ and both R_i and $\mathbb{C}_{E \mathbb{Z}_{q^2}}(R_i)$ are S -invariant. By Corollary 6.3 and the remarks following that corollary, there exists $x \in U^0$ such that R_i^x is different from R_i and $\mathbb{C}_{E \mathbb{Z}_{q^2}}(R_i)$. Now $Y_i^x = Y_j$ for some j , and since $R_i^x \neq R_i$ we have $i \neq j$. Hence $[R_i, R_i^x] \leq [Y_i \mathbb{Z}_{q^2}, Y_j \mathbb{Z}_{q^2}] = [Y_i, Y_j] = 1$ so $R_i^x \leq \mathbb{C}_{E \mathbb{Z}_{q^2}}(R_i)$ and equality must hold as these groups have the same order. But this contradicts the choice of x , and proves that $R_i = E \mathbb{Z}_{q^2}$ for all i . Hence $E \mathbb{Z}_{q^2} \leq Y_i \mathbb{Z}_{q^2}$ for all i . Therefore, $E = [E, S] \leq [E \mathbb{Z}_{q^2}, S] \leq [Y_i \mathbb{Z}_{q^2}, S] = [Y_i, S] \leq Y_i$, and we have $E \leq Y_i$ for all i . Since E is nonabelian and Y_i centralizes Y_j for $i \neq j$ we must have $t = 1$. This completes the proof of Lemma 6.7. \square

The proof of the main result of this section (Theorem 6.12) will follow once it is shown that $\mathbb{F}^*(X)$ does not have the form given by Lemma 6.7 for every choice of the simple group $\bar{Y} = Y/\mathbb{Z}(Y)$.

The next result, which is an improvement of Lemma 6.7, uses the Schreier ‘‘conjecture’’. Its proof is similar to that of Lemma I.8.8 and will be omitted. Recall that S is the normal subgroup of U^0 isomorphic to $SL(2, F_0)$.

LEMMA 6.8. *Assume Hypothesis 6.6 holds for X . Then $\mathbb{F}^*(X) = \mathbb{Z}(X)Y$ where Y is quasisimple and contains E . If $q^2 > 9$ or if $q^2 = 9$ and $\text{Aut}(\bar{Y})/\text{Inn}(\bar{Y})$ has order prime to 3, where $\bar{Y} = Y/\mathbb{Z}(Y)$, then $ES \leq Y$.*

LEMMA 6.9. *Let K denote a field of characteristic 0 or a finite field of characteristic $\neq p$. Then*

- (a) *Any faithful representation of E over K has degree at least q^2 .*
- (b) *Any faithful representation of $ES/\mathbb{Z}(E)$ over K has degree at least $q^2 - 1$.*

Proof. Without loss of generality, we may assume that the field K is a splitting field for all subgroups of E and $ES/\mathbb{Z}(E)$. Since E is extraspecial, part (a) is immediate, as every irreducible representation of E is either linear or has degree q^2 .

The group $U(2, F)$ permutes the vectors of $V = F \times F$, and the nonzero vectors of V fall into orbits of the form $\eta^{-1}(a)$ for $a \in F_0^\#$ and $\eta^{-1}(0) - \{0\}$. Here η denotes the map $V \rightarrow F_0$ given by

$$\eta(v) = \nu v \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \bar{v}^T$$

(as in the proof of Lemma 4.4). A representative of $\eta^{-1}(a)$ for $a \neq 0$ may be chosen of the form $(s\nu, s)$, and since C centralizes $(s\nu, s)$ and $U(2, F) = C \cdot SL(2, F_0)$, S is transitive on $\eta^{-1}(a)$. For $a = 0$, $\eta^{-1}(0) - \{0\}$ consists of the nonzero isotropic vectors. These are precisely the elements of $\cup_{c \in F^\times} (cV_0)^\#$, where $V_0 = F_0 \times F_0$. As S acts transitively on the sets $(cV_0)^\#$, the set $\eta^{-1}(0) - \{0\}$ decomposes into a union of $q + 1$ orbits under S , each of size $q^2 - 1$. As $|\eta^{-1}(a)| = (q + 1)q(q - 1)$, it follows that every orbit of S on $V^\#$ has cardinality at least $q^2 - 1$.

Since the commutator map gives a bilinear pairing of $E/\mathbb{Z}(E) \simeq V$ with itself, we have $V \simeq \hat{V}$ as S -modules, and hence every orbit of S on the nonprincipal characters of $E/\mathbb{Z}(E) \simeq V$ has cardinality at least $q^2 - 1$. This also applies to Brauer characters over a field of characteristic $\neq p$, and Lemma 6.9(b) now follows by Clifford's theorem. \square

If Y satisfies the conclusion of Lemma 6.7, then $\bar{Y} = Y/\mathbb{Z}(Y)$ is simple and $\mathbb{Z}(E) \leq \mathbb{Z}(Y) \leq Y'$ so that p divides the order of the Schur multiplier of \bar{Y} . The simple groups having a Schur multiplier divisible by an odd prime p are listed below (see [6]):

$$\begin{aligned} &PSL(n, r) \quad \text{where } p|(n, r - 1) \quad \text{or} \quad (n, r) = (2, 9), \\ &PSU(n, r) \quad \text{where } p|(n, r + 1) \quad \text{or} \quad (n, r) = (4, 3), \\ &\overline{E_6(r)} \quad \text{where } p = 3|(r - 1), \\ &{}^2\overline{E_6(r)} \quad \text{where } p = 3|(r + 1). \end{aligned}$$

A bar is used to denote the quotient of the universal Chevalley group (or its twisted type) by its center. In addition to these four infinite families, there are 10 other exceptional groups, and in each of these cases $p = 3$:

$$\begin{aligned} &A_7, \quad M_{22}, \quad J_3, \quad O'S, \\ &G_2(3), \quad \text{McL}, \quad \text{Suz}, \\ &SO(7, 3), \quad \text{Fi}_{22} = M(22), \quad \text{Fi}'_{24} = M(24)'. \end{aligned}$$

LEMMA 6.10. *Let X satisfy Hypothesis 6.6 and let Y be as in Lemma 6.7. Then $\bar{Y} = Y/\mathbb{Z}(Y)$ is not one of the groups $PSL(n, r)$, $PSU(n, r)$, $\overline{E_6(r)}$ or ${}^2\overline{E_6(r)}$.*

Proof. Suppose first that \bar{Y} is either $PSL(n, r)$ or $PSU(n, r)$ where $p \nmid r$. The group Y is a homomorphic image of the covering group of \bar{Y} which is either $SL(n, r)$ except for $(n, r) = (3, 4)$, or $SU(n, r)$ except for $(n, r) = (6, 2)$. The Schur multiplier of $PSL(3, 4)$ is $\mathbb{Z}_3 \times \mathbb{Z}_4 \times \mathbb{Z}_4$ while that of $PSU(6, 2)$ is $\mathbb{Z}_3 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. Since $|\mathbb{Z}(Y)|$ is odd, Y is necessarily a homomorphic image of either $SL(3, 4)$ or $SU(6, 2)$ in these exceptional cases, even though these last two groups are not covering groups of \bar{Y} .

Let L denote either $SL(n, r)$ or $SU(n, r)$ so that Y is a homomorphic image of L . Then there exists $W \leq Z(L)$ such that $L/W \cong Y$ and $p \mid |Z(L):W|$. Choose $W_0 \leq W$ so that $p \nmid |W_0|$, and $|W:W_0|$ is a power of p . As $E \leq Y$ and $E' = Z(E) \leq Z(L)$, there is a subgroup \tilde{E} of L containing W such that $\tilde{E}/W \cong E$ and $(\tilde{E}/W)' = Z(\tilde{E}/W)$ has order p in $Z(L)/W$. Denote this subgroup by W_1/W . Hence \tilde{E}' covers $W_1 \bmod W$.

Since $|W_1:W_0|$ is a power of p and W_1/W_0 is cyclic, \tilde{E}' must cover $W_1 \bmod W_0$. However, $\tilde{E}/Z(\tilde{E})$ has exponent p , and hence so does \tilde{E}' . Thus W_1/W_0 has order p and this implies $W_0 = W$. Therefore, $p \nmid |W|$. Since $(|W|, |\tilde{E}:W|) = 1$, W is complemented in \tilde{E} and hence E is isomorphic to a subgroup of L . Now L is a linear group of degree n over a finite field of characteristic $\neq p$ (which is in fact a splitting field for all subgroups of L), and the embedding of E in L gives a faithful representation of E of degree n . By Lemma 6.9(a) we have $q^2 \leq n$. In particular, $n \geq 9$.

Suppose we are in the case $L = SL(n, r)$. By an argument appearing in the proof of Lemma I.8.9, every representation of Y has degree at least $r^{n-1} - 1$, so $r^{n-1} - 1 \leq q^2 \leq n$. As $p \mid (r-1)$ we have $4^{n-1} - 1 \leq n$, so $n \leq 1$, a contradiction.

Suppose then that $L = SU(n, r)$. Let $k = \lceil n/2 \rceil$ so that $k \geq 4$. By another argument appearing in Lemma I.8.9 again, $(r^2)^{k-1} - 1 \leq q^2 \leq n$. Hence $2^{2k-2} - 1 \leq n = 2k$ or $2k + 1$, and this leads to the contradiction $n \leq 5$.

If \tilde{Y} is $PSL(n, r)$ where $p \mid r$, then $p = 3$ and $(n, r) = (2, 9)$, so that $\tilde{Y} \cong PSL(2, 9) \cong A_6$. This case is easily eliminated as $|E/Z(E)|$ does not divide $|A_6|$. If \tilde{Y} is $PSU(n, r)$ where $p \mid r$, then $\tilde{Y} \cong PSU(4, 3)$, and this case will be considered last.

Suppose now \tilde{Y} is $E_6(r)$ or ${}^2E_6(r)$. Except for ${}^2E_6(2)$, the Schur multiplier of these groups is cyclic of order 3 and $Y \cong E_6(r)$ or ${}^2E_6(r)$ except possibly when $r = 2$ in the twisted case. However, ${}^2E_6(2)$ has a Schur multiplier isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, and since $|Z(Y)|$ is odd, Y must be isomorphic to ${}^2E_6(2)$ in this exceptional case. Notice that in all cases $p = 3$.

The group $E_6(r)/Z(E_6(r))$ acts faithfully on the 78-dimensional Lie algebra of type E_6 over $GF(r)$. As ${}^2E_6(r) \leq E_6(r^2)$, ${}^2E_6(r)/Z({}^2E_6(r))$ also has a 78-dimensional representation, although the field is $GF(r^2)$. By Lemma 6.8, if $q^2 > 9$ then $q^2 - 1 \leq 78$. However, $p = 3$ so q is a power of 3 and so $q^2 = 9$. Thus, either $E_6(r)$ or ${}^2E_6(r)$ has a complex representation of degree 9.

In the case of $E_6(r)$, we have $SL(6, r) = A_5(r) \leq E_6(r)$, and the smallest degree of any faithful complex representation of $SL(6, r)$ is at least $r^5 - 1$. Thus, $r^5 - 1 \leq q^2 = 9$, a contradiction. For ${}^2E_6(r)$, we have $SL(3, r^2) \leq SU(6, r) = {}^2A_5(r) \leq {}^2E_6(r)$, and this leads to the contradiction $r^4 - 1 \leq 9$.

Assume finally that $\tilde{Y} \cong PSU(4, 3)$ and hence that Y is a 3-fold cover of $PSU(4, 3)$. The highest power of 3 dividing the order of \tilde{Y} is 3^6 , and hence $q^2 = 9$. Moreover, 3 does not divide the outer automorphism group of $PSU(4, 3)$, so SE embeds as a subgroup of Y . Let $V = E/Z(E)$ so that SV embeds as a subgroup of $\tilde{Y} = PSU(4, 3)$. A contradiction will be reached by showing that this last embedding is impossible.

The group $U(4, 3)$ may be described as $\{M \in GL(4, 9) \mid M\bar{C}M^T = C\}$ where $C = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$, I is the 2×2 identity matrix, and $\bar{}$ is the automorphism of $GF(9)$ fixing $GF(3)$. It is not hard to check that matrices of the form $\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}$ where $A = \begin{pmatrix} a & \alpha \\ 0 & b \end{pmatrix}$, for $a, b \in GF(3)$ and $\alpha \in GF(9)$, form an abelian subgroup of $SU(4, 3)$. Let \tilde{V} denote the image of this subgroup in $PSU(4, 3)$. We also have that matrices of the form

$$M = \begin{pmatrix} X & 0 \\ 0 & (\bar{X}^T)^{-1} \end{pmatrix}$$

belong to $U(4, 3)$ for all X in $GL(2, 9)$. Let $SL^\pm(2, 9) = \{X \in GL(2, 9) \mid \det X = \pm 1\}$.

Now $M \in SU(4, 3)$ if and only if $X \in SL^\pm(2, 9)$. Let H denote the image of

$$\left\{ \begin{pmatrix} X & 0 \\ 0 & (\bar{X}^T)^{-1} \end{pmatrix} \middle| X \in SL^\pm(2, 9) \right\}$$

in $PSU(4, 3)$. Notice that when $X = iI$ (where $i^2 = -1$), then $\det X = -1$ and M is a scalar matrix. Hence $H \simeq SL^\pm(2, 9)/\langle iI \rangle \simeq SL(2, 9)/\langle -I \rangle = PSL(2, 9) (\simeq A_6)$. By direct calculation H normalizes \tilde{V} , and the full normalizer of \tilde{V} in $PSU(4, 3)$ is $H\tilde{V}$.

Let P_0 be the subgroup of H corresponding to $X = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix}$ for $\beta \in GF(9)$, and set $P = P_0\tilde{V}$. Thus P is a Sylow 3-subgroup of $PSU(4, 3)$, and direct computation shows

$$P' = [P_0, \tilde{V}] = \left\{ \begin{pmatrix} I & a & \alpha \\ & \bar{a} & 0 \\ 0 & & I \end{pmatrix} \middle| a \in GF(3), \alpha \in GF(9) \right\}$$

as well as

$$C_P(P') = \tilde{V}.$$

The group SV embeds as a subgroup of $PSU(4, 3)$. Let Q be a Sylow 3-subgroup of SV (hence $V \leq Q$). Replacing SV by a conjugate if necessary, we may assume $Q \leq P$. By consideration of orders, we have $|P:Q| = 3$, and hence $Q \trianglelefteq P$. As V is the unique abelian subgroup of Q having index 3, V is characteristic in Q and hence is normal in P . As $|P:V| = 9$, we have $V \cong P'$, and as V is abelian, we have $C_P(P') \cong V$. From the previous paragraph, then, $V = \tilde{V}$.

As S normalizes $V = \tilde{V}$, S embeds as a subgroup of $H\tilde{V}$. However, a Sylow 2-subgroup of S is Q_8 , while that of $H\tilde{V}$ is D_8 . This final contradiction completes the proof of Lemma 6.10. \square

LEMMA 6.11. *Let X satisfy Hypothesis 6.6 and let Y be as in Lemma 6.7. Then $\bar{Y} = Y/Z(Y)$ is not one of the groups $A_7, M_{22}, J_3, O'S, G_2(3), \text{McL}, \text{Suz}, SO(7, 3), \text{Fi}_{22} = M(22), \text{Fi}'_{24} = M(24)$.*

Proof. Suppose \bar{Y} is one of these ten groups. In each of these cases, $p = 3$, and hence q is a power of 3. By Lemma 6.7, $E \trianglelefteq Y$ and hence $E/Z(E)$ embeds in \bar{Y} . In particular $q^4 \mid |\bar{Y}|$, and this quickly eliminates the groups A_7 and M_{22} .

If $q^2 = 9$ then Y is a linear group of degree 9. Since $|Z(Y)|$ is odd, $|Y|$ can have no prime divisor greater than 11 (by [2, Thm. 1]), and hence \bar{Y} must be McL .

If $q^2 > 9$ then $ES/Z(E)$ embeds as a subgroup of \bar{Y} and hence $q^5 \mid |\bar{Y}|$, and so $3^{10} \mid |\bar{Y}|$. Hence \bar{Y} is Fi'_{24} and q^2 is either 81 or 729.

Suppose first $q^2 = 9$ and $\bar{Y} = \text{McL}$. Now McL contains a subgroup isomorphic to $PSU(4, 3)$ which in turn contains a Sylow 3-subgroup of McL , say P . Replacing $ES/Z(E)$ by a conjugate if necessary, we may assume that a Sylow 3-subgroup of $ES/Z(E)$ is contained in P . By the argument at the end of the proof of the preceding lemma, $E/Z(E)$ is then $C_P(P')$, and the normalizer of $E/Z(E)$ in McL contains a subgroup isomorphic to A_6 which acts faithfully and irreducibly on $E/Z(E)$. Since the dimension of this module is only 4, it must be absolutely irreducible. The commutator map of E may be used to construct a nondegenerate skew-symmetric form on $E/Z(E)$ that is invariant under the action of A_6 .

Up to isomorphism, A_6 has only one absolutely irreducible module of dimension 4 over a field of characteristic 3. This module is M/N where M and N are the unique maximal and minimal submodules of the standard 6-dimensional permutation module for A_6 . The permutation module supports a nondegenerate symmetric form stabilized by A_6 , and the radical of M with respect to this form is N . Therefore, M/N also

supports a symmetric and nondegenerate form stabilized by A_6 . However, $M/N \simeq E/\mathbb{Z}(E)$ as A_6 -modules. This is a contradiction, as no absolutely irreducible module for a group can support both a nondegenerate symmetric and a skew-symmetric form that is invariant under the group.

Suppose now \bar{Y} is Fi_{24} where $q^2 = 81$ or 729 . Since $Fi_{22} \leq Fi_{23}$ and the Schur multiplier of Fi_{23} is not divisible by 3, Fi_{22} appears as a subgroup of Y . Let ψ be the character of the linear group Y so that $\psi(1) = 81$ or 729 . The desired contradiction will be reached by considering the character $\psi|_{Fi_{22}}$.

The entire character table of Fi_{22} is known and listed in [7]. Suppose first $\psi(1) = 81$. Then $\psi|_{Fi_{22}} = 3 \cdot 1_{Fi_{22}} + \chi_{78}$ where χ_{78} is the unique irreducible character of Fi_{22} of degree 78. A Sylow 5-subgroup of Fi_{22} is also one for Y , and the character table of Fi_{22} implies that all elements of order 5 in Fi_{22} are conjugate. Now ES embeds in Y where $S \simeq SL(2, 9)$. Let $s \in S$ have order 5. Similarly, let $g \in Fi_{22}$ have order 5. As $\chi_{78}(g) = 3$, the remarks above imply $\psi(s) = \psi(g) = 6$. However by Theorem 2.5(c) we have $\psi(s) = 1$.

Suppose finally that $\psi(1) = 729$. Then $\psi|_{Fi_{22}} = a \cdot 1_{Fi_{22}} + b\chi_{78} + b\chi_{429}$ for some nonnegative integers a, b, c where χ_{429} is the unique irreducible character of Fi_{22} of degree 429. Again, we have $ES \leq Y$ where $S \simeq SL(2, 27)$. Thus, S contains an element s of order 13, and since 13 divides the order of Fi_{22} and Y to the first power only, s is conjugate to an element g of Fi_{22} of order 13. Now $\chi_{78}(g) = \chi_{429}(g) = 0$, so $\psi(s) = \psi(g) = a$. Calculating $\psi(s)$ directly using Theorem 2.5(a), we have $\psi(s) = 1$ so that $a = 1$. This implies $728 = 78b + 429c$, which is a contradiction, as 3 does not divide 728. This concludes the proof of Lemma 6.11. \square

The last two lemmas imply that no linear group X exists which satisfies the conditions of Hypothesis 6.6. This strengthens Lemma 6.5 to the following theorem, which is the main result of this section.

THEOREM 6.12. *If X is a finite subgroup of $SL(q^2, \mathbb{C})$ containing G^0 where q is an odd prime power, then X is contained in G_2 .*

When q is a power of 2, the linear group $\rho(G)$ as constructed in § 3 is irreducible and primitive, and so some version of Theorem 6.12 is valid in this case as well. However, no claim is made here about the existence of a *unique* maximal unimodular linear group containing $\rho(G)$. Notice that when $q^2 = 4$, $\rho(G)$ is not unimodular (Theorem 3.4), and linear groups with determinant ± 1 would have to be considered. This (mild) exceptional behavior when $q^2 = 4$ contrasts that of the non-Hermitian case of [3] when the field is either $GF(2)$ or $GF(4)$. In that case the resulting linear group is not even primitive.

Appendix. The Molien series $\Phi_{G_0}(X)$ and $\Phi_G(X)$ are listed here for the fields $GF(q^2) = GF(4)$, $GF(9)$ and $GF(16)$. Recall that $G_0 = \langle -1 \rangle \times U(2, F)$ for all characteristics, $G = \langle -1 \rangle \times \tilde{G}$ where $\tilde{G} = U(2, F) \rtimes E$ (as defined in § 2) for characteristic $p > 2$, and $G = U(2, F) \rtimes E$ (as defined in § 3) for characteristic 2.

Theorems 4.6, 4.7 and 5.3 were used to calculate these series, and the calculations themselves were done on a microcomputer. All coefficients turn out to be integral, providing a built-in (partial) check of the results.

Unnormalized codes:

$$q^2 = 4, \quad \Phi_{G_0}(X) = \frac{1 + 3X^6}{(1 - X^2)^2(1 - X^6)^2},$$

$$q^2 = 9, \quad \Phi_{G_0}(X) = \frac{\sum_{i=0}^{18} a_i^{(9)} X^{2i}}{(1 - X^2)^3(1 - X)^4(1 - X^6)^3(1 - X^8)(1 - X^{12})},$$

$$q^2 = 16, \quad \Phi_{G_0}(X) = \frac{\sum_{i=0}^{20} a_i^{(16)} (X^{2i} + X^{84-2i}) + a_{21}^{(16)} X^{42}}{(1-X^2)^8(1-X^6)^4(1-X^{10})^3(1-X^{30})}.$$

Normalized codes:

$$q^2 = 4, \quad \Phi_G(X) = \frac{1 - X^4 + X^8}{(1 - X^2)(1 - X^4)(1 - X^6)(1 - X^{12})}$$

$$= \frac{1 + X^{12}}{(1 - X^2)(1 - X^6)(1 - X^8)(1 - X^{12})},$$

$$q^2 = 9, \quad \Phi_G(X) = \frac{\sum_{i=0}^{17} b_i^{(9)} X^{6i}}{(1 - X^6)^3(1 - X^{12})^2(1 - X^{18})^3(1 - X^{24})},$$

$$q^2 = 16, \quad \Phi_G(X) = \frac{\sum_{i=0}^{28} b_i^{(16)} (X^{2i} + X^{116-2i}) + b_{29}^{(16)} X^{58}}{(1 - X^2)^2(1 - X^4)^6(1 - X^6)^4(1 - X^{10})(1 - X^{20})^2(1 - X^{30})}.$$

TABLE A.1
Table of coefficients

i	$a_i^{(9)}$	$a_i^{(16)}$	$b_i^{(9)}$	$b_i^{(16)}$	$b_{i+20}^{(16)}$
0	1	1	1	1	424034
1	-1	-5	-1	-1	508755
2	4	24	25	-5	595717
3	20	73	167	4	680859
4	54	549	791	22	760700
5	102	2105	2459	33	829807
6	219	6975	5731	93	886525
7	306	18308	10135	359	929723
8	399	41684	14594	1226	957882
9	506	82248	16936	3158	967420
10	483	147269	16196	6912	
11	448	236796	12623	13734	
12	389	352683	7963	25638	
13	251	482896	3866	43940	
14	155	623000	1438	69934	
15	88	752772	341	104645	
16	23	870049	45	149424	
17	8	958295	1	204318	
18	1	1023591		269737	
19		1063239		343516	
20		1085823			
21		1092050			

REFERENCES

[1] L. DORNHOFF, *Group Representation Theory, Part A*, Marcel Dekker, New York, 1971.
 [2] W. FEIT, *On finite linear groups*, J. Algebra, 5 (1967), pp. 378-400.
 [3] S. M. GAGOLA, JR., *Weight enumerators of normalized codes*, this Journal, 2 (1981), pp. 347-380.
 [4] A. M. GLEASON, *Weight polynomials of self-dual codes and the MacWilliams identities*, in Act. Congr. Int. Math., vol. 3 (1970), pp. 211-215, Gauthier-Villars, Paris, 1971.
 [5] D. GORENSTEIN, *Finite Groups*, Harper and Row, New York, 1968.
 [6] R. L. GRIESS, JR., *Schur multipliers of the known finite simple groups*, II, AMS Proceedings of Symposia in Pure Mathematics, vol. 37 (1980), pp. 279-282.

- [7] D. C. HUNT, *Character tables of certain finite simple groups*, Bull. Austral. Math. Soc., 5 (1971), pp. 1–42.
- [8] B. HUPPERT, *Endliche Gruppen I*, Springer-Verlag, Berlin/Heidelberg/New York, 1967.
- [9] I. M. ISAACS, *Character Theory of Finite Groups*, Academic Press, New York, 1976.
- [10] F. J. MACWILLIAMS, A. M. ODLYZKO, N. J. A. SLOANE AND H. N. WARD, *Self-dual codes over $GF(4)$* , J. Combin. Theory, Ser. A, 25 (1978), pp. 288–318.
- [11] N. J. A. SLOANE, *Error-correcting codes and invariant theory: New applications of a nineteenth century technique*, Amer. Math. Monthly, 84 (1977), pp. 82–107.

ON SOME PROBLEMS IN THE DESIGN OF PLANE SKELETAL STRUCTURES*

KŌKICHI SUGIHARA†

Abstract. Two-dimensional frameworks composed of rods and joints are studied from a matroid theoretical point of view, and three problems are solved. First, an efficient algorithm is proposed for deciding whether or not a framework has redundant rods. Second, a solution is given to the problem of how to use a redundant rod for the construction of a “strong” framework in the sense that the rigidity is not violated if any rod is broken. Third, a method is given for the optimal construction of rigid structures under some constraints on available rods.

1. Introduction. Rigidity of skeletal structures has recently been studied actively, and many new and interesting results have been obtained. For example, Laman [11] found a necessary and sufficient condition for a graph to be rigid when its arcs and vertices are made of rigid rods and rotatable joints, Asimow and Roth [1], [2] established a powerful approach to rigidity in terms of edge functions, and many others characterized rigidity of various kinds of structures such as rectangular grids (Bolker and Crapo [3]), tensegrity structures (Connelly [5], and Roth and Whiteley [14]), and bipartite structures (Bolker and Roth [4]).

It seems, however, that in most of those works rigidity is treated from purely mathematical interest. From an engineering point of view, there are many problems left unconsidered. Especially, problems concerning efficient algorithms for the design of structures with specified properties have scarcely been considered.

The present work is an application of matroid theory to some problems arising out of the attempt to design two-dimensional rigid structures.

In § 2 we follow Crapo [6] and Lovász and Yemini [13] and define generic independence for a two-dimensional structure. The result is a certain matroid on the underlying graph of the structure. We also review Laman’s theorem [11], which characterizes this matroid.

Based on these preliminaries, we shall solve three problems. The first problem is how to determine whether or not a structure has redundant rods. This problem was theoretically solved by Laman [11], but his method is not practical because its time complexity is an exponential function of the size of the structure. Recently Lovász and Yemini [13] found a polynomial time algorithm for the recognition of nonredundant structures, though they did not present its time complexity explicitly. In § 3 we shall present a new theorem which enables us to construct an efficient algorithm for this problem. The second problem is how to use a redundant rod efficiently. Redundant rods can make the structure “stronger” in the sense that the structure remains rigid even if some rods are broken. In § 4 we shall consider the most efficient way of making use of a redundant rod against a breakdown of a rod. The third problem is how to use rods efficiently for the construction of a rigid structure. In practical situations we usually have restrictions on materials. As an example of such situations, we shall in § 5 consider the problem of how to connect given points into a rigid structure by rods of given lengths.

Finally, let us point out some of the problems we do not address.

First, the analogue of Laman’s theorem in three or more dimensions is not yet known (see Asimow and Roth [1] and Crapo [6]).

* Received by the editors September 2, 1981, and in final revised form September 7, 1982.

† Department of Information Science, Faculty of Engineering, Nagoya University, Furō-chō, Chikusa-ku, Nagoya, Japan 464.

Second, while we restrict our consideration to independent/dependent relations of rods, it is also important to study the structure of the assignment of degrees of freedom to joint positions, which forms an integral polymatroid (see Lovász [12] and Sugihara [15]). Furthermore, we can extract many integral polymatroids from general rigid/flexible structures in which solid members are connected by various kinds of joints such as ball joints, pin joints, piston joints, etc.; there are many unsolved problems there.

Third, though our consideration is restricted to the topological properties of the underlying graphs, the reliability of actual structures depends on the positions of joints and stiffness of rods, and consequently a quantitative approach is also necessary for their analysis.

2. Preliminaries. We consider two-dimensional frameworks composed of rigid rods and rotatable joints. Each joint connects the end vertices of two or more rods in such a way that the mutual angles of the rods can change freely if the other ends are not constrained. A framework of this kind is called a *plane skeletal structure*. Let V and E denote the set of joints and that of rods, respectively. (If there are any end vertices that are not connected with others, we shall include them in V as joints with single rods.) Regarding V and E as a vertex set and an edge set, we obtain a finite undirected graph $G = (V, E)$ without loops or multiple edges. G is called the *underlying graph* of the plane skeletal structure.

Let S be a plane skeletal structure with underlying graph $G = (V, E)$, and (x_i, y_i) be the Cartesian coordinates of vertex (=joint) v_i ($\in V$). Furthermore, let $|V| = n$ and $|E| = m$, where $|X|$ denotes the number of elements of finite set X . An edge (=rod) connecting v_i and v_j constrains the movement of S in such a way that the distance between the vertices is constant:

$$(1) \quad (x_i - x_j)^2 + (y_i - y_j)^2 = \text{const.}$$

Differentiating with respect to time t , we get

$$(2) \quad (x_i - x_j)(\dot{x}_i - \dot{x}_j) + (y_i - y_j)(\dot{y}_i - \dot{y}_j) = 0,$$

where the dot denotes the differentiation with respect to t . Equation (2) implies that the relative velocity should be perpendicular to the rod, that is, no rod is stretched or compressed. Gathering the equations associated with each edge in E , we get a system of linear equations

$$(3) \quad H\mathbf{w} = \mathbf{0},$$

where H is an $m \times 2n$ constant matrix and \mathbf{w} is a column vector of unknown variables $\mathbf{w} = {}^t(\dot{x}_1 \dot{y}_1 \cdots \dot{x}_n \dot{y}_n)$ (t denotes transpose). A vector \mathbf{w} is called an *infinitesimal displacement* of S if it satisfies (3). The infinitesimal displacements of S form a linear subspace of \mathbf{R}^{2n} . The rigid motions in a plane yield a three-dimensional subspace of this linear space. S is called *rigid* if the infinitesimal displacements of S form a three-dimensional linear space. (This definition of rigidity is due to Laman [11]; Asimow and Roth [1] and Connelly [5] propose other definitions.)

The rigidity of a structure depends on the positions of joints. The structure shown in Fig. 1(a) is rigid, while the one shown in (b), which has the same underlying graph, is not rigid; the assignment of velocities indicated by the arrows (the vertices without arrows are assumed to have zero velocities) forms an infinitesimal displacement because it does not violate (3). Similarly, though the structures in (c) and (d) have the same underlying graph, (c) is rigid and (d) is not. Note that an infinitesimal displacement does not always correspond to an actual movement of a mechanism; the structure in

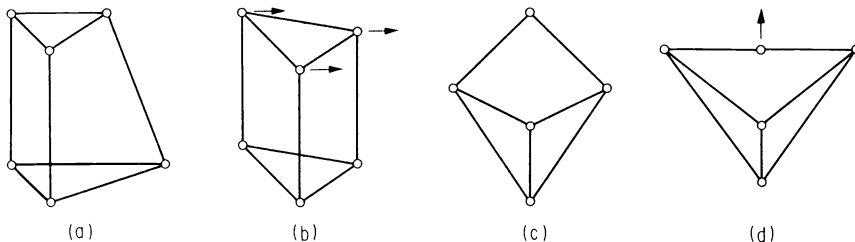


FIG. 1. Rigid and nonrigid structures.

(b) deforms mechanically, whereas the structure in (d) does not. The structure in (d) is categorized as nonrigid only because our definition of rigidity requires the absence of infinitesimal deformations.

The vertices of a structure S are in *general position* if $x_1, y_1, \dots, x_n, y_n$ are algebraically independent over the rational field. When the vertices are in general position, the definition of algebraic dependence shows that a subdeterminant of the matrix H is 0 if and only if it is identically 0 when we consider $x_1, y_1, \dots, x_n, y_n$ as variables. Therefore, if the vertices are in general position, the linear independence of the equations in (3) depends only on the underlying graph, and consequently the rigidity also depends only on the graph. In what follows, we shall consider structures whose vertices are in general position.

Suppose that $G = (V, E)$ is the underlying graph of structure S whose vertices are in general position. G is called *stiff* if S is rigid. For any $X \subseteq E$, let $\rho_G(X)$ be the rank of the submatrix of H consisting of the rows associated with the edges in X . The value $\rho_G(X)$ is called the *generic rank* of X . X is called *generically independent* if $\rho_G(X) = |X|$, and *generically dependent* if $\rho_G(X) < |X|$. G is called *generically independent* (resp. *dependent*) if E is generically independent (resp. dependent).

It can be seen that ρ_G is a rank function of a matroid (Welsh [14]), and hence (E, ρ_G) is the matroid on E defined by the rank function ρ_G .

For any subset X of E , let $V(X)$ be the set of terminal vertices of edges in X and define μ_G by

$$(4) \quad \mu_G(X) = 2|V(X)| - |X| - 3.$$

Laman [11] proved a basic theorem on rigidity which is equivalent to the following.

LAMAN'S THEOREM. *The graph $G = (V, E)$ is generically independent if and only if $\mu_G(X) \geq 0$ for any nonempty subset X of E .*

Laman's theorem and the definition of rigidity imply immediately the following.

COROLLARY L.1. *$G = (V, E)$ is stiff if and only if there exists $E' \subseteq E$ such that $|E'| = 2|V| - 3$ and $\mu_G(X) \geq 0$ for every nonempty subset X of E' .*

COROLLARY L.2. *$G = (V, E)$ is stiff and generically independent if and only if*

- (a) $\mu_G(E) = 0$, and
- (b) $\mu_G(X) \geq 0$ for every nonempty subset X of E .

3. Efficient recognition of generic independence. If we simply use Laman's theorem, it would take 2^m steps to determine whether or not a graph is generically independent, where m is the number of edges. In the present section we shall construct a polynomially bounded algorithm for the recognition of generic independence.

Let $B = (U_1, L, U_2)$ be a bipartite graph with node sets U_1, U_2 and arc set L . A subset L' of L is called a *complete matching* with respect to U_1 if the terminal nodes of arcs in L' are distinct and if every node in U_1 is a terminal node of some arc in

L' . For $X \subseteq U_1$, let $\Gamma(X)$ be the set of those nodes in U_2 that are connected to nodes in X by some arc in L . It is known that G has a complete matching with respect to U_1 if and only if $|X| \leq |\Gamma(X)|$ for every $X \subseteq U_1$, and an efficient algorithm for finding a complete matching is also known (Hopcroft and Karp [7]).

Let $G = (V, E)$ be a graph with n vertices and m edges. For each vertex v_i of G , let p_i, q_i be distinct symbols. Then let $B(G) = (N_1, A, N_2)$ be the bipartite graph whose node sets N_1, N_2 and arc set A are defined by

$$\begin{aligned} N_1 &= E, \\ N_2 &= \{p_1, q_1, \dots, p_n, q_n\}, \\ A &= \{(e, p_i), (e, q_i), (e, p_j), (e, q_j) | e = \{v_i, v_j\} \in E\}. \end{aligned}$$

Let t_1, t_2, t_3 be three distinct symbols. Then for any $1 \leq i < j \leq n$, let $B_{ij}(G) = (\bar{N}_1, A_{ij}, N_2)$ be the bipartite graph constructed from $B(G)$ by the addition of three nodes and three arcs in the following way.

$$\begin{aligned} \bar{N}_1 &= N_1 \cup \{t_1, t_2, t_3\}, \\ A_{ij} &= A \cup \{(t_1, p_i), (t_2, q_i), (t_3, p_j)\}. \end{aligned}$$

For any $Z \subseteq \bar{N}_1$, we shall denote by $\Gamma_{ij}(Z)$ the set of nodes in N_2 that are connected to elements of Z by arcs in A_{ij} . Note that $2|V(X)| = |\Gamma_{ij}(X)|$ for any $X \subseteq E$. Then we can prove the following theorem.

THEOREM 1. *The graph $G = (V, E)$ is generically independent if and only if, for any i and j ($1 \leq i < j \leq n$), $B_{ij}(G) = (\bar{N}_1, A_{ij}, N_2)$ has a complete matching with respect to \bar{N}_1 .*

Proof. Suppose that G is generically independent. Let $Z = X \cup Y$ ($X \subseteq E, Y \subseteq \{t_1, t_2, t_3\}$) be any subset of \bar{N}_1 . If $X = \emptyset$, then $|\Gamma_{ij}(Z)| = |\Gamma_{ij}(Y)| = |Y| = |Z|$. If $X \neq \emptyset$, then

$$|\Gamma_{ij}(Z)| \geq 2|V(X)| \geq |X| + 3 \geq |Z|,$$

where the first inequality follows from the definition of $B_{ij}(G)$ and the second from Laman's theorem. Thus $|\Gamma_{ij}(Z)| \geq |Z|$ in every case, and hence $B_{ij}(G)$ has a complete matching with respect to \bar{N}_1 .

Next suppose that, for any $1 \leq i < j \leq n$, $B_{ij}(G)$ has a complete matching with respect to \bar{N}_1 . Then, $|Z| \leq |\Gamma_{ij}(Z)|$ is satisfied for any $Z \subseteq \bar{N}_1$. Let X be any nonempty subset of E . Then $V(X)$ contains at least two vertices, say v_k and v_l ($1 \leq k < l \leq n$). Because $B_{kl}(G)$ has a complete matching, we get

$$|X \cup \{t_1, t_2, t_3\}| \leq |\Gamma_{kl}(X \cup \{t_1, t_2, t_3\})|.$$

Since $\Gamma_{kl}(X \cup \{t_1, t_2, t_3\}) = \Gamma_{kl}(X)$, we obtain

$$2|V(X)| = |\Gamma_{kl}(X)| \geq |X \cup \{t_1, t_2, t_3\}| = |X| + 3.$$

By Laman's theorem we conclude that G is generically independent. Q.E.D.

A complete matching of the bipartite graph $B = (U_1, L, U_2)$ can be found in $O((|L| + |U_1| + |U_2|)|U_1|^{1/2})$ steps (Hopcroft and Karp [7]) and consequently a complete matching of $B_{ij}(G) = (\bar{N}_1, A_{ij}, N_2)$ can be found in $O(m^{1.5})$ steps, because $|\bar{N}_1| = m + 3$, $|N_2| = 2n$, $|A_{ij}| = 4m + 3$, and $n \leq 2m$, where m and n are the number of edges and that of vertices, respectively, of G . The number of $B_{ij}(G)$'s is proportional to m^2 . Therefore, by the simplest implementation of Theorem 1, we can decide whether G is generically independent or not in $O(m^{3.5})$ steps.

This order of time complexity can still be lessened. Recall that the bipartite graph $B_{ij}(G)$'s are very similar to each other. Once a complete matching of any one of $B_{ij}(G)$'s is found, complete matchings of the other $B_{ij}(G)$'s can be efficiently obtained by slight modifications of this complete matching. On the basis of this observation, Imai [8] found quite recently an $O(m^2)$ algorithm for the recognition of the generic independence of G .

4. Strong structures. Let $G = (V, E)$ be an underlying graph of structure S . Edge $e (\in E)$ is *redundant* in G if $\rho_G(E) = \rho_G(E - \{e\})$. G is *redundant* if G has a redundant edge. G is *globally redundant* if every edge of G is redundant. G is *strong* if G is stiff and globally redundant. A strong graph remains stiff if any edge is deleted. G is *C-strong* if G is strong and $|E| = 2|V| - 2$. Since a stiff graph contains at least $2|V| - 3$ edges, a C-strong graph affords us the most efficient way of bracing a structure with one redundant rod. In the present section we shall investigate C-strong graphs.

It is easy to see the following.

THEOREM 2. *The graph $G = (V, E)$ is C-strong if and only if E is a circuit of the matroid (E, ρ_G) .*

For any $n (\geq 4)$ we can construct a C-strong graph with n vertices. Let $V = \{1, 2, \dots, n\}$, and G be a graph constructed first by connecting vertices 1, 2, 3 by the three edges and next, for each $i = 4, 5, \dots, n$, connecting vertex i to vertices $i - 1$ and $i - 2$, as illustrated in Fig. 2(a). G is stiff because it is composed of triangles sharing edges. Since G has $2n - 3$ edges, G is generically independent by Corollary L.1. Let G' be the graph obtained by the addition of edge $\{1, n\}$ to G , as shown in Fig. 2(b). G' is generically dependent because it has $2n - 3$ edges, and moreover the deletion of any edge from G' yields a graph that satisfies the condition in Laman's theorem. Therefore, G' is C-strong.

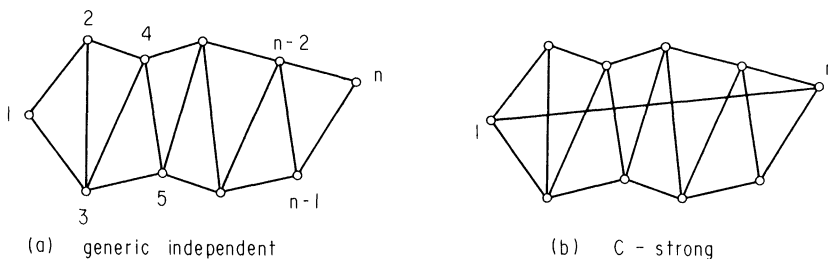


FIG. 2. C-strong graph.

Now suppose that G is any graph which is stiff and generically independent. Is it possible to make a C-strong graph by the addition of one edge to G ? If so, how shall we brace it? Since $G = (V, E)$ is stiff and generically independent, $\mu_G(X) \geq 0$ for every nonempty subset X of E , where μ_G is defined in (4). Let $\mathcal{E}(G)$ be the collection of maximal proper subsets X of E such that $\mu_G(X) = 0$, that is,

$$\mathcal{E}(G) = \{X | X \subsetneq E, \mu_G(X) = 0, \text{ and } \mu_G(Y) > 0 \text{ for any } Y \text{ such that } X \subsetneq Y \subsetneq E\}.$$

For every $X \in \mathcal{E}(G)$, $(V(X), X)$ is a stiff subgraph of G . Let $\bar{G} = (V, \bar{E}(G))$ be the graph having vertex set V and edge set

$$\bar{E}(G) = \{\{v_i, v_j\} | v_i, v_j \in V(X) \text{ for some } X \in \mathcal{E}(G)\};$$

that is, two vertices in \bar{G} are connected by an edge if and only if they belong to the same stiff proper subgraph of G . Obviously $E \subseteq \bar{E}(G)$. We get the next theorem.

THEOREM 3. *Let $G = (V, E)$ be stiff and generically independent. An addition of new edge e ($e \notin E$) to G yields a C -strong graph if and only if $e \notin \bar{E}(G)$.*

Proof. Let $E' = E \cup \{e\}$ and $G' = (V, E')$. E' is generically dependent because $|E'| = |E| + 1 = 2|V| - 2$.

First, suppose that $e \in \bar{E}(G)$. Then, there exists $Y \in \mathcal{G}(G)$ such that $V(\{e\}) \subseteq V(Y)$, and hence $V(Y) = V(Y \cup \{e\})$. Since $2|V(Y)| = |Y| - 3$, we get $2|V(Y \cup \{e\})| = |Y \cup \{e\}| - 2$ and consequently $Y \cup \{e\}$, which is a proper subset of E' , is generically dependent. Therefore, G' is not C -strong.

Next, suppose that $e \notin \bar{E}(G)$. Assume that, for some $e' \in E$, $E \cup \{e\} - \{e'\}$ is generically dependent. Then, there exists $X \subseteq E - \{e'\}$ such that $2|V(X \cup \{e\})| = |X \cup \{e\}| - 2 = |X| - 3$. Since $2|V(X)| \geq |X| - 3$, we get $|V(X \cup \{e\})| = |V(X)|$, and consequently $2|V(X)| = |X| - 3$ and $V(X \cup \{e\}) = V(X)$. Hence, $e \in \bar{E}(G)$, which is a contradiction. Therefore, $E \cup \{e\} - \{e'\}$ is generically independent for any $e' \in E$, and hence G' is C -strong. Q.E.D.

Let $G = (V, E)$ be the graph shown in Fig. 3(a). G is stiff and generically independent with six vertices and nine edges. $\mathcal{G}(G)$ consists of the two subsets $E - \{a, b\}$ and $E - \{h, i\}$. Because $V - V(E - \{a, b\}) = \{1\}$ and $V - V(E - \{h, i\}) = \{6\}$, graph \bar{G} is as shown in 3(b). Therefore, the only way to make a C -strong graph is to add an edge connecting 1 and 6, as shown in Fig. 3(c). By similar argument we can see that the bracing in Fig. 2(b) is the only way to make a C -strong graph from Fig. 2(a).

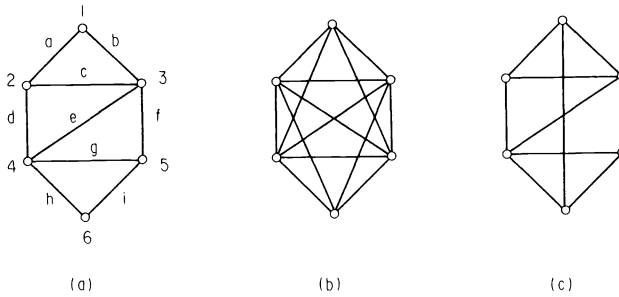


FIG. 3. Construction of a C -strong graph.

Two more examples are shown in Fig. 4, where the graphs illustrated by the solid lines are stiff and generically independent, and the broken lines represent new edges to be added for the construction of C -strong graphs.

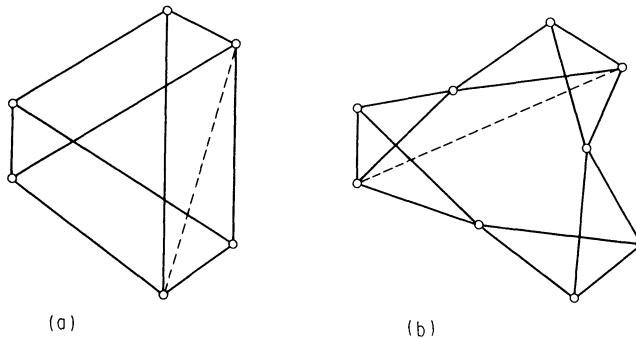


FIG. 4. Some examples of C -strong graphs.

Next, let G be the graph shown in Fig. 5. $\mathcal{E}(G)$ consists of the three subsets $E - \{a, b\}$, $E - \{c, d\}$ and $E - \{h, i\}$. It is easily shown that any pair of vertices in V belongs to $V(E - \{a, b\})$, $V(E - \{c, d\})$ or $V(E - \{h, i\})$, and hence \bar{G} is a complete graph. Therefore we can not make G C -strong by adding an edge. Two new edges are necessary to make G strong.

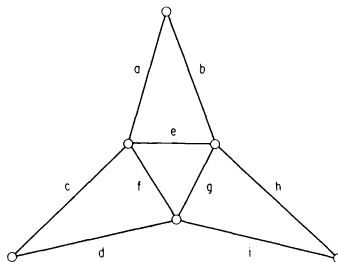


FIG. 5. Graph which requires two new edges to become strong.

5. Design of rigid structures under rod-length constraints. In the present section we shall consider the following problem.

Problem 1. Suppose that we are given a graph $G = (V, E)$, a partition $\{E_i\}_{i=1}^p$ of E and nonnegative integers h_1, \dots, h_p . Find subset X_i of E_i for $i = 1, \dots, p$ such that graph $(V, X_1 \cup X_2 \cup \dots \cup X_p)$ is stiff and generically independent and $|X_i| \leq h_i$.

This problem arises, for example, in the following situation. Let V be a finite set of vertices whose locations are specified in general position on a plane. Let E be the set of pairs of vertices that can be connected by rods. Suppose that h_i ($i = 1, \dots, p$) is the number of available rods of length d_i ($d_1 < d_2 < \dots < d_p$). We consider the most efficient way of using these rods to connect the vertices into a rigid structure. It would be wasteful for us to connect close vertices by a long rod. Therefore, when the distance from v_j to v_k is $d(v_j, v_k)$, it seems most efficient to connect v_j and v_k by using part of a rod of length d_i where $d_{i-1} < d(v_j, v_k) \leq d_i$. Let us define

$$E_1 = \{\{v_j, v_k\} | \{v_j, v_k\} \in E, d(v_j, v_k) \leq d_1\},$$

$$E_i = \{\{v_j, v_k\} | \{v_j, v_k\} \in E, d_{i-1} < d(v_j, v_k) \leq d_i\}$$

for $i = 2, \dots, p$. Then Problem 1 is equivalent to finding a way of constructing a rigid structure with a minimum number of rods subject to the constraints on the number of available rods of each length.

Let σ_i be the nonnegative, integer-valued function on 2^{E_i} defined by

$$\sigma_i(X) = \min \{|X|, h_i\},$$

for $X \subseteq E_i$, and let σ be the function on 2^E defined by

$$\sigma(X) = \sum_{i=1}^p \sigma_i(X \cap E_i)$$

for $X \subseteq E$. Then, (E_i, σ_i) is a matroid and consequently (E, σ) is a matroid. $((E_i, \sigma_i)$ is a uniform matroid of rank h_i , and (E, σ) is the union matroid of the matroids (E_i, σ_i) ($i = 1, \dots, p$); see Welsh [16].)

Note that a subset $X = X_1 \cup \dots \cup X_p$ ($X_i \subseteq E_i$) of E satisfies the condition $|X_i| \leq h_i$ if and only if X is an independent set of the matroid (E, σ) . We have already seen

that $X \subseteq E$ is generically independent if and only if X is an independent set of the matroid (E, ρ_G) . Therefore, Problem 1 can be reduced to Problem 2.

Problem 2. Find a maximal subset X of E which is independent in both of the matroids (E, ρ_G) and (E, σ) .

Let X^* be a solution to Problem 2. If $|X^*| = 2|V| - 3$, then X^* is a solution to Problem 1. If otherwise, Problem 1 does not have any solution.

A solution to Problem 2 is called a *maximal common independent set*, and an efficient algorithm for finding one has already been established in a more general framework by Iri and Tomizawa [10]. Moreover, a partially ordered structure called a *principal partition* that accompanies this problem (Iri [9]) gives us much information about the solutions; if the solutions to Problem 1 exist, it tells us a family of all the solutions, and if they do not exist, it gives us information about the second best way for constructing the rigid structure.

Acknowledgments. The author would like to express his appreciation to Professor Masao Iri of the University of Tokyo for valuable discussions. He is also grateful to the referee for his kind comments, which made the paper more readable.

REFERENCES

- [1] L. ASIMOW AND B. ROTH, *The rigidity of graphs*, Trans. Am. Math. Soc., 245 (1978), pp. 279–289.
- [2] ———, *The rigidity of graphs II*, J. Math. Anal. Appl., 68 (1979), pp. 171–190.
- [3] E. D. BOLKER AND H. CRAPO, *Bracing rectangular frameworks I*, SIAM J. Appl. Math., 36 (1979), pp. 473–490.
- [4] E. D. BOLKER AND B. ROTH, *When is a bipartite graph a rigid framework?*, Pacific J. Math., 90 (1980), pp. 27–44.
- [5] R. CONNELLY, *The rigidity of certain cabled frameworks and the second-order rigidity of arbitrarily triangulated convex surfaces*, Advances in Math., 37 (1980), pp. 272–299.
- [6] H. CRAPO, *Structural rigidity*, Structural Topology, 1 (1979), pp. 26–45.
- [7] J. E. HOPCROFT AND R. M. KARP, *An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs*, SIAM J. Comput., 2 (1973), pp. 225–231.
- [8] H. IMAI, *Transversal polymatroids with negative defect and network flow*, preprint, Univ. of Tokyo, 1982.
- [9] M. IRI, *A review of recent work in Japan on principal partitions of matroids and their applications*, Ann. New York Acad. Sci., 319 (1979), pp. 306–319.
- [10] M. IRI AND N. TOMIZAWA, *An algorithm for finding an optimal independent assignment*, J. Oper. Res. Japan, 19 (1976), pp. 32–57.
- [11] G. LAMAN, *On graphs and rigidity of plane skeletal structures*, J. Eng. Math., 4 (1970), pp. 331–340.
- [12] L. LOVÁSZ, *Matroid matching and some applications*, J. Combin. Theory, B28 (1980), pp. 208–236.
- [13] L. LOVÁSZ AND Y. YEMINI, *On generic rigidity in the plane*, this Journal, 3 (1982), pp. 91–98.
- [14] B. ROTH AND W. WHITELEY, *Tensegrity frameworks*, Trans. Am. Math. Soc., 265 (1981), pp. 419–446.
- [15] K. SUGIHARA, *A unifying approach to descriptive geometry and mechanisms*, Discrete Appl. Math. (to appear).
- [16] D. J. A. WELSH, *Matroid Theory*, Academic Press, London, 1976.

A CLASS OF BALANCED MATRICES ARISING FROM LOCATION PROBLEMS*

ARIE TAMIR†

Abstract. A $(0, 1)$ -matrix is balanced if it contains no square submatrix of odd order whose row and column sums are all two. Given two collections, $S = \{T_1, \dots, T_m\}$ and $Q = \{T'_1, \dots, T'_n\}$, of neighborhood subtrees of a tree T , let $A(S, Q) = (a_{ij})$ be the incidence matrix with $a_{ij} = 1$ if and only if T_i intersects T'_j . It is shown that $A(S, Q)$ is balanced. This balancedness is then used to exhibit the existence of a polynomial algorithm to certain location problems.

1. Introduction. In his paper [1], Berge defined a $(0, 1)$ -matrix to be balanced if it contains no square submatrix of odd order whose row and column sums are all two. He then characterized a balanced matrix in terms of the existence of integral solutions to certain linear programs whose constraints are defined by a balanced matrix. Berge's results were then refined and extended by Lovasz [8] and Fulkerson, Hoffman and Oppenheim [5].

In this work a special class of balanced matrices is presented. This class arises from location problems on tree networks.

Assume that an undirected tree $T = (N, E)$, with N and E denoting the sets of nodes and edges respectively, is embedded in the Euclidean plane, so that the edges are line segments whose endpoints are the nodes and the edges intersect one another only at nodes. Moreover, each edge of T has a positive Euclidean length. This embedding enables us to talk about points, not necessarily nodes, on the edges. For any two points x, y on T let $d(x, y)$ denote the distance between x and y , measured along the edges of T . $P(x, y)$ will denote the set of points on the simple path connecting x and y . T will also be used to denote the (infinite) set of points on T .

A subtree of T is a connected subset of the set T . A subtree, T_i , is called a *neighborhood subtree* if there exist a point $x_i \in T$ and $r_i \geq 0$ such that $T_i = \{x \mid x \in T, d(x, x_i) \leq r_i\}$. x_i is called the center of T_i .

Let $S = \{T_1, \dots, T_k\}$ be a finite collection of subtrees of T , and define the intersection graph, $G(S)$, as follows. $G(S)$ has k nodes, corresponding to the k subtrees in S . Two nodes of $G(S)$ are connected by an edge if and only if the respective subtrees intersect. Defining a clique to be a maximal complete subgraph, let $A(S)$ be the node clique incidence matrix of $G(S)$, where nodes correspond to rows and cliques appear as the columns. The graph $G(S)$ has been shown in [2] to be chordal, i.e., for any circuit of order at least four there exists an edge, not of the circuit, which connects two nodes of the circuit. It is also proved in [2] that any chordal graph is realizable as the intersection graph of subtrees of a tree. Furthermore $G(S)$ is known to be perfect and its respective matrix $A^T(S)$ is therefore perfect, [10]. Perfectness is weaker than balancedness. In fact, even matrices $A^T(S)$ arising from chordal graphs $G(S)$ are not necessarily balanced. This is illustrated by the following chordal graph.

Example 1. Let T be as in Fig. (1a) and define the collection of subtrees $S = \{T_1, T_2, T_3, T_4, T_5, T_6\}$ as follows. $T_1 = P(v_1, v_2)$, $T_2 = P(v_2, v_3)$, $T_3 = P(v_3, v_1)$,

* Received by the editors July 11, 1980, and in final revised form August 2, 1982.

† Department of Statistics, Tel Aviv University, Ramat Aviv, Israel.

$T_4 = \{v_2\}$, $T_5 = \{v_3\}$ and $T_6 = \{v_1\}$. The respective intersection graph, $G(S)$, is given in Fig. 1(b), and

$$A^T(S) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Considering the submatrix of $A^T(S)$ defined by the first three columns and the last three rows we observe that $A^T(S)$ is not balanced. (We note in passing that $G(S)$ is also realizable as the intersection of neighborhoods in R^2 .)

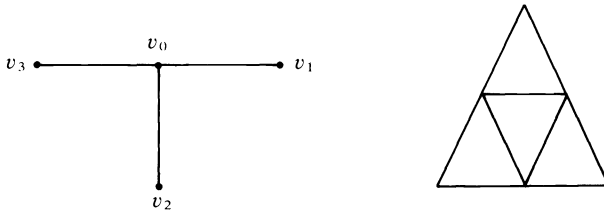


FIG. 1

In this paper we focus on collections of neighborhood subtrees and show that they, unlike collections of arbitrary subtrees do give rise to balanced matrices.

Let $S = \{T_1, \dots, T_m\}$ and $Q = \{T'_1, \dots, T'_n\}$ be two sets of neighborhood subtrees in T . Define the $m \times n$ incidence matrix $A(S, Q) = (a_{ij})$ by $a_{ij} = 1$ if and only if the intersection $T_i \cap T'_j$ is nonempty and $a_{ij} = 0$, otherwise. We will prove that $A(S, Q)$ is balanced. (In particular, $A(S)$ is balanced when S consists of neighborhood subtrees.) This result is then applied to exhibit the existence of polynomial algorithms for certain location models.

2. Balancedness and intersection graphs.

LEMMA 1. Let $\{x_1, \dots, x_k\}$, $k \geq 3$, be a set of distinct points on T . Define $x_{k+1} = x_1$. There exist indices $1 \leq i_1 < i_2 < i_3 \leq k$ such that the paths $P(x_{i_1}, x_{i_1+1})$, $P(x_{i_2}, x_{i_2+1})$ and $P(x_{i_3}, x_{i_3+1})$ of the tree T intersect at some point $y \in T$.

Proof. We define the indices i_1, i_2, i_3 and the point $y \in T$ as follows. First let $x_{i_1} = x_1$ and $x_{i_2} = x_2$. Now y is chosen to be the closest point to x_3 on the path $P(x_1, x_2)$. It remains to define i_3 . If $k = 3$ set $i_3 = 3$ and the result clearly holds. Thus let $k > 3$. Define

$$j = \begin{cases} k - 1 & \text{if } y \notin P(x_i, x_3) \text{ for all } 3 < i \leq k - 1, \\ \min \{i | 3 \leq i < k - 1, y \in P(x_{i+1}, x_3)\} & \text{otherwise.} \end{cases}$$

Suppose first that $j = k - 1$. If $y \in P(x_k, x_3)$ then $y \in P(x_{k-1}, x_k)$, since $y \notin P(x_{k-1}, x_3)$. Set $i_3 = k - 1$ and then the paths $P(x_1, x_2)$, $P(x_2, x_3)$ and $P(x_{k-1}, x_k)$ intersect at y . If $y \notin P(x_k, x_3)$ then $y \in P(x_k, x_1)$ since $y \in P(x_1, x_3)$. Set $i_3 = k$ and the paths $P(x_1, x_2)$, $P(x_2, x_3)$ and $P(x_k, x_1)$ intersect at y .

Now suppose $j < k - 1$. Set $i_3 = j$. From the definition of j it follows that the paths $P(x_1, x_2)$, $P(x_2, x_3)$ and $P(x_j, x_{j+1})$ intersect at y . This completes the proof.

We are now ready to present the main result.

THEOREM 1. Let $S = \{T_1, \dots, T_m\}$ and $Q = \{T'_1, \dots, T'_n\}$ be two sets of neighborhood subtrees in T . Let $A(S, Q) = (a_{ij})$ be the incidence matrix satisfying $a_{ij} = 1$ if and

only if $T_i \cap T'_j$ is nonempty. Then $A(S, Q)$ does not contain a square submatrix of size $k \geq 3$ which has no identical columns, and its row and column sums equal to two.

Proof. Let $T_i = \{x | d(x_i, x) \leq r_i\}$, $i = 1, \dots, m$, and $T'_j = \{x | d(y_j, x) \leq s_j\}$, $j = 1, \dots, n$. Then $a_{ij} = 1$ if and only if $d(x_i, y_j) \leq r_i + s_j$. Suppose that $A(S, Q)$ contains a square submatrix $B = (b_{ij})$ of size $k \geq 3$ which has no identical columns and its row and column sums are all equal to two. Without loss of generality suppose that this is the submatrix defined by the first k columns and k rows of $A(S, Q)$. Also, suppose that $b_{ij} = 1$ if and only if $i = j, j - 1$ or $(i, j) = (1, k)$. First we note that $x_i \neq x_j$ for $1 \leq i \neq j \leq k$. Since, otherwise, we would have $T_i \subseteq T_j$ or $T_i \supseteq T_j$ which contradicts the fact that no row vector of B is greater than or equal to another row vector of B . Considering $\{x_1, \dots, x_k\}$, let $(x_{i_j}, x_{i_{j+1}})$, $j = 1, 2, 3$, be the three pairs obtained from the previous lemma, and let y be the point on the path connecting x_{i_j} to $x_{i_{j+1}}$, $j = 1, 2, 3$.

The matrix B expresses the intersection relations between $\{T_i\}_{i=1}^k$ and $\{T'_i\}_{i=1}^k$. Each column of B contains exactly two 1's. Furthermore, these two 1's appear consecutively (mod k), and one of them is a diagonal element. Therefore, for $j = 1, 2, 3$ there exist T'_{i_j} intersecting T_{i_j} and $T_{i_{j+1}}$ (exclusively). Without loss of generality suppose that

$$s_{i_1} - d(y, y_{i_1}) \leq s_{i_2} - d(y, y_{i_2}) \leq s_{i_3} - d(y, y_{i_3}).$$

Since $y \in P(x_{i_1}, x_{i_{1+1}})$ it follows that $y \in P(z, y_{i_1})$ where z is either x_{i_1} or $x_{i_{1+1}}$. Let $z = x_{i_1}$ then

$$\begin{aligned} 0 \leq r_{i_1} + s_{i_1} - d(x_{i_1}, y_{i_1}) &= r_{i_1} + s_{i_1} - d(x_{i_1}, y) - d(y, y_{i_1}) \\ &\leq r_{i_1} + s_{i_j} - d(x_{i_1}, y) - d(y, y_{i_j}) \leq r_{i_1} + s_{i_j} - d(x_{i_1}, y_{i_j}), \quad j = 1, 2, 3. \end{aligned}$$

Hence, we obtained the contradiction that the neighborhood subtree T_{i_1} intersects the three neighborhood subtrees T'_{i_j} , $j = 1, 2, 3$. This contradicts the fact that row i_1 of B contains exactly two 1's. A similar contradiction is obtained if we take $z = x_{i_{1+1}}$. Therefore the proof is now complete.

We note that since a point on T is a neighborhood subtree, Q for example may be a collection of points on T . Indeed, this is the special case arising from the location model considered in the next section.

COROLLARY 1. $A(S, Q)$ defined as in Theorem 1 is balanced. In particular, the node clique incidence matrix of the intersection graph corresponding to the collection of neighborhood subtrees S is balanced.

Proof. The first part is obvious from Theorem 1. Let $A(S)$ be the node clique incidence matrix of the intersection graph $G(S)$. It is shown in [3] that all the subtrees corresponding to a clique of $G(S)$ have a point in T contained in all of them. (The maximality of the clique ensures that this point is contained in no other subtree of the collection.) Thus there exists a set of points Y in T such that $A(S) = A(S, Y)$, and the result follows from Theorem 1.

Theorem 1 and Example 1 present one property which is satisfied by intersection graphs realizable by collections of neighborhood subtrees but not by chordal graphs which are known, [2], to be realized by collections of subtrees. Next, we demonstrate another property which is met by our class of balanced matrices but not by matrices arising from general chordal graphs.

Given that $A(S, Q)$ is balanced, it then follows from [5] that all the extreme points of the polyhedron $\{z | A(S, Q)z \geq e, z \geq 0\}$ are integral, (e is the vector of all 1's). As noted in the introduction, the node clique incidence matrix of a general

chordal graph, which is not an intersection graph of neighborhood subtrees, is not necessarily balanced. Hence the results of [5] do not induce the above integrality property of the respective polyhedron defined by a general chordal graph. Indeed, the above integrality property, which is weaker than balancedness, is not shared by a general chordal graph.

Example 2. Let G be the chordal graph in Fig. 2. Let A be the node clique incidence matrix of G (with nodes corresponding to rows).

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

The polyhedron $\{x | Ax \geq e, x \geq 0\}$ possesses the nonintegral extreme point, $x = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 1, 1, 1)$.

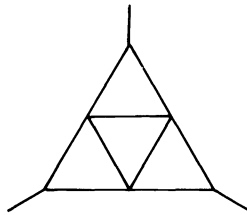


FIG. 2

Chordal graphs satisfy the following weaker integrality property.

THEOREM 2. *Let G be a chordal graph and let $A = (a_{ij})$ be its node clique incidence matrix with nodes corresponding to rows. If the equality constrained set covering polyhedron, $\{x | Ax = e, x \geq 0\}$, is nonempty, it is a singleton consisting of a 0–1 vector.*

Proof. In fact we prove that for any integer vector f the system $Ax = f$ has at most one solution. Furthermore, if it exists, this solution is integer.

The proof is by induction on the number of nodes in G . The result is trivial for a graph consisting of one or two nodes.

Now, let G be a chordal graph. Using the induction hypothesis we may assume that G is connected. Then G contains a simplicial node, [2], i.e., a node, say i , that belongs to exactly one clique. Therefore, i is associated with a unit row of A , and if $a_{ij} = 1$ we must set $x_j = f_i$ if $Ax = f$. We can then eliminate the i th equation of the system $Ax = f$. Let $N(i)$ be the set of neighbors of i in G . $N(i)$ is nonempty since G is connected.

Let G' be the induced subgraph obtained by omitting node i and all edges connecting it to members in $N(i)$. The node clique incidence matrix of G' , A' , is a submatrix of A , defined as follows. If the complete subgraph induced by the nodes in $N(i)$ is maximal in G' , then A' is the submatrix obtained by deleting the i th row of A . Otherwise, A' is obtained by deleting the i th row and the j th column of A .

Suppose first that A' is obtained by deleting only the i th row of A . Since G' is an induced subgraph it is chordal. By the induction hypothesis, the system $(Ax)_k = f_k$,

$\forall k \neq i$, is either inconsistent or else it has a unique solution, x' , which is also integer. Thus the original system $Ax = f$ is consistent if and only if x' exists and $x'_j = f_i$. The uniqueness of x' as a solution to the subsystem implies its uniqueness with respect to the system $Ax = f$.

Next, suppose A' has one less column than A . The system $Ax = f$ may be written as $(Ax)_k = f_k, \forall k \neq i, x_j = f_i$. Substituting f_i for x_j in each equation $(Ax)_k = f_k, \forall k \neq i$, we obtain exactly the subsystem corresponding to A' . Now we use the chordality of G' , and apply the induction hypothesis on the subsystem to conclude the validity of the result for the system $Ax = f$.

3. The location model. Given the tree T defined in the introduction, suppose that two finite subsets of T, Σ and Δ are specified. $\Sigma = \{y_1, \dots, y_n\}$ is called the supply set and $\Delta = \{x_1, \dots, x_m\}$ is the demand set. The demand points are to be served by centers which can be located only at points of Σ . Each demand point, x_i , must have at least a_i centers established at a distance not greater than $r_i \geq 0$ from it. Due to capacity constraints at most b_j centers can be located at y_j . The cost of establishing any center at y_j is $v_j \geq 0$. The problem is to find the minimum budget required for setting centers meeting the demand constraints.

We note that if T is replaced by a general (planar) network even a special case of the above model is known to be NP-hard, [6]. Turning back to a tree network, the demand constraints imply that for each $x_i, i = 1, \dots, m$, at least a_i centers should be set at the neighborhood subtree $T_i = \{x | d(x, x_i) \leq r_i\}$. Defining $S = \{T_1, \dots, T_m\}$, the location problem is formulated as

$$\begin{aligned}
 & \text{Minimize } \sum_{j=1}^n v_j z_j \\
 (1) \quad & \text{s.t. } Az \geq a, \\
 & b \geq z \geq 0 \text{ and integer,}
 \end{aligned}$$

where $A = A(S, \Sigma), a = (a_1, \dots, a_m), b = (b_1, \dots, b_n)$ and $e = (1, \dots, 1)$.

Certain instances of (1) have been considered in the literature. The special case of equal setting costs, v_j , for the centers and $a_i = 1, i = 1, \dots, m, b_j = \infty, j = 1, \dots, n$, can be solved in linear time by a modified version of the algorithm in [6]. A generalization of the latter special case, allowing arbitrary integer values for a_i is solved in [3]. There, the problem is reduced to finding a minimum cover of the nodes of $G(S), S = \{T_1, \dots, T_m\}$, by cliques, and observing the chordality of $G(S)$. The cliques are induced by the supply points. Applying the perfectness of $G(S)$ a dispersion location problem which is dual to this special case is also defined in [3]. Using only the perfectness property of A^T , the case considered in [3] was maximal in the sense that perfectness is equivalent to the existence of an integer solution to the linear program $\min \{\sum_{j=1}^n z_j | Az \geq a, z \geq 0\}$ for all nonnegative integers a , [4], [10].

The results of the previous sections, where the balancedness of $A = A(S, \Sigma)$ is proved, enable us to extend the class of "solvable" cases of (1).

We start with the special case of (1), where all the setting costs, v_j , are equal. This case is called the multiple coverage problem. Using [1] we note that the balancedness of A is equivalent to the existence of an integer solution to the linear program $\min \{\sum_{j=1}^n z_j | Az \geq a, b \geq z \geq 0\}$, for all nonnegative integer vectors a, b . Thus, the multiple coverage problem can be solved polynomially using Khachian's algorithm, [7], for linear programs. Also, we have constructed a direct algorithm for the multiple coverage model. Since this algorithm is based on simple extensions of the main ideas

embedded in the algorithms of [3], [6], we skip the description of our procedure. (The interested reader can obtain the detailed scheme from the author.) We mention that if, for example, the supply and demand sets consist only of nodes of the tree T , then the complexity of our direct algorithm is $O(n^2)$, where n is the number of nodes of T .

Secondly we consider the special case of (1) where $a_i = 1, i = 1, \dots, m$. (The constraints $z \leq b$ can be assumed to be redundant in this case.) The results in [5] ensure that all the extreme points of $\{z | Az \geq e, z \geq 0\}$ are integral. Thus, again the problem can be solved polynomially using Khachian's algorithm, provided the v_j are rational. (Khachian's algorithm may find an optimal solution which is not extreme and therefore may not be integer. However, an optimal extreme point to a linear program can always be generated in polynomial time if some optimal solution is available.) In the next section we will present a direct algorithm for solving this case.

We now summarize the results on the location model (1). To our knowledge no efficient algorithms to solve (1) are available. Verifying whether this problem is polynomially solvable will require a different approach than the one presented above for the special cases. This is due to the fact that the integer solution to (1) may not be optimal to the relaxed linear program. This is illustrated by the following.

Example 3. Let T be given by Fig. 3. Suppose that $\Sigma = \Delta = \{x_1, x_2, x_3, x_4\}$ with $d(x_i, x_4) = 1, i = 1, 2, 3$. Also let $r_i = 1, i = 1, 2, 3, 4$. Finally set $b = e, a_i = v_i = 1, i = 1, 2, 3$, and $a_4 = v_4 = 2$. We then have that the solution to (1) is 3 while the optimal objective of the relaxed linear program is 2.5.

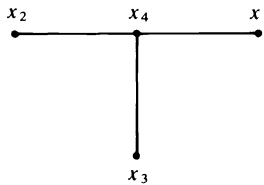


FIG. 3

Combining the results of the previous section with those of [1], [5], [7] our work shows the existence of efficient algorithms when either $a = e$, or the setting costs, v_j , are equal.

Finally we note another solvable case of (1), which is not implied by the above. If the matrix A is totally unimodular the model can now be solved efficiently by [7], if all data are rational. Total unimodularity is achieved, for example, by a tree which is a simple path. In this case the resulting graph is an interval graph.

4. Solving the location problem. In this section we present a direct algorithm for solving the location problem (1) described in the previous section, with $a_i = 1, i = 1, \dots, m$. To simplify the presentation we consider here the following special case. Given the tree $T = (N, E)$ with N and E the sets of nodes and edges respectively, suppose that $\Sigma = \Delta = N$, i.e., demand and supply occur at the nodes only. Given $r_i \geq 0, i \in N$ we wish to minimize the budget for setting centers such that each demand point is covered by a center, i.e., each $i \in N$ is at a distance of at most r_i from some center.

To present the algorithm we first assume that the tree is rooted at some distinguished node, say v . For each node $j \in N$ define $B(j)$ as the set of descendants of j , i.e., the entire set of nodes having j on the path connecting them with v . In particular $j \in B(j)$. Also define $S(j)$ to be the set of "sons" of j , i.e., the nodes having j as the immediate predecessor on the path connecting them with v . $T(j)$ will denote the minimal subtree containing $B(j)$.

Let $j \in N$. Suppose that a center already exists at some node in $N - B(j)$ whose distance from j is t . (If more than one center exists in $N - B(j)$ consider only the closest to j .) This center may clearly cover some nodes of $B(j)$. Suppose, further, that no centers exist in $B(j)$. Now define $h(j, t, s)$ to be the minimum budget required to cover the nodes of $T(j)$, given that new centers are set at $B(j)$ only, with the closest being at a distance s from j , and the closest existing center in $N - B(j)$ is at a distance t from j .

Let $D(j)(F(j))$ be the set of distances from j to the members in $B(j)(N - B(j))$. Also the value $s = \infty(t = \infty)$ indicates that no center is set at $B(j)(N - B(j))$. Define $\bar{D}(j) = D(j) \cup \{\infty\}$ and $\bar{F}(j) = F(j) \cup \{\infty\}$. Then $h(j, t, s)$ is defined only for $s \in \bar{D}(j)$ and $t \in \bar{F}(j)$. Furthermore, we compute $h(j, t, s)$ only for $t \leq s$ since $h(j, t, s) = h(j, s_j^*, s)$ for all $t \geq s$ in $\bar{F}(j)$. (s_j^* is the smallest element in $\bar{F}(j)$ which is not smaller than s .)

Defining $H(j, t, s) = \min_{p \geq s} h(j, t, p)$, we obtain $H(j, t, s) = H(j, t, \lceil s \rceil_j)$, where $\lceil s \rceil_j$ is the smallest element in $\bar{D}(j)$ which is not smaller than s . The answer to the location problem is given by $H(v, \infty, 0)$. Our algorithm is based on a recursive computation of $h(j, t, s)$ leading to $H(v, \infty, 0)$.

Starting with the tips of the rooted tree we obtain the following recursion for $t \in \bar{F}(j)$, $s \in \bar{D}(j)$ and $t \leq s$.

If j is a tip, then $h(j, t, 0) = v_j$ and

$$h(j, t, \infty) = \begin{cases} 0 & \text{if } t \leq r_j, \\ \infty & \text{if } t > r_j. \end{cases}$$

Suppose j is not a tip; then

$$h(j, t, \infty) = \begin{cases} 0 & \text{if } d(i, j) + t \leq r_i \text{ for all } i \in B(j), \\ \infty & \text{otherwise,} \end{cases}$$

$$h(j, t, 0) = v_j + \sum_{i \in S(j)} H(i, d(i, j), 0),$$

and for $0 \neq s \in D(j)$

$$h(j, t, s) = \begin{cases} \infty & \text{if } t > r_j, \\ \min_{\substack{i \in S(j) \text{ and} \\ s - d(i, j) \in D(i)}} \left\{ h(i, t + d(i, j), s - d(i, j)) + \sum_{\substack{k \in S(j) \\ k \neq i}} H(k, t + d(k, j), s - d(k, j)) \right\} & \text{otherwise,} \end{cases}$$

when $t \leq r_j$.

Simplifying the expression for $t \leq r_j$ we obtain

$$h(j, t, s) = \sum_{k \in S(j)} H(k, t + d(k, j), \lceil s - d(k, j) \rceil_k) + \min_{\substack{i \in S(j) \text{ and} \\ s - d(i, j) \in D(i)}} \{ h(i, t + d(i, j), s - d(i, j)) - H(i, t + d(i, j), s - d(i, j)) \}.$$

Having established the recursive relations leading to the optimal solution we next demonstrate that the complexity of the suggested algorithm is $O(n^3)$ when n is the number of nodes of T .

In the initial phase we generate and sort each one of the sets $D(j), F(j), j \in N$. This will consume $O(n^2 \log n)$ time.

Now, given j we show that the total effort needed to compute $h(j, t, s)$ for all $t \in \bar{F}(j)$ and $s \in \bar{D}(j), t \leq s$, is $O(n^2 |S(j)|)$.

First, for each $k \in S(j)$ compute $\lceil s - d(k, j) \rceil_k$, for all $s \in D(j)$. This will enable us to use previously computed values of the functions $h(k, \cdot, \cdot)$ and $H(k, \cdot, \cdot)$ for $k \in S(j)$. Since $D(j)$ and $D(k)$ are already sorted this step is done in $O(n)$ time for each $k \in S(j)$, or in $O(n|S(j)|)$ for all $k \in S(j)$.

Next, for each $s \in D(j)$ the set of indices with $s - d(i, j) \in D(i)$ is found. Like the preceding step this is performed in $O(n|S(j)|)$ time for all $s \in D(j)$.

Finally we turn to a given pair (t, s) with $t \in F(j)$, $s \in D(j)$. Using the recursive relations and the information acquired in the previous steps $h(j, t, s)$ is computed in $O(|S(j)|)$ time. Thus the effort for computing $h(j, t, s)$ and $H(j, t, s)$ for all pairs (t, s) , $t \in F(j)$, $s \in D(j)$ is $O(n^2|S(j)|)$, and the bound for the entire algorithm becomes $O(n^3)$.

It is easily verified that this bound is not affected if one also wishes to find the optimal locations of the centers yielding the minimum budget. The space required for implementing the algorithm is also $O(n^3)$.

We have provided an efficient procedure to solve the location problem where it is required to minimize the budget for covering each demand point. This procedure can now be used to solve the following related problem.

Suppose that the total budget available for setting centers at the supply points is $B > 0$. Given this constraint one wishes to establish centers such that the maximum distance from a demand point to its nearest center is minimized.

It is clear that the minimum of the maximum distance is an element in the set

$$R = \{d(x_i, y_j) \mid x_i \in \Delta, y_j \in \Sigma\}.$$

Hence the optimal value is the minimum element $r \in R$ such that the minimum budget, needed to ensure that each demand point is covered within a radius r does not exceed B . The procedure described above will be used to determine for any given r whether the respective budget exceeds B . To find the optimal value we can use the sophisticated search on the set R which is used in [9] to find an optimal element in the case where the setting costs, v_j are equal.

Note added in proof. We note that a special case of Theorem 1 is proved in R. Giles, *A balanced hypergraph defined by certain subtrees of a tree*, *Ars Combinatoria*, 6 (1978), pp. 179–183.

REFERENCES

- [1] C. BERGE, *Balanced matrices*, *Math. Programming*, 2 (1972), pp. 19–31.
- [2] P. BUNEMAN, *A characterization of rigid circuit graphs*, *Discrete Math.*, 9 (1974), pp. 205–212.
- [3] R. CHANDRASEKARAN AND A. TAMIR, *Polynomially bounded algorithms for locating p centers on a tree*, *Math. Programming*, 22 (1982), pp. 304–315.
- [4] D. R. FULKERSON, *Blocking and anti-blocking pairs of polyhedra*, *Math. Programming*, 1 (1971), pp. 168–194.
- [5] D. R. FULKERSON, A. J. HOFFMAN AND R. OPPENHEIM, *On balanced matrices*, *Math. Programming Study*, 1 (1974), pp. 120–132.
- [6] 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.
- [7] L. G. KHACHIAN, *A polynomial algorithm in linear programming*, *Dokl. Akad. Nauk USSR*, 244, 5, Feb. (1979).
- [8] L. LOVASZ, *Normal hypergraphs and the perfect graph conjecture*, *Discrete Math.*, 2 (1972), pp. 253–267.
- [9] N. MEGIDDO, A. TAMIR, E. ZEMEL AND R. CHANDRASEKARAN, *An $O(n \log^2 n)$ algorithm for the k th longest path in a tree with applications to location problems*, *SIAM J. Comput.*, 10 (1981), pp. 328–337.
- [10] M. W. PADBERG, *Perfect zero-one matrices*, *Math. Programming*, 6 (1974), pp. 180–196.

ISOLATING ERROR EFFECTS IN SOLVING ILL-POSED PROBLEMS*

C. MARK AULICK† AND THOMAS M. GALLIE‡

Abstract. Many ill-posed problems are reduced to a matrix equation, usually very ill-conditioned, which is then solved using the smoothing techniques of regularization. Any such smoothing will introduce bias into the calculated solution in the sense that if the data were exact, the calculated solution will not be the “exact” solution. Since this calculated solution is also affected by error in the data, we show how these two error effects may be isolated and considered separately. Using a very general form of the regularization technique, we derive exact formulas for each error component which illustrates the dependence of each upon the different variables and parameters of the problem.

1. Introduction. When solving the matrix equation

$$(1.1) \quad Ax = \hat{b},$$

where A is a known $m \times n$ ill-conditioned matrix of rank n , x is to be calculated, and $\hat{b} = b - \varepsilon$ is an m -vector of data values subject to error (ε is the vector of measurement error), some sort of smoothing usually is performed in order to make the calculated solution \hat{x} less sensitive to the error ε . This smoothing action, however, will also contribute to the total error of the solution by introducing “bias”; that is, as ε tends to 0, the calculated solution \hat{x} will not approach the true solution x_0 (defined uniquely by $Ax_0 = b$). This effect has been noted by many researchers (see, for example, [JACK79] and [VARA73]), especially with regard to the smoothing technique of *regularization*.

We would like to consider two different forms of regularizing functions and how they affect the two different components of the total error $e = x_0 - \hat{x}$. We shall call these error components *regularization error* (which we shall denote by e_R) and *noise amplification error* (denoted by e_N). The first of these components deals with the bias introduced by smoothing; the second involves the effects of ε on the calculated solution \hat{x} .

We define these two components as follows: we assume that \hat{x} is given as a linear function of the data vector \hat{b} : $\hat{x} = C\hat{b} + d$. The total error of the solution is $x_0 - \hat{x}$, where x_0 is the unknown “true solution”; hence the total error e is equal to $x_0 - Cb - d + C\varepsilon$. To obtain the “bias” (e_R), we set $\varepsilon = 0$ and obtain $e_R = x_0 - (Cb + d)$. Then $e_N = e - e_R = C\varepsilon$.

Separate consideration of these error effects may be useful in trying to select a method for a particular application or in comparing the behavior of methods when varying certain problem or solution parameters. Since $\|e\| \cong \|e_R\| + \|e_N\|$, it is likely that a method which “balances” the norms of the two error components will perform well in terms of minimizing the total error of the method. It is also possible that insight gained from this study will lead to ways of improving the performance of known solution methods or to the development of new ones.

* Received by the editors June 7, 1982. This research was supported in part by the National Institutes of Health under grant HL 11307. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26–29, 1982.

† Department of Mathematics and Computer Science, Louisiana State University, Shreveport, Louisiana 71115.

‡ Department of Computer Science, Duke University, Durham, North Carolina 27706.

The regularization methods we consider arise from reformulating the original problem [1.1] into a constrained minimization:

$$(1.2) \quad \text{minimize } \|Lx - k\|^2,$$

$$(1.3) \quad \text{subject to } \|Ax - \hat{b}\|^2 \leq \mu^2,$$

where L , k and μ^2 are “parameters”; A and \hat{b} are from the original problem [1.1]. The choices of L and k will determine the particular type of smoothing to be done, and μ^2 will reflect something of the measurement error in \hat{b} . Since the optimum of [1.2] gives inequality in [1.3] only if k is very close to Lx_0 or if μ^2 is too large, we assume that the minimum occurs when the constraint [1.3] is tight. (See [AULI81a, b] and [TWOM65].)

Using the technique of Lagrange multipliers, it can be shown in the case of the equality constraint that the desired solution is

$$(1.4) \quad \hat{x} = (A^T A + tL^T L)^{-1} (A^T \hat{b} + tL^T k)$$

for some positive constant t . Since A is of full rank, the indicated matrix is invertible.

At this point we remark that we have assumed that b is in the range of A . This means that in the absence of error the exact solution to [1.1] is $x_0 = A^+ b$, where A^+ denotes the pseudo-inverse of A [PENR56]. Although in practice the assumption that A is full-rank may not be valid, it will give a starting point from which we can proceed in deriving our results.

Two tools which are central to our analysis are the singular value decomposition (s.v.d.) [LAWS74] and the generalized singular value decomposition (g.s.v.d.) [VANL74], [PAIG81]. To summarize these briefly, the s.v.d. of a matrix A allows us to write

$$(1.5) \quad A = QSR^T,$$

where Q and R are orthogonal matrices, and S is a $m \times n$ diagonal matrix of nonnegative entries arranged in nonincreasing order. The diagonal elements of S are called the *singular values* of A and are equal to the positive square roots of the eigenvalues of $A^T A$.

The g.s.v.d. allows us to relate two matrices in the following way: given A and L , where A is $m \times n$ and L is $p \times n$, we can write

$$(1.6) \quad A = XS_1 Z^{-1},$$

$$(1.7) \quad L = YS_2 Z^{-1},$$

where X and Y are orthogonal, Z is nonsingular and has columns of unit length, and S_1 and S_2 are diagonal matrices of nonnegative entries (S_1 and S_2 are $m \times n$ and $p \times n$, respectively). Furthermore, the diagonal elements of S_1 are in nonincreasing order. If L is the identity, then S_2 is also the identity with $Z = Y$ and $S_1 = S$; formulas [1.6] and [1.7] reduce to [1.5] in this case.

Among other things, these decompositions may be used to calculate the pseudo-inverses of A and L . From [1.5], we get

$$(1.8) \quad A^+ = RS^+ Q^T = R \operatorname{diag} \left(\frac{1}{\gamma_i} \right) Q^T,$$

where $S = \operatorname{diag} [\gamma_i]$; the γ_i 's are the nonzero singular values of A . Similarly, if we let

$S_1 = \text{diag} [\alpha_i]$ and $S_2 = \text{diag} [\beta_i]$, with the α_i 's and β_i 's nonzero, using [1.6] and [1.7] we can write

$$(1.9) \quad A^+ = Z \text{diag} [1/\alpha_i] X^T,$$

$$(1.10) \quad L^+ = Z \text{diag} [1/\beta_i] Y^T.$$

These formulas will be used widely in the derivations of our results.

2. Error expressions when $k = 0$. When the vector k is zero, we have the most common form of regularization:

$$(2.1) \quad \hat{x} = C\hat{b} = (A^T A + tL^T L)^{-1} A^T \hat{b}.$$

Our definitions of e_R and e_N yield the following:

$$e_R = x_0 - Cb = x_0 - (A^T A + tL^T L)^{-1} A^T b,$$

$$e_N = C\varepsilon = (A^T A + tL^T L)^{-1} A^T \varepsilon.$$

We first give results for the general case (L arbitrary) and consider separately the case that $L = I$ (see [TIKH65]).

Although Theorem 1 is not new (see [BJOR79]), we shall need this result later and so we include a proof, illustrating some of the techniques we use.

THEOREM 1. *The inverse matrix C in [2.1] is equal to $Z \text{diag} [\alpha_i / (\alpha_i^2 + t\beta_i^2)] X^T$, where α_i, β_i, Z and X are given from [1.9] and [1.10].*

Proof. Let $A = XS_1Z^{-1}$ and $L = YS_2Z^{-1}$ be the g.s.v.d. of A and L . Then the inverse matrix C is

$$C = (Z^{-T} [S_1^T S_1 + tS_2^T S_2] Z^{-1})^{-1} Z^{-T} S_1^T X^T.$$

Since S_1 and S_2 are diagonal matrices, the expression inside the square brackets is also a diagonal matrix; namely $\text{diag} [\alpha_i^2 + t\beta_i^2]$. Since none of the diagonal elements are zero, the matrix to be inverted is nonsingular and we have

$$C = Z \text{diag} \left[\frac{1}{\alpha_i^2 + t\beta_i^2} \right] Z^T Z^{-T} S_1^T X^T = Z \text{diag} \left[\frac{1}{\alpha_i^2 + t\beta_i^2} \right] S_1^T X^T.$$

Since S_1 is diagonal, this final expression reduces to

$$(2.2) \quad C = Z \text{diag} \left[\frac{\alpha_i}{\alpha_i^2 + t\beta_i^2} \right] X^T$$

as desired.

This result gives a means for determining when the inversion of $(A^T A + tL^T L)$ may be accomplished "safely" even when A is rank-deficient. Furthermore, this expression provides additional insight when we use it in our definitions of e_R and e_N .

THEOREM 2. *The regularization error e_R is equal to*

$$Z \text{diag} \left[\frac{t\beta_i^2}{\alpha_i(\alpha_i^2 + t\beta_i^2)} \right] X^T b.$$

Proof.

$$\begin{aligned}
 e_R &= x_0 - Cb = x_0 - Z \operatorname{diag} \left[\frac{\alpha_i}{\alpha_i^2 + t\beta_i^2} \right] X^T b \\
 &= x_0 - \left(Z \operatorname{diag} [1/\alpha_i] X^T b - Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i(\alpha_i^2 + t\beta_i^2)} \right] X^T b \right) \\
 &= (x_0 - A^+ b) + Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i(\alpha_i^2 + t\beta_i^2)} \right] X^T b.
 \end{aligned}$$

Since we have assumed that b is in the range of A , the term $x_0 - A^+ b$ is zero, and the desired result follows.

COROLLARY 2.1. *If $L = I$, then $e_R = R \operatorname{diag} [t/(\gamma_i(\gamma_i^2 + t))] Q^T b$, where R, Q and γ_i are from (1.5).*

This theorem and its corollary may give useful insight into how the calculated solution behaves as a function of t . We note first that as t approaches zero, we are doing less and less smoothing and hence less bias appears in the answer. Also, as t approaches infinity, all information is “smoothed out” of the data and \hat{x} will go to zero; hence e_R will approach $x_0 = A^+ b$.

We also see the effects of the β_i 's in the formula. As previously noted, the α_i 's are arranged in nonincreasing order down the diagonal of A . However, in practice the β_i 's are usually arranged in “approximately” nondecreasing order down the diagonal of S_2 ; hence large β_i 's will tend to be associated with small α_i 's, and vice versa. The further importance of this arrangement will be seen when we consider the results of the following theorem and its corollary.

THEOREM 3. *The noise amplification error e_N is equal to $C\varepsilon = Z \operatorname{diag} [\alpha_i/(\alpha_i^2 + t\beta_i^2)] X^T \varepsilon$.*

Proof. This follows directly from the definition of e_N and Theorem 1.

COROLLARY 3.1. *If $L = I$, then $e_N = R \operatorname{diag} [\gamma_i/(\gamma_i^2 + t)] Q^T \varepsilon$.*

We observe from Theorem 3 the damping effects of the generalized singular values of L . Since the large β_i 's will tend to be associated with the small α_i 's, the noise-amplifying effects of the reciprocals of these small values will be lessened. Also, since the small or zero β_i 's will usually be associated with the large α_i 's, the “signal-carrying” effects of the large generalized singular values of A will not be impaired too much.

One possible criticism of these results is that they require exact knowledge of b or ε , quantities usually unavailable. However, numerical experiments we have performed [AULI81a] indicate that \hat{b} may be substituted for b in Theorem 2, and the results are close to those calculated using the exact b if t is not too small. Dealing with the requirement of Theorem 3 (knowing ε) is more difficult. If a bound for $\|\varepsilon\|$ is known, we can derive an upper bound for $\|e_N\|$, but typically such bounds are quite loose.

Finally, we remark that the results from Theorems 1, 2 and 3 do not require any matrix inversion to vary different “parameters”—notably, the regularization parameter t . Thus it would be a simple matter to find the value of t for which the norms of e_R and e_N are balanced.

3. Error expressions when $k \neq 0$. Only a few solution methods have appeared in the literature in which k is not zero (e.g., [TWOM65]). However, such methods have an intuitive attraction since it is often the case that we have an idea of the shape of the solution or its expected value. (Refer to [AULI81b] for further discussion.)

We observe that the intuitive interpretation of the function in [1.2] is to constrain $L\hat{x}$ to be “close to” k . Thus as k is close to Lx_0 (where x_0 is the “true” solution), we would like to have the regularization error, or bias, of the solution be small. In this case the definition of e_R gives

$$e_R = x_0 - (A^T A + tL^T L)^{-1} (A^T b + tL^T k).$$

The formula for e_N , however, remains the same as in the previous section; thus Theorem 3 and Corollary 3.1 still apply.

Intuition suggests that as k improves as an approximation to Lx_0 , then e_R should decrease. This is true.

THEOREM 4. *When $k \neq 0$, e_R is a homogeneous linear function of $x_0 - L^+ k$:*

$$e_R = Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i + t\beta_i^2} \right] Z^{-1} (x_0 - L^+ k).$$

Proof. Using the generalized singular-value decomposition, we have

$$(A^T A + tL^T L)^{-1} = Z \operatorname{diag} [1/(\alpha_i^2 + t\beta_i^2)] Z^T.$$

Using this expression, we have

$$\begin{aligned} e_R &= x_0 - \left(Z \operatorname{diag} \left[\frac{\alpha_i}{\alpha_i^2 + t\beta_i^2} \right] X^T b + tZ \operatorname{diag} \left[\frac{\beta_i}{\alpha_i^2 + t\beta_i^2} \right] Y^T k \right) \\ &= x_0 - Z \operatorname{diag} [1/\alpha_i] X^T b + Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i(\alpha_i^2 + t\beta_i^2)} \right] X^T b \\ &\quad - Z \operatorname{diag} \left[\frac{t\beta_i}{\alpha_i^2 + t\beta_i^2} \right] Y^T k. \end{aligned}$$

Since $Z \operatorname{diag} [1/\alpha_i] X^T = A^+$, the first two terms cancel and we are left with

$$\begin{aligned} (3.1) \quad e_R &= Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i(\alpha_i^2 + t\beta_i^2)} \right] X^T b - Z \operatorname{diag} \left[\frac{t\beta_i}{\alpha_i^2 + t\beta_i^2} \right] Y^T k \\ &= Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i^2 + t\beta_i^2} \right] Z^{-1} (Z \operatorname{diag} [1/\alpha_i] X^T b - Z \operatorname{diag} [1/\beta_i] Y^T k) \\ &= Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i^2 + t\beta_i^2} \right] Z^{-1} (A^+ b - L^+ k) \\ (3.2) \quad &= Z \operatorname{diag} \left[\frac{t\beta_i^2}{\alpha_i^2 + t\beta_i^2} \right] Z^{-1} (x_0 - L^+ k). \end{aligned}$$

The value of k which minimizes the norm of $x_0 - L^+ k$ is $k = Lx_0$. Thus our intuition, which tells us that if k is a good estimate of Lx_0 , then e_R should be small, is borne out. However, it is not necessarily true that $k = Lx_0$ is the value of k which minimizes e_R . If L has rank less than n , where n is the number of columns in L , then $L^+ Lx_0$ is not equal to x_0 , although $x_0 - L^+ k$ is of minimal norm when $k = Lx_0$ (in the sense that there is no other value of k for which $L^+ k$ is closer to x_0). However, since Z is not generally orthogonal, this value of k does not guarantee a minimal norm for e_R . It is possible that there is another k for which $Z^{-1}(x_0 - L^+ k)$ is closer to zero.

COROLLARY 4.1. *If $\operatorname{rank}(L) = n$ and if $k = Lx_0$, then $e_R = 0$.*

Proof. If $\operatorname{rank}(L) = n$, then $L^+ L = I_n$ and $x_0 - L^+ k = x_0 - L^+ Lx_0 = x_0 - x_0 = 0$. The result then follows from Theorem 4.

This result may be especially important for the Twomey method in which $L = I$. Thus as k improves as an approximation to x_0 , e_R will decrease to zero. Also, for this particular method, since $L = I$, Z is orthogonal and the result is even stronger since an orthogonal linear transformation is norm-preserving.

Although the expression [3.2] in the proof of Theorem 4 appears to require knowledge of x_0 , such is not the case. We could write $x_0 = A^+b'$ and leave it at that; however, estimating b by \hat{b} in this case will not work as well since the error in \hat{b} will tend to be greatly magnified by A^+ . It might be better to use [3.1], since some "smoothing" is being applied and replacement of b by \hat{b} will not have too severe an effect (especially if t is not too small).

Finally, we remark that the expression for e_R derived in Theorem 4 reduces to the formula for e_R in Theorem 2 if $k = 0$. All that is needed to see this is the fact that $Z^{-1}x_0 = Z^{-1}A^+b = \text{diag}[1/\alpha_i]X^Tb$. This may be easily confirmed by considering the generalized singular-value decomposition of A^+ .

4. Conclusion. We have derived some results which may be useful in analyzing and comparing solution techniques for ill-conditioned linear systems. Although our focus has been on regularization methods, we believe that our definitions and analysis techniques may be profitably applied to other methods (e.g., the truncated singular-value decomposition [HANS71], [VARA73]).

REFERENCES

- [AULI81a] C. MARK AULICK, *Numerical techniques for inverse problems*, Ph.D. dissertation, Duke University, Durham, NC, 1981.
- [AULI81b] ———, *A generalized approach to regularization*, Tech Rep. CS-1981-5, Duke University Department of Computer Science, Durham, NC, 1981.
- [BJOR79] AKE BJORK, AND LARS ELDEN, *Methods in numerical algebra for ill-posed problems*, Report LiTH-MAT-R-33-1979, Linkoping University, Sweden, 1979. Paper presented at the International Symposium on Ill-Posed Problems: Theory and Practice, held at the University of Delaware, October 1979.
- [HANS71] RICHARD J. HANSON, *A numerical method for solving Fredholm integral equations of the first kind using singular values*, SIAM J. Numer. Anal., 8 (1971), pp. 616–622.
- [JACK79] DAVID D. JACKSON, *The use of a priori data to resolve non-uniqueness in linear inversion*, Geophysical J. Royal Astron. Soc., 57 (1979), pp. 137–157.
- [LAWS74] CHARLES L. LAWSON, AND RICHARD J. HANSON, *Solving Least-Squares Problems*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.
- [PAIG81] C. C. PAIGE, AND M. A. SAUNDERS, *Toward a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.
- [PENR56] R. PENROSE, *On best approximate solutions of linear matrix equations*, Proc. Cambridge Philos. Soc., 52 (1956), pp. 17–19.
- [TIKH65] A. N. TIKHONOV, *Incorrect problems of linear algebra and a stable method for their solution*, Soviet Math. Doklady, 6 (1965), pp. 988–991.
- [TWOM65] S. TWOMEY, *The application of numerical filtering to the solution of integral equations encountered in indirect sensing measurements*, J. Franklin Inst., 279 (1965), pp. 95–109.
- [VANL76] CHARLES F. VAN LOAN, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.
- [VARA73] JAMES M. VARAH, *On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems*, SIAM J. Numer. Anal., 10 (1973), pp. 259–267.

MATRIX DIAGONAL STABILITY AND ITS IMPLICATIONS*

ABRAHAM BERMAN† AND DANIEL HERSHKOWITZ†

Abstract. Relations between diagonal stability, stability, positiveness of principal minors and semipositiveness are described for several classes of matrices. In particular, it is shown that for matrices whose nondirected graph is acyclic, positiveness of principal minors is equivalent to diagonal stability.

Key words. Diagonally stable, stable, P -matrix, semipositive, forest, tree, tridiagonal, totally nonnegative, oscillation, ω -matrices

Introduction. The matrices in this paper are real and square. Following [4] and [3] we consider four classes of matrices:

- $\mathcal{A} = \{A; \text{there exists a positive definite diagonal matrix } D \text{ such that } AD + DA^T \text{ is positive definite}\}$ —the *diagonally stable matrices* [1], also known as the *Volterra-Lyapunov stable matrices* [5],
- $\mathcal{L} = \{A; \text{there exists a positive definite matrix } X \text{ such that } AX + XA^T \text{ is positive definite}\}$ —the *(positive) stable matrices*,
- $\mathcal{P} = \{A; \text{all the principal minors of } A \text{ are positive}\}$ —the P -matrices,
- $\mathcal{S} = \{A; \text{there exists a positive vector } x \text{ such that } Ax \text{ is positive}\}$ —the *semipositive matrices* [7], [15].

Diagonally stable matrices play an important role in various applications, for example, predator-prey systems and economic models (see for example [10], [14] and the references in [1]). A useful characterization of such matrices is that $A \in \mathcal{A}$ if and only if for every nonzero symmetric positive semidefinite matrix B , the matrix BA has a positive diagonal element.

Usually, positive stable matrices are defined as matrices whose eigenvalues have positive real part. Such a condition is of great importance in the study of equilibrium states of physical systems. The definition of \mathcal{L} , used above to point out the relation to \mathcal{A} , is the classical characterization of stable matrices due to Lyapunov [13]. Matrices with positive principal minors appear in economics and mathematical programming, e.g., [2], while semipositive matrices are of interest in numerical analysis, e.g., [15].

In this paper we study the inclusion relations between the four classes, continuing the work begun in [4]. To facilitate the description of these relations we use the letters \mathcal{A} , \mathcal{L} , \mathcal{P} and \mathcal{S} also to denote the properties of being a diagonally stable matrix, a stable matrix, a P -matrix and a semipositive matrix, respectively. For example, $\mathcal{P} \Rightarrow \mathcal{S}$ (e.g., [7]), means that a P -matrix is semipositive.

The main results of the paper deal with matrices whose graphs are acyclic, in particular, tridiagonal matrices. They are given in § 3. Section 2 contains introductory results, mostly known. Some partial results and open questions on ω -matrices (see the definition in § 4) and on matrices with real spectra are introduced in the last section.

2. Introductory results. The relations between the four properties will now be described for several classes of matrices, using implication diagrams. The completeness of the diagrams is demonstrated by examples.

* Received by the editors June 1, 1982, and in revised form September 20, 1982. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26-29, 1982. This research was supported by the Fund for Promotion of Research at the Technion.

† Department of Mathematics, Technion-Israel Institute of Technology, Haifa 32000, Israel.

THEOREM 1.

a. In general

$$\mathcal{L} \Leftarrow \mathcal{A} \Rightarrow \mathcal{P} \Rightarrow \mathcal{S},$$

b. For Z -matrices, i.e., matrices with nonpositive off-diagonal entries,

$$\mathcal{A} \Leftrightarrow \mathcal{L} \Leftrightarrow \mathcal{P} \Leftrightarrow \mathcal{S}.$$

c. For symmetric matrices

$$\mathcal{A} \Leftrightarrow \mathcal{L} \Leftrightarrow \mathcal{P} \Rightarrow \mathcal{S}.$$

d. For triangular matrices

$$\mathcal{A} \Leftrightarrow \mathcal{L} \Leftrightarrow \mathcal{P} \Rightarrow \mathcal{S}.$$

e. for normal matrices

$$\mathcal{A} \Leftrightarrow \mathcal{L} \Rightarrow \mathcal{P} \Rightarrow \mathcal{S}.$$

The absence of an implication in the above relations means that a counterexample exists.

Proof. The implications are well known (e.g., [4], [3], [5] and the references included there). The Z -matrices which satisfy any of the equivalent properties \mathcal{A} , \mathcal{L} , \mathcal{P} or \mathcal{S} are the nonsingular M -matrices (e.g., [2]). The symmetric matrices which satisfy any of the equivalent properties \mathcal{A} , \mathcal{L} or \mathcal{P} are the positive definite matrices.

Consider the following examples.

Example 1.

$$A = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{spec}(A) = \{-\sqrt{2}, +\sqrt{2}\},$$

Example 2.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

Example 3.

$$A = \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix} \quad \text{spec}(A) = \{1, 1\}$$

Example 4.

$$A = \begin{pmatrix} 1 & 4 & 0 \\ 0 & 1 & 4 \\ 4 & 0 & 1 \end{pmatrix},$$

$$\text{spec}(A) = \{5, -1 \pm 2\sqrt{3}i\}, \quad AA^T = A^T A = \begin{pmatrix} 17 & 4 & 4 \\ 4 & 17 & 4 \\ 4 & 4 & 17 \end{pmatrix}.$$

Example 1 shows that $\mathcal{S} \not\Rightarrow \mathcal{P}$ in (a), (c) and (e). Example 2 shows the same in (d). Example 3 shows that $\mathcal{L} \not\Rightarrow \mathcal{S}$ in (a) and Example 4 shows that $\mathcal{P} \not\Rightarrow \mathcal{L}$ in (a) and (e). \square

3. Matrices whose graph is a forest. A class of matrices which appears in many applications and has interesting properties is the one of *tridiagonal* matrices, also called Jacobi matrices [9]; i.e., matrices A such that $|i - j| > 1 \Rightarrow a_{ij} = 0$. The main result of this section is that in this case P -matrices are stable and even diagonally stable.

This, in fact, is a property of the combinatorial structure of the matrix.

Recall that an acyclic graph is called a *forest*. A connected forest is a *tree*. The nondirected graph $G(A)$ of an $n \times n$ matrix A has n vertices $1, \dots, n$ and an edge between i and j , $i \neq j$, if and only if $a_{ij} \neq 0$ or $a_{ji} \neq 0$. The directed graph $DG(A)$ of such a matrix has n vertices $1, \dots, n$ and an arc from i to j if and only if $a_{ij} \neq 0$. A matrix A is *treediagonal* [12] if $G(A)$ is a tree. If $G(A)$ is a linear path then A is cogredient (equivalent via a simultaneous permutation of rows and columns) to a tridiagonal matrix.

THEOREM 2. *If all the principal minors of A are positive and if the nondirected graph of A is a forest, then A is diagonally stable. Thus, for matrices whose nondirected graph is a forest*

$$\mathcal{L} \Leftarrow \mathcal{A} \Leftrightarrow \mathcal{P} \Rightarrow \mathcal{S}.$$

Here too, the absence of an implication means the existence of a counterexample.

Proof. Example 1 shows that $\mathcal{S} \not\Rightarrow \mathcal{L}$ and Example 3 shows that $\mathcal{L} \not\Rightarrow \mathcal{P}$.

The fact that $\mathcal{P} \Rightarrow \mathcal{A}$ is proved by induction on the order of the matrix A .

The claim is trivial for $n = 1$. Assume it holds for matrices of order less than n and let A be of order n .

Case 1. A is a combinatorial symmetric, i.e., $a_{ij} \neq 0 \Leftrightarrow a_{ji} \neq 0$. In this case we show that A is diagonally stable by constructing a positive definite diagonal matrix D such that $AD + DA^T$ is positive definite. If $G(A)$ is not connected then A is cogredient to the direct sum $A_1 + A_2$, where A_1 and A_2 satisfy the induction assumption. Let D_1 and D_2 be positive definite diagonal matrices such that $A_1 D_1 + D_1 A_1^T$ and $A_2 D_2 + D_2 A_2^T$ are positive definite. Then $(A_1 + A_2)(D_1 + D_2) + (D_1 + D_2)(A_1 + A_2)^T$ is positive definite.

When $G(A)$ is a tree, D can be constructed by the following algorithm which assigns positive numbers to the vertices:

Initial step: Set $d_1 = 1, d_2 = \dots = d_n = 0$.

Step $k + 1$: If d_i became positive in step k , i and j are neighbors in $G(A)$, ($G(A)$ contains an edge between i and j) and $d_j = 0$, set

$$d_j = \left| \frac{a_{ji}}{a_{ij}} \right| d_i.$$

Final step: If $d_i > 0, i = 1, \dots, n$, set

$$D = \text{diag} \{d_i\}, \text{ STOP.}$$

The algorithm reaches the final step since $G(A)$ is connected. Also the number d_j is well defined as there is a unique path from 1 to j since $G(A)$ is a tree. (If $1 = i_1, \dots, i_{k+1} = j$ is this unique path, then

$$d_j = \left| \frac{a_{i_2 i_1}}{a_{i_1 i_2}} \dots \frac{a_{i_{k+1} i_k}}{a_{i_k i_{k+1}}} \right|.$$

The matrix $C = AD$ is also a P -matrix, and since $G(A)$ is a tree, $|c_{ij}| = |c_{ji}|$ for all $i \neq j$. If for all i and j , $a_{ij} a_{ji} \geq 0$, then C is symmetric and thus positive definite so A is indeed diagonally stable. Otherwise, $C + C^T$ is cogredient to a direct sum of principal submatrices of $2C$ which are P -matrices, so $C + C^T$ is positive definite and again A is diagonally stable.

Case 2. There exist two indices i and j such that $a_{ij} = 0$ and $a_{ji} \neq 0$. In this case $DG(A)$ contains no path from i to j since $G(A)$ is acyclic. Thus A is reducible, i.e., cogredient to a matrix of the form

$$(1) \quad \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}.$$

where the blocks A_1 and A_3 are square. Since diagonal stability is not affected by simultaneous permutation of rows and columns we may assume that A is in the form of (1).

We now show that a reducible matrix in this form where A_1 and A_3 are diagonally stable is also diagonally stable. To do it we use the criterion mentioned in the introduction. Suppose B is a symmetric positive definite matrix such that $(BA)_{ii} \leq 0$, $i = 1, \dots, n$. Partition

$$B = \begin{pmatrix} B_1 & B_2 \\ B_2^T & B_3 \end{pmatrix}$$

in conformity with (1). The main diagonal of B_1A_1 is nonpositive. Then $B_1 = 0$ for A_1 is diagonally stable by the induction assumption. But then $B_2 = 0$ and

$$BA = \begin{pmatrix} 0 & 0 \\ 0 & B_3A_3 \end{pmatrix}.$$

Again, A_3 is diagonally stable by the induction assumption, thus $B_3 = 0$ so $B = 0$, proving that A is diagonally stable. \square

Examples 1 and 2 used in Theorem 2 are of 2×2 treediagonal matrices. Thus we have the following corollary.

COROLLARY. *The inclusion diagram of Theorem 2 holds for treediagonal matrices, tridiagonal matrices and 2×2 matrices.*

Remarks.

- 1) The equivalence $\mathcal{A} \Leftrightarrow \mathcal{P}$ for 2×2 matrices is well known (e.g., [10]).
- 2) A special case of combinatorial symmetric matrices, where $a_{ij}a_{ji} \leq 0$, is studied in [15]. Note that in this case $AD + DA^T$ is a diagonal matrix.

4. Matrices with real spectra and ω -matrices. The triangular and the symmetric matrices treated in Theorem 1 have real spectra. For matrices with real eigenvalues in general we have the following diagram.

$$\begin{array}{ccc} \mathcal{A} & \Rightarrow & \mathcal{P} \begin{array}{l} \nearrow \mathcal{L} \\ \searrow \mathcal{S} \end{array} \\ & \leftarrow ? = & \end{array}$$

Here $\mathcal{P} \Rightarrow \mathcal{L}$ since every real eigenvalue of a P -matrix must be positive (e.g. [11]). The matrices of Examples 1 and 3 have real spectra. Thus $\mathcal{S} \nrightarrow \mathcal{L}$ and $\mathcal{L} \nrightarrow \mathcal{S}$.

*The question whether $\mathcal{P} \Rightarrow \mathcal{A}$ for matrices with real eigenvalues is open.*¹

A special case of matrices with real spectra is the class of the totally nonnegative matrices. A matrix is called *totally nonnegative* (*totally positive*) if all its minors are nonnegative (positive) [9]. A totally nonnegative matrix is called an *oscillation matrix* if some power of it is totally positive [9].

It is known that the eigenvalues of a totally nonnegative matrix are nonnegative and those of an oscillation matrix are positive. For totally nonnegative matrices we have the following implications:

$$(2) \quad \begin{array}{c} \mathcal{A} \Rightarrow \mathcal{L} \Leftrightarrow \mathcal{P} \Rightarrow \mathcal{S} \\ \Leftarrow ? = \end{array}$$

A stable totally nonnegative matrix A has positive spectrum since it has no zero eigenvalue. This holds by [8] for every principal submatrix of A . Therefore $\mathcal{L} \Rightarrow \mathcal{P}$. Example 2 demonstrates that $\mathcal{S} \not\Rightarrow \mathcal{P}$.

*The question whether $\mathcal{P} \Rightarrow \mathcal{A}$ for totally nonnegative matrices is open.*¹

Oscillation matrices are P -matrices (e.g., [9]), thus semi-positive and stable, by (2). *The question whether they, or even totally positive matrices, are diagonally stable is open.*¹

A matrix is called an ω -matrix if each of its principal submatrices has at least one real eigenvalue and if $\beta \subseteq \alpha$ implies that $l(A[\alpha]) \leq l(A[\beta])$, where $l(B)$ denotes the minimal real eigenvalue of a matrix B . It is well known (see [6]) that Z -matrices and totally nonnegative matrices are ω -matrices.

The implications diagram for ω -matrices is

$$\begin{array}{c} \mathcal{A} \Rightarrow \mathcal{L} \Rightarrow \mathcal{P} \Rightarrow \mathcal{S} \\ \Leftarrow ? = \quad \Leftarrow ? = \end{array}$$

The implication $\mathcal{L} \Rightarrow \mathcal{P}$ follows from the definition of an ω -matrix, while $\mathcal{S} \not\Rightarrow \mathcal{P}$ by Example 1.

The question whether $\mathcal{P} \Rightarrow \mathcal{L}$, for ω -matrices, is suggested in [6]. *The question whether $\mathcal{L} \Rightarrow \mathcal{A}$ for such matrices is also open.*¹

¹Note added in proof. The question whether $\mathcal{P} \Rightarrow \mathcal{A}$ for oscillation matrices (and thus for totally nonnegative matrices, ω -matrices and for matrices with real spectra) is answered in the negative in [16] using as an example the matrix B given in [17, p. 163].

REFERENCES

[1] G. P. BARKER, A. BERMAN AND R. J. PLEMMONS, *Positive diagonal solutions to the Lyapunov equations*, Lin. Multi. Alg., 5 (1978), pp. 249–256.
 [2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
 [3] A. BERMAN, R. S. VARGA AND R. C. WARD, *ALPP: matrices with nonpositive off-diagonal entries*, Lin. Alg. Appl., 21 (1978), pp. 233–244.
 [4] A. BERMAN AND R. C. WARD, *ALPP: classes of stable and semipositive matrices*, Lin. Alg. Appl., 21 (1978), pp. 163–174.
 [5] G. W. CROSS, *Three types of matrix stability*, Lin. Alg. Appl., 20 (1978), pp. 253–263.
 [6] G. M. ENGEL AND H. SCHNEIDER, *The Hadamard–Fischer inequality for a class of matrices defined by eigenvalue monotonicity*, Lin. Multi. Alg., 4 (1976), pp. 155–176.
 [7] M. FIEDLER AND V. PTÁK, *Some generalizations of positive definiteness and monotonicity*, Numer. Math., 9 (1966), pp. 163–172.
 [8] S. FRIEDLAND, *Weak interlacing properties of totally positive matrices*, unpublished.
 [9] F. R. GANTMACHER AND M. G. KREIN, *Oscillation matrices and kernels and small vibrations of mechanical systems*, Translation Series, U.S. Atomic Energy Commission, 1961.
 [10] B. S. GOH, *Global stability in two species interactions*, J. Math. Biol., 3 (1976), pp. 313–318.
 [11] R. B. KELLOG, *On complex eigenvalues of M and P matrices*, Numer. Math., 19 (1972), pp. 170–175.
 [12] D. J. KLEIN, *Treediagonal matrices and their inverses*, Lin. Alg. Appl., 42 (1982), pp. 109–117.

- [13] A. LYAPUNOV, *Problème général de la stabilité du mouvement*, Ann. Math. Studies, 17, Princeton Univ. Press, Princeton, NJ, 1947.
- [14] R. REDHEFFER AND Z. ZHIMING, *A class of matrices connected with Volterra prey-predator equations*, this Journal, 3 (1982), pp. 122–134.
- [15] J. VANDERGRAFT, *Applications of partial orderings to the study of positive definiteness, monotonicity and convergence*, SIAM J. Numer. Anal., 9 (1972), pp. 97–104.
- [16] S. BJAŁAS AND J. GARLOFF, *Intervals of P-matrices and an extremal property of the determinant*, Freiburger Intervall-Berichte, to appear.
- [17] T. MARKHAM, *A semigroup of totally nonnegative matrices*, Linear Alg. Appl., 3 (1970), pp. 157–164.

APPROXIMATION ALGORITHMS FOR MAXIMIZING THE NUMBER OF SQUARES PACKED INTO A RECTANGLE*

B. S. BAKER[†], A. R. CALDERBANK[†], E. G. COFFMAN, JR.[†] AND J. C. LAGARIAS[†]

Abstract. We consider the NP-hard problem of packing into a specified rectangle a maximum number of squares from a given set. We define two related approximation algorithms and derive bounds on the worst case performance of the packings they produce.

1. Introduction. As described in a recent survey [10], the past decade has seen many significant results in the combinatorial analysis of one- and two-dimensional packing problems. Our interest focuses on the two-dimensional problems, which are the less studied of the two. The basic versions of the two-dimensional problem that bear on this paper involve packing a given collection or list L of squares (S_1, S_2, \dots, S_n) into an enclosing rectangle, R , such that the sides of the enclosed squares are parallel to the sides of R , and no two squares overlap. Three principal variations of this problem are:

(1) For any given number $A > 0$ determine the width and length of R so that it has the least area sufficient to pack all lists of squares whose cumulative area does not exceed A [11]. Special cases of the more general problem in which rotations of squares are allowed have also been studied [8], [9].

(2) The width of R is assumed fixed, and the object is to pack the squares so as to minimize the other dimension. This problem, to be called the *height* problem, has also been extended to the case of lists of rectangles [1], [3], [5], [12].

(3) Assuming R is fixed, pack into R the largest possible number of squares from L . This we will term the *subset* problem.

An interesting, related problem is treated in [4], where an unbounded collection of identical enclosing rectangles is given and the object is to pack a given but arbitrary list of rectangles into as few of these enclosing rectangles as possible. This problem differs from the subset problem in that the efficient packing of rectangles not packed in R does not have to be considered in the subset problem.

Both the height and subset problems have a complexity at least that of their NP-hard, one-dimensional counterparts. For this reason simple but effective approximation algorithms have been proposed and analyzed for the height problem; the remaining sections of this paper will concern the description and analysis of two such algorithms for the subset problem. As in the earlier studies we shall compare the worst case performance of approximate packings relative to optimal packings.

The corresponding one-dimensional bin packing problem is studied in [6] and [7]. Indeed, the algorithms that we analyze can be viewed as two-dimensional analogues of the so-called Next-Fit-Increasing rule analyzed in [6].

From an engineering point of view the applications of two-dimensional packing problems are many and quite varied. To name just a few, they include stock-cutting, VLSI chip design, loading carriers in transportation systems, and multiprogram scheduling in computer operating systems. For further discussion we refer the reader to [7] and [10].

The remainder of the paper is organized as follows. In the next section we define and illustrate two similar algorithms for the subset problem. In § 3 we compare the

* Received by the editors March 5, 1982, and in revised form August 24, 1982.

[†] Bell Laboratories, Murray Hill, New Jersey 07974.

worst case packings of one of these algorithms with an optimal packing. In § 4 the asymptotic results are shown to apply to the other algorithm. In § 5, the worst case performance of the two algorithms is analyzed with respect to the height problem. In the final section some remarks are made concerning generalizations.

2. Approximation algorithms. The study of effective algorithms for the subset problem focuses naturally on those rules which pack subsets of smallest squares; i.e., a square outside the packing is not smaller than a square in the packing. Thus, we consider rules that pack in order of increasing square size.

Two basic classes of algorithms have been studied for the height problem: “bottom-up” [3] and “level-oriented” [5] algorithms. Algorithms in each class pack the squares in sequence as they are drawn from a given list L . When a square is packed by a bottom-up algorithm it is placed, left-justified at the lowest possible level in the current (partial) packing. For our purposes, we define the *Bottom-Up-Increasing*

$$\begin{array}{cccccccc}
 S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & S_8 & \dots \\
 \frac{1}{5} \times \frac{1}{5} & \frac{1}{5} \times \frac{1}{5} & \frac{1}{5} \times \frac{1}{5} & \frac{4}{15} \times \frac{4}{15} & \frac{4}{15} \times \frac{4}{15} & \frac{4}{15} \times \frac{4}{15} & \frac{1}{3} \times \frac{1}{3} & \frac{7}{15} \times \frac{7}{15} & \dots
 \end{array}$$

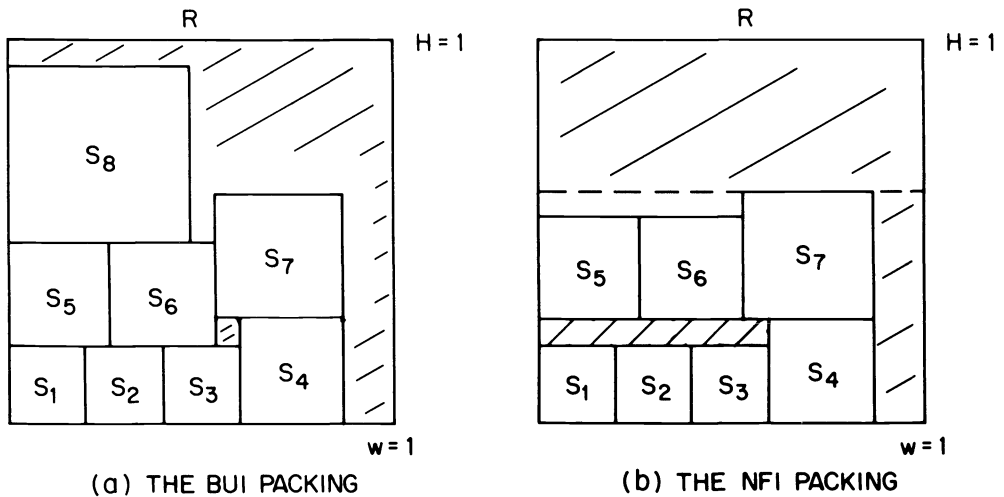


FIG. 1. Example packings.

(BUI) rule as that bottom-up rule which packs the squares in order of increasing size. The BUI rule is applied to the subset problem as follows. The packing process is terminated whenever: i) all squares are packed or, ii) the next square to be packed is larger than any hole in the current packing.

An example is shown in Fig. 1(a). Hereafter, the enclosing rectangle R will be treated as a bin whose width is normalized to 1 and whose height is given by H .

As the name implies, level-oriented packings are arranged in levels, with the bottom of the bin serving as the first level. The Next-Fit rule for the height problem begins, as does the BUI rule, by packing squares left-justified across the first level until a square S is encountered that will not fit in the space remaining at the right end. A horizontal line is then drawn through the top of the largest square packed on this level, thus forming the second level. Beginning with S , packing on level 2 proceeds in an identical fashion, terminating when a square is encountered that will not fit in

the remaining space on this level. Once again, the top of the largest square on level 2 defines level 3. This process continues until all squares are packed.

We shall be concerned with the Next-Fit-Increasing (NFI) rule, which packs the squares in an order of increasing size. Again, the NFI rule is applied to the subset problem by modifying the termination rule. The process terminates whenever: i) all squares are packed or ii) the height of the next square to be considered exceeds the height available below H and above the level on which it would have to be packed.

Fig. 1(b) shows the NFI rule at work on the same list given in Fig. 1(a). The structures of the packings in the figures suggest that the BUI and NFI rules are not substantially different; they both appear to distribute the squares in rows containing the same collections of squares, except possibly for the last row of the BUI packing. Indeed, we shall prove subsequently that the two rules have the same asymptotic performance relative to both the height and subset criteria. Also, for any finite list the performance, in both senses, of the BUI rule is at least as good as that of the NFI rule. Proofs of these results are delayed until after the next section, so that the analysis of the NFI rule can be used to advantage.

3. Performance bounds for the NFI rule. Let $L = (S_1, \dots, S_n)$ be a list of squares to be packed by the NFI rule into a bin of width 1 and height H . Let $0 < s(i) \leq 1$ be the width of S_i , $1 \leq i \leq n$, and for indexing convenience assume $s(1) \leq s(2) \leq \dots \leq s(n)$. For given H and L let $N_{\text{NFI}}(L, H)$ be the number of squares packed by the NFI rule, and let $N_{\text{OPT}}(L, H)$ be the number packed by an optimal rule. L and H may be suppressed from this notation when they are clear from context. Define the asymptotic bound

$$Q_{\text{NFI}} = \lim_{H \rightarrow \infty} \left\{ \max_L \left(\frac{N_{\text{OPT}}(L, H)}{N_{\text{NFI}}(L, H)} \right) \right\},$$

assuming that the limit exists.

THEOREM 1. *We have $Q_{\text{NFI}} = \frac{4}{3}$.*

Remark. If H is small then it is possible to find lists that NFI packs less efficiently than indicated in the above bound. For example, in Fig. 2 we have $N_{\text{OPT}}(L, H) = \frac{7}{5}N_{\text{NFI}}(L, H)$. Indeed, if $H = 2/M$ and $s(1) = \dots = s(M-1) = 1/M - \epsilon$, $s(M) = 1/M$, and $s(M+1) = \dots = s(2M-1) = 1/M + \epsilon$, then $N_{\text{OPT}}(L, H) = (2 - 1/M)N_{\text{NFI}}(L, H)$. We see from Theorem 1 that the effect of this errant behavior does not persist as the size of the problem increases. \square

Proof. We begin with a proof that the asymptotic bound cannot be less than $\frac{4}{3}$.

LEMMA 1. *We have $Q_{\text{NFI}} \geq \frac{4}{3}$.*

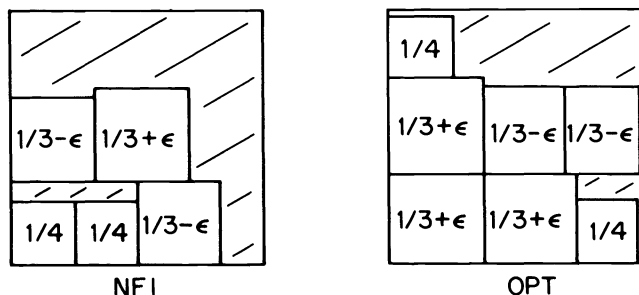


FIG. 2. Example for small H .

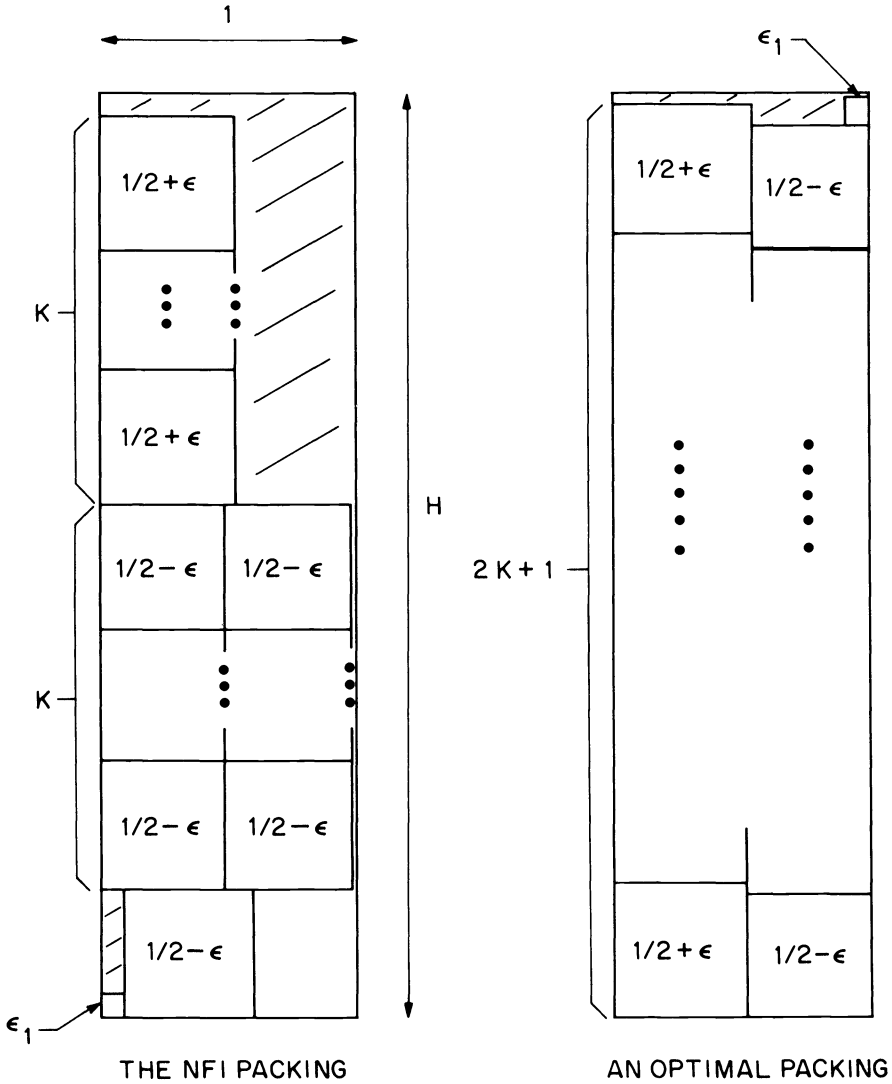


FIG. 3. Worst case example.

Proof. Given any height H we exhibit a list L of squares for which $N_{\text{OPT}}(L, H) \geq \frac{4}{3}N_{\text{NFI}}(L, H)$. If $H = (2K + 1)/2 + M$, where $0 < M < \frac{1}{2}$, and K is an integer, then we choose $0 < \epsilon < M/(2K + 1)$, $\epsilon_1 > 2\epsilon$, and the example given in Fig. 3, which yields

$$\frac{N_{\text{OPT}}(L, H)}{N_{\text{NFI}}(L, H)} = \frac{4K + 3}{3K + 2} \geq \frac{4}{3}.$$

When H is not of the above form there are similar constructions. We leave these as an exercise for the reader. \square

The reverse inequality is proved at the end of a sequence of lemmas. In each lemma the height H is fixed but arbitrary as is the list of squares. We label the levels $x(1), x(2), \dots$ of the NFI packing in order of increasing height. Let $p(i)$ and $q(i)$ respectively denote the side of the largest and smallest squares in level $x(i)$. Let $U(i)$

be the unused area at the right-hand side of level $x(i)$ and let U be the total unused area in the NFI packing. Let L_N be the number of levels containing exactly N squares and let L_N^* be the number of levels containing at least N squares. Let ρ be the side of the smallest square that the NFI algorithm fails to pack and let $\Delta = N_{OPT} - N_{NFI}$.

LEMMA 2. Let $N = \lfloor 1/\rho \rfloor$. Then

$$\frac{N_{OPT}}{N_{NFI}} \leq 1 + \frac{2}{K} + \frac{1}{N},$$

where K is the number of levels in the NFI packing. If $H \geq 6$ and $N \geq 4$ then $K \geq 24$ and $N_{OPT} \leq \frac{4}{3}N_{NFI}$.

Proof. The unused area at the top of the NFI packing is $H - H_1 \leq \rho$, where H_1 is the height of the top of the largest square in the last level. The contribution from level $x(i)$ to the unused area in the region $x \leq 1 - \rho$, $y \leq H_1$ is bounded above by $(1 - \rho)(p(i) - q(i))$. (See Fig. 4.) Since $U(i) \leq \rho^2$ we have

$$U \leq \left(\sum_{i=1}^K (p(i) - q(i)) \right) (1 - \rho) + K\rho^2 + \rho.$$

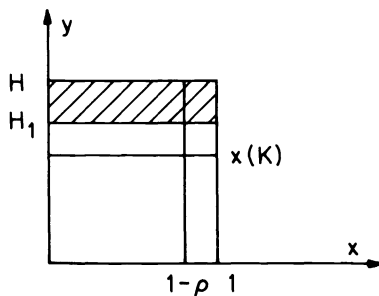


FIG. 4. Illustrating wasted area.

Since $q(i + 1) \geq p(i)$ we have $U \leq (1 - \rho)(p(K) - q(1)) + K\rho^2 + \rho$. Now $U \geq \Delta\rho^2$, so

$$\Delta < \frac{2}{\rho} + (K - 1).$$

Since $\rho > 1/(N + 1)$ we have $\Delta \leq 2N + K$. Using $N_{NFI} \geq NK$, the result follows. \square

The remainder of the proof considers the cases for $\lfloor 1/\rho \rfloor = 1, 2,$ and 3 . For each case, we will need several bounds on the size of Δ . Ultimately, we will show that in each case, the minimum of the bounds on Δ is always small enough to give the desired result.

In order to obtain the first bounds on Δ , we need to bound the area of the unused space at the right sides of the levels.

LEMMA 3. We have

$$(1) \quad \sum_{i=1}^{L_4^*} U(i) \leq \frac{1}{25}(L_4^* - 1) + 3\rho^2$$

and

$$(2) \quad \sum_{i=L_4^*+1}^{L_3^*} U(i) \leq \frac{1}{16}(L_3^* - 1) + 2\rho^2.$$

Proof. If $q(L_4^*) \leq \frac{1}{5}$ then $U(i) \leq \frac{1}{25}$ for $i < L_4^*$ and we have

$$\sum_{i=1}^{L_4^*} U(i) \leq U(L_4^*) + \frac{1}{25}(L_4^* - 1) \leq \rho^2 + \frac{1}{25}(L_4^* - 1).$$

If $q(L_4^*) > \frac{1}{5}$, then let b be the greatest index for which $q(b) \leq \frac{1}{5}$. If $b < i \leq L_4^*$ then $x(i)$ contains exactly 4 squares and $U(i) \leq p(i)(1 - 3q(i) - p(i))$. If $i < b$ then $U(i) \leq p(i)q(i+1)$ and so $U(i) \leq q(b)^2 \leq \frac{1}{25}$. It follows that

$$\sum_{i=1}^{L_4^*} U(i) \leq U(b) + \frac{1}{25}(b-1) + \sum_{i=b+1}^{L_4^*} (p(i) - 4p(i)^2) + 3 \sum_{i=b+1}^{L_4^*} p(i)(p(i) - q(i)).$$

If $i > b$ then $p(i) = \frac{1}{5} + \epsilon_i$, where $\epsilon_i > 0$, and we have $p(i) - 4p(i)^2 \leq \frac{1}{25}$. Hence

$$\begin{aligned} \sum_{i=1}^{L_4^*} U(i) &\leq p(b)q(b+1) + \frac{1}{25}(L_4^* - b + (b-1)) + 3p(L_4^*)(p(L_4^*) - q(b+1)) \\ &\leq \frac{1}{25}(L_4^* - 1) + 3\rho^2. \end{aligned}$$

The proof of (2) is similar and we omit the details. \square

LEMMA 4. If $\frac{1}{3} < \rho \leq \frac{1}{2}$ then

$$(3) \quad \Delta \leq 11 + L_2 + \frac{9}{16}(L_3 - 1) + \frac{9}{25}(L_4^* - 1).$$

If $\frac{1}{4} < \rho \leq \frac{1}{3}$ then

$$(4) \quad \Delta \leq 11 + L_3 + \frac{16}{25}(L_4^* - 1).$$

Proof. If $i > L_3^*$ then $x(i)$ contains exactly 2 squares and $U(i) \leq \rho^2$. By the proof of Lemma 2 the unused area above and in the interior of the packing is bounded above by 2ρ . Combining these facts with inequalities (1) and (2), we have

$$U \leq L_2\rho^2 + \frac{1}{25}(L_4^* - 1) + 3\rho^2 + \frac{1}{16}(L_3 - 1) + 2\rho^2 + 2\rho.$$

Since $\Delta\rho^2 \leq U$ and since $\rho > \frac{1}{3}$ we have

$$\Delta \leq 11 + L_2 + \frac{9}{16}(L_3 - 1) + \frac{9}{25}(L_4^* - 1).$$

The proof of (4) is similar and we omit the details. \square

Lemmas 2, 3, and 4 rest only on area arguments. For the height arguments of subsequent lemmas, we need a little more notation. Let H_N be the height of those levels containing exactly N squares and let H_N^* be the height of those levels containing at least N squares.

LEMMA 5. $H_{N+1}^* \leq p(L_{N+1}^*) + 1/(N+1)(L_{N+1}^* - 1)$.

Proof. The level $x(L_{N+1}^*)$ contains at least $(N+1)$ squares and so $q(L_{N+1}^*) \leq 1/(N+1)$. If $i < L_{N+1}^*$ then $p(i) \leq q(L_{N+1}^*) \leq 1/(N+1)$. The result follows. \square

The levels $x(i)$, $L_{N+1}^* + 1 \leq i \leq L_N^*$, all contain exactly N squares. Let $T(i)$ be the sum of the sides of the squares in level $x(i)$ and let $T_N = T(L_{N+1}^* + 1) + \dots + T(L_N^*)$.

LEMMA 6. $T_N - T(L_{N+1}^* + 1) \geq NH_N - (N-1)p(L_N^*) - p(L_{N+1}^* + 1)$.

Proof. Since $T(i) \geq p(i) + (N-1)p(i-1)$, for $L_{N+1}^* + 2 \leq i \leq L_N^*$, we have

$$\begin{aligned} T_N - T(L_{N+1}^* + 1) &\geq \sum_{i=L_{N+1}^*+2}^{L_N^*} (p(i) + (N-1)p(i-1)) \\ &\geq N \left(\sum_{i=L_{N+1}^*+1}^{L_N^*} p(i) \right) - (N-1)p(L_N^*) - p(L_{N+1}^* + 1). \end{aligned}$$

Since $H_N = \sum_{i=L_{N+1}^*+1}^{L_N^*} p(i)$, the result follows. \square

LEMMA 7. If $\frac{1}{3} < \rho \leq \frac{1}{2}$ then

$$(5) \quad \Delta \leq 10 + L_2 + \frac{9}{4}(L_4^* - 1)$$

and

$$(6) \quad \Delta \leq 8 + 2(L_3 - 1) + \frac{3}{2}(L_4^* - 1).$$

Proof. Consider the 3 vertical lines $x = \frac{1}{4}$, $x = \frac{1}{2}$, $x = \frac{3}{4}$, drawn parallel to the sides of the bin. The combined length of the lines is $3H$ and $H \leq \rho + H_2 + H_3 + H_4^*$. Every square of side greater than $\frac{1}{4}$ covers at least one line. If $i \geq L_4^* + 2$ then every square in $x(i)$ has side greater than $\frac{1}{4}$. By Lemma 6 we have

$$(7) \quad \begin{aligned} T_2 &\geq 2H_2 - p(L_2^*) - p(L_3^* + 1) + T(L_3^* + 1) \\ &\geq 2H_2 - p(L_2^*), \end{aligned}$$

and

$$T_3 - T(L_4^* + 1) \geq 3H_3 - 2p(L_3^*) - p(L_4^* + 1).$$

If Z is the length of the 3 lines not covered by squares in the NFI packing then

$$Z \leq 3(\rho + H_2 + H_3 + H_4^*) - T_2 - (T_3 - T(L_4^* + 1)) \leq 7\rho + H_2 + 3H_4^*.$$

Since $Z \geq \Delta\rho$, it follows from Lemma 5 that

$$\Delta\rho \leq 7\rho + L_2\rho + 3(p(L_4^*) + \frac{1}{4}(L_4^* - 1)).$$

Since $\rho > \frac{1}{3}$ we have

$$\Delta \leq 10 + L_2 + \frac{9}{4}(L_4^* - 1).$$

We prove (6) by considering the two vertical lines $x = \frac{1}{3}$, $x = \frac{2}{3}$ drawn parallel to the sides of the bin. If $i \geq L_3^* + 2$ then every square in $x(i)$ has side greater than $\frac{1}{3}$ and covers at least one line. If Z' is the length of the 2 lines not covered by squares in the NFI packing then

$$Z' \leq 2(\rho + H_2 + H_3 + H_4^*) - (T_2 - T(L_3^* + 1)).$$

By Lemma 5 and by (7) above we have

$$Z' \leq 4\rho + 2(p(L_3^*) + \frac{1}{3}(L_3 - 1)) + 2(p(L_4^*) + \frac{1}{4}(L_4^* - 1)).$$

Since $\rho > \frac{1}{3}$ and since $Z' \geq \Delta\rho$ we have

$$\Delta \leq 8 + 2(L_3 - 1) + \frac{3}{2}(L_4^* - 1). \quad \square$$

LEMMA 8. If $\frac{1}{4} < \rho \leq \frac{1}{3}$ then

$$(8) \quad \Delta \leq 9 + 3(L_4^* - 1).$$

If $\rho > \frac{1}{2}$ then

$$(9) \quad \Delta \leq 3 + L_2 + \frac{2}{3}(L_3^* - 1),$$

and

$$(10) \quad \Delta \leq 6 + L_1 + \frac{4}{3}(L_3^* - 1).$$

Proof. The proofs of (8), (9), and (10) are similar to those given in Lemma 7 and we shall not give the details. Equation (8) is proved by considering the lines $x = \frac{1}{4}$, $x = \frac{1}{2}$, and $x = \frac{3}{4}$. Equation (10) is proved by considering the lines $x = \frac{1}{3}$ and $x = \frac{2}{3}$. In

the proof of (9) we consider the two vertical lines $x = p(L_2^* + 1)$, and $x = 1 - p(L_2^* + 1)$, and we observe that if $i > L_2^* + 1$ then the square in level $x(i)$ covers both vertical lines. \square

LEMMA 9. *If $\frac{1}{4} < \rho \leq \frac{1}{3}$ then*

$$(11) \quad N_{\text{OPT}} < \frac{14}{11}N_{\text{NFI}} + 11.$$

Proof. By (4) and (8) we have

$$\frac{N_{\text{OPT}}}{N_{\text{NFI}}} \leq \frac{N_{\text{NFI}}}{N_{\text{NFI}}} + \min\left(\frac{3L_4^*}{N_{\text{NFI}}}, \frac{L_3 + 16L_4^*/25}{N_{\text{NFI}}}\right) + \frac{11}{N_{\text{NFI}}}.$$

If $L_4^* = 0$ then we are done. Suppose $L_4^* \neq 0$ and set $L_3 = \alpha L_4^*$. Then since $N_{\text{NFI}} \geq 3L_3 + 4L_4^*$ we have

$$\frac{N_{\text{OPT}}}{N_{\text{NFI}}} \leq 1 + \min(f_1(\alpha), f_2(\alpha)) + \frac{11}{N_{\text{NFI}}},$$

where $f_1(\alpha) = 3/(3\alpha + 4)$ and $f_2(\alpha) = (\alpha + \frac{16}{25})/(3\alpha + 4)$. We observe that f_1 is a decreasing function of α and f_2 is an increasing function of α . Setting $f_1(\alpha) = f_2(\alpha)$ gives $\alpha = \frac{59}{25}$. Since $f_1(\frac{59}{25}) = \frac{75}{277} < \frac{3}{11}$ we have

$$N_{\text{OPT}} < \frac{14}{11}N_{\text{NFI}} + 11. \quad \square$$

LEMMA 10. *If $\rho > \frac{1}{2}$ then*

$$(12) \quad N_{\text{OPT}} \leq \frac{4}{3}N_{\text{NFI}} + 5.$$

Proof. By (9) and (10) we have

$$\frac{N_{\text{OPT}}}{N_{\text{NFI}}} \leq \frac{N_{\text{NFI}}}{N_{\text{NFI}}} + \frac{5}{N_{\text{NFI}}} + \min\left(\frac{L_2 + 2L_3^*/3}{N_{\text{NFI}}}, \frac{L_1 + 4L_3^*/3}{N_{\text{NFI}}}\right).$$

Note that $N_{\text{NFI}} \geq L_1 + 2L_2 + 3L_3^*$. If $L_2 = 0$ then $N_{\text{OPT}} \leq (\frac{11}{9})N_{\text{NFI}} + 5$ and we are done. Suppose $L_2 \neq 0$, let $L_1 = \alpha L_2$ and let $L_3^* = \beta L_2$. Then

$$\frac{N_{\text{OPT}}}{N_{\text{NFI}}} \leq 1 + \frac{5}{N_{\text{NFI}}} + \min(h_1(\alpha, \beta), h_2(\alpha, \beta))$$

where

$$h_1(\alpha, \beta) = \frac{1 + 2\beta/3}{\alpha + 2 + 3\beta} \quad \text{and} \quad h_2(\alpha, \beta) = \frac{\alpha + 4\beta/3}{\alpha + 2 + 3\beta}.$$

Now h_1 is a decreasing function of α and h_2 is an increasing function of α . Setting $h_1(\alpha, \beta) = h_2(\alpha, \beta)$ we obtain $\alpha = 1 - 2\beta/3$, and so

$$\min(h_1(\alpha, \beta), h_2(\alpha, \beta)) \leq \frac{1 + 2\beta/3}{3 + 7\beta/3} \leq \frac{1}{3}.$$

The result follows. \square

LEMMA 11. *If $\frac{1}{3} < \rho \leq \frac{1}{2}$ then*

$$(13) \quad N_{\text{OPT}} < (1 + \frac{74}{234})N_{\text{NFI}} + 11.$$

Proof. By (3), (5) and (6) we have

$$\frac{N_{\text{OPT}}}{N_{\text{NFI}}} \leq \frac{N_{\text{NFI}}}{N_{\text{NFI}}} + \frac{11}{N_{\text{NFI}}} + \min\left(\frac{L_2 + 9L_3/16 + 9L_4^*/25}{N_{\text{NFI}}}, \frac{2L_3 + 3L_4^*/2}{N_{\text{NFI}}}, \frac{L_2 + 9L_4^*/4}{N_{\text{NFI}}}\right).$$

If $L_3 = 0$ then

$$\frac{N_{OPT}}{N_{NFI}} \leq 1 + \frac{11}{N_{NFI}} + \min\left(\frac{L_2 + 9L_4^*/25}{N_{NFI}}, \frac{3L_4^*/2}{N_{NFI}}\right),$$

and $N_{NFI} \geq 2L_2 + 4L_4^*$. If $L_4^* = 0$ then we are done. If $L_4^* \neq 0$ then set $L_2 = \alpha L_4^*$. By an argument similar to Lemma 9 we have

$$\frac{N_{OPT}}{N_{NFI}} \leq 1 + \frac{11}{N_{NFI}} + \min\left(\frac{\alpha + 9/25}{2\alpha + 4}, \frac{3/2}{2\alpha + 4}\right) \leq \frac{5}{4} + \frac{11}{N_{NFI}}.$$

We may suppose that $L_3 \neq 0$. Setting $L_2 = \alpha L_3$ and $L_4^* = \beta L_3$ we have

$$\frac{N_{OPT}}{N_{NFI}} \leq 1 + \frac{11}{N_{NFI}} + \min(g_1(\alpha, \beta), g_2(\alpha, \beta), g_3(\alpha, \beta)),$$

where

$$g_1(\alpha, \beta) = \frac{\alpha + 9/16 + 9\beta/25}{2\alpha + 3 + 4\beta}, \quad g_2(\alpha, \beta) = \frac{2 + 3\beta/2}{2\alpha + 3 + 4\beta}, \quad g_3(\alpha, \beta) = \frac{\alpha + 9\beta/4}{2\alpha + 3 + 4\beta}.$$

Now g_1 is an increasing function of α and g_2 is a decreasing function of α . Setting $g_1(\alpha, \beta) = g_2(\alpha, \beta)$ gives $\alpha = \frac{23}{16} + \frac{57}{50}\beta$. Thus

$$\min(g_1(\alpha, \beta), g_2(\alpha, \beta)) \leq m_1(\beta) = \frac{2 + 3\beta/2}{94/16 + 314\beta/50}.$$

If $\beta \geq 6$ then $m_1(\beta) < \frac{11}{43}$. If $\beta < 6$ then $g_3(\alpha, \beta)$ is an increasing function of α . Setting $g_2(\alpha, \beta) = g_3(\alpha, \beta)$ gives $\alpha = 2 - 3\beta/4$. Thus if $\beta < 6$ then

$$\min(g_2(\alpha, \beta), g_3(\alpha, \beta)) \leq m_2(\beta) = \frac{2 + 3\beta/2}{7 + 5\beta/2}.$$

Now m_2 is an increasing function of β ; m_1 is a decreasing function of β ; and equality occurs at $\beta = 900/(16 \times 189)$. Since $m_2(900/(16 \times 189)) = 7398/23418 < 74/234$ the result follows. \square

Theorem 3.1 now follows directly from Lemma 1 and Lemmas 9, 10 and 11. \square

4. Comparison of the NFI and BUI rules. In this section we shall verify that asymptotic results for the NFI rule in both the height and subset problems carry over to the BUI rule. It is convenient to consider the height problem first. Consistent with the literature we shall use $NFI(L)$ and $OPT(L)$ as the respective heights of the NFI and an optimum packing of list L .

THEOREM 2. For any list $L = (S_1, \dots, S_n)$, with $s(i) \leq s(i + 1)$, $1 \leq i < n$,

$$(14) \quad BUI(L) \leq NFI(L) \leq BUI(L) + s(n).$$

Proof. A subsequence S_j, \dots, S_k in L constitutes a row in the BUI packing of L if:

1. S_j rests against the left side of the bin; for each $j \leq i < k$, the left side of S_i touches the right side of S_{i-1} , but if $k < n$, the left side of S_{k+1} does not touch the right side of S_k on the right;
2. The top heights (levels of the top edges) of S_j, \dots, S_k form a nondecreasing sequence such that the top heights of S_j and S_k differ by less than $s(k)$.

The following preliminary result establishes the close correspondence between BUI and NFI packings.

LEMMA 12. *Let $x(1), \dots, x(K)$ be the sets of squares in the $K \geq 1$ levels of an NFI packing of list L . The BUI packing of L consists of K rows, where row i , $1 \leq i \leq K$, consists of just those squares in $x(i)$, all squares in row 1 rest on the bottom of the bin, and each square in row i , $i > 1$, touches a square below it in row $i - 1$.*

Proof (by induction). The result follows easily for $x(1)$, since this set of squares must also rest on the bottom of the bin in the BUI packing, in a left-to-right increasing order by size.

Thus, suppose the result holds for $x(1), \dots, x(i - 1)$ and consider the first square S_j in $x(i)$. Since $x(i - 1)$ is a row, S_j cannot fit to the right of S_{j-1} . S_j cannot fit in any hole below any square in row $i - 1$, because it is at least as large as S_{j-1} which the BUI rule could not pack into any such hole. Thus, S_j must be packed above and/or to the left of S_{j-1} . Since the top heights of squares in row $i - 1$ are nondecreasing, the BUI rule must pack S_j against the left side of the bin so that it touches one or more squares in row $i - 1$.

Since the total variation in the nondecreasing top heights in row $i - 1$ is less than $p(i - 1) = s(j - 1)$, which is no greater than $s(j) = q(i)$, the top height of S_j must exceed the greatest top height in row $i - 1$. It follows readily that squares S_j, \dots, S_k must be packed contiguously, left-to-right, and touching squares in row $i - 1$ until a square S_{k+1} is encountered not fitting in the space remaining at the right of this bin, or until there are no squares left to pack. From the NFI packing we know that if $k < n$ the cumulative width of squares in $x(i)$ is in $(1 - s(k + 1), 1]$. Thus, $x(i) = \{S_j, \dots, S_k\}$ and property 1 of a row is established.

Next, since the top heights in row $i - 1$ are nondecreasing, and the squares in $x(i)$ are packed in an order of increasing size, the top heights of the squares S_j, \dots, S_k in the BUI packing are also nondecreasing. Finally, since the top height of S_j exceeds the greatest top height in row $i - 1$, it must exceed the bottom height of S_k . Thus, the top heights of S_j and S_k differ by less than $s(k)$ and property 2 of a row is established for $\{S_j, \dots, S_k\} = x(i)$. \square

From the monotonicity property in Lemma 12, it is clear that the height of the last square S_n defines $\text{BUI}(L)$, and that the bottom of S_n must be at a height at least that of the top of the left most square in $x(K - 1)$. Thus, using $p(K) = s(n)$, Lemma 12 implies

$$(15) \quad \text{BUI}(L) \geq \sum_{i=1}^{K-1} q(i) + p(K).$$

Since $q(i) \geq p(i - 1)$, $i > 1$, we can write

$$(16) \quad \text{BUI}(L) \geq q(1) + \sum_{i=1}^{K-2} p(i) + p(K).$$

By definition of the NFI rule $\text{NFI}(L) = \sum_{i=1}^K p(i)$. Substituting from (16), we have $\text{NFI}(L) \leq \text{BUI}(L) - q(1) + p(K - 1) \leq \text{BUI}(L) + s(n)$, thus proving the upper bound in (14). The lower bound in (14) can be obtained readily from Lemma 12. \square

We turn now to a comparison of BUI and NFI for the subset problem.

THEOREM 3. *Let the NFI packing of a sublist of L in a bin of height H have $K \geq 1$ levels. Then*

$$(17) \quad 1 \leq \frac{N_{\text{BUI}}(L, H)}{N_{\text{NFI}}(L, H)} \leq \frac{K + 1}{K}.$$

Proof. From Lemma 12 it is readily verified that $N_{\text{BUI}}(L, H) \geq N_{\text{NFI}}(L, H)$. Let x^* denote the set of squares in L packed by the BUI rule but not by the NFI rule. We analyze two cases.

Case 1. $x(K)$ is a row in the BUI packing. In this case all squares in x^* have bottom edges at a height at least that of the top of the first square in row K , and this height in turn is at least $\sum_{i=1}^K q(i)$. We now verify that $H - \sum_{i=1}^K q(i) < 2\rho$, where ρ is the width of the smallest square not packed by the NFI rule; i.e. a smallest square in x^* .

First, using $q(i) \geq p(i-1)$, $i > 1$, we have $\text{NFI}(L) = \sum_{i=1}^K p(i) \leq p(K) + \sum_{i=1}^K q(i) - q(1)$, and hence $\text{NFI}(L) - \sum_{i=1}^K q(i) < \rho$. Next, by definition, $H - \text{NFI}(L) < \rho$, so that

$$H - \sum_{i=1}^K q(i) = H - \text{NFI}(L) + \left(\text{NFI}(L) - \sum_{i=1}^K q(i) \right) < 2\rho.$$

It follows immediately that the BUI rule packs at most one extra row of squares. Therefore,

$$(18) \quad \frac{\text{BUI}(L, H)}{\text{NFI}(L, H)} \leq \frac{\text{NFI}(L, H) + r}{\text{NFI}(L, H)},$$

where r is the number of squares in x^* . Since the number of squares per level is nonincreasing, we have $\text{NFI}(L, H) \geq rK$. Maximization of (18) thus yields (17).

Case 2. $x(K)$ is not a row in the BUI packing. In this case the first square in x^* must be narrow enough to fit on level K in the NFI packing, but it is too tall. Thus, $H - \sum_{i=1}^{K-1} p(i) < \rho$. Again using $q(i) \geq p(i-1)$, $i > 1$, we have $H - \sum_{i=2}^K q(i) < \rho$ and hence $H - \sum_{i=1}^{K-1} q(i) < \rho + q(K) < 2\rho$. Clearly, from Lemma 12 all squares in x^* must have bottom heights at least $\sum_{i=1}^{K-1} q(k)$. Thus, x^* consists of at most one extra row of squares and (17) follows as before. \square

5. The height problem. We obtain a tight asymptotic bound for the worst case performance of the NFI rule for the height problem. The bound will be calculated from the following infinite series. For any positive integer r , let

$$t_1(r) = r + 1, \quad t_2(r) = r + 2, \\ t_{i+1}(r) = t_i(r)[t_i(r) - 1] + 1, \quad i \geq 2.$$

For example, the first two sequences begin with

$$2, 3, 7, 43, 1807 \quad \text{and} \quad 3, 4, 13, 157, 24493.$$

Let

$$\gamma_r = \sum_{i=1}^{\infty} \frac{1}{t_i(r) - 1} \quad \text{and} \quad \gamma_r^* = \frac{r-1}{r} + \gamma_r.$$

The first few values of γ_r^* are approximately $\gamma_1^* = 1.691 \dots$, $\gamma_2^* = 1.423 \dots$, $\gamma_3^* = 1.302 \dots$.

THEOREM 4. For any list $L = (S_1, \dots, S_n)$, with $s(i) \leq s(i+1)$ ($1 \leq i < n$), if $r = \lfloor 1/s(n) \rfloor$, then

$$\text{BUI}(L) \leq \text{NFI}(L) \leq \gamma_r^* \text{OPT}(L) + 8.4s(n).$$

Moreover, the multiplicative constant γ_r^* is the smallest possible.

Proof. From Theorem 2, $\text{BUI}(L) \leq \text{NFI}(L)$. The bound on $\text{NFI}(L)$ follows from a modification of the analysis of the Next-Fit-Decreasing algorithm in [2]. Define a

weighting function W_r as follows. For $x \in (1/(k + 1), 1/k)$, $k \geq r$,

$$W_r(x) = \begin{cases} \frac{1}{k} & \text{if } k = t_i(r) - 1 \text{ for some } i \geq 1, \\ \frac{k+1}{k}x & \text{otherwise.} \end{cases}$$

We claim that

$$\gamma_r^* \text{OPT}(L) \geq \sum_{i=1}^n s(i)W_r(s(i)) \geq \text{NFI}(L) - 8.4s(n).$$

The first inequality follows from the proof in [2] that for any set S of real numbers in the interval $(0, 1/r)$ summing to at most 1, $\sum_{x \in S} W(x) \leq \gamma_r^*$. Divide an optimal packing of L into horizontal strips by drawing a horizontal line through the top and bottom of each square. Within a strip of height h , the sum of $hs(i)$ over all $s(i)$ intersecting the strip is at most $h\gamma_r^*$, and summing over all such strips we have

$$\sum_{i=1}^n s(i)W_r(s(i)) \leq \gamma_r^* \text{OPT}(L).$$

To prove the second inequality, we begin with some notation.

If $k = t_j(r) - 1$ for some $j \geq 1$, we say that $(1/(k + 1), 1/k)$ is a γ_r -interval, and a square whose size is in such an interval is a γ_r -square.

If there are $K \geq 1$ levels in the NFI packing of L , let $x(i)$ denote the set of squares packed in level i , $1 \leq i \leq K$. Define $p(i)$ to be the size of the largest square in $x(i)$, $1 \leq i \leq K$, and define $p(0) = 0$. For $1 \leq i \leq K$, define $A_r(i) = p(i) \sum_{S(j) \in x(i)} W_r(s(j))$. Note that for $1 \leq i \leq K$,

$$\begin{aligned} A_r(i) &\leq \sum_{S(j) \in x(i)} s(j)W_r(s(j)) + [p(i) - p(i - 1)] \max_{1 \leq j \leq n} \frac{W_r(s(j))}{s(j)} \\ &\leq \sum_{S(j) \in x(i)} s(j)W_r(s(j)) + \frac{r+1}{r} [p(i) - p(i - 1)] \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^K A_r(i) &\leq \sum_{j=1}^n s(j)W_r(s(j)) + \frac{r+1}{r} \sum_{i=1}^K [p(i) - p(i - 1)] \\ &\leq \sum_{j=1}^n s(j)W_r(s(j)) + \frac{r+1}{r} s(n). \end{aligned}$$

Thus, it will be sufficient to show that

$$\sum_{i=1}^K A_r(i) \leq \text{NFI}(L) - 7.4s(n) + \frac{s(n)}{r}.$$

Equivalently, if we define the shortfall in level i to be $p(i) - A_r(i)$, then it will be sufficient to show that the total shortfall over all levels is at most

$$7.4s(n) - \frac{s(n)}{r}.$$

For $1 \leq i \leq K$, let $m(i)$ be an integer such that the size of the smallest square in $x(i)$ lies in the interval $(1/(m(i)+1), 1/m(i))$.

In order to bound the total shortfall, we partition the levels other than level K into three groups. Level i is in Group 1 if $m(i) = m(i+1)$, Group 2 if $m(i) > m(i+1)$ and $x(i)$ contains at least one γ_r -square, and Group 3 if $m(i) > m(i+1)$ and $x(i)$ contains no γ_r -squares.

Note that if level i is in Group 1, it contains $m(i)$ squares in the interval $(1/(m(i)+1), 1/m(i))$ and

$$A_r(i) \geq m_i \left(\frac{1}{m_i} \right) p(i) = p(i).$$

Thus, the total shortfall of levels in Group 1 is 0.

Suppose level i is in Group 2. Since the smallest square in $x(i+1)$ did not fit in level i ,

$$\sum_{s(j) \in x(i)} s(j) > 1 - \frac{1}{m(i+1)}$$

and

$$A_r(i) \geq p(i) \left[1 - \frac{1}{m(i+1)} \right] \min_{1 \leq j \leq n} \frac{W_r(s(j))}{s(j)} \geq p(i) \left[1 - \frac{1}{m(i+1)} \right].$$

If the smallest γ_r -square (if any) packed in a Group 2 level after level i is in the interval $(1/t_i(r), 1/(t_i(r)-1))$, then the shortfall for level i is at most $p(i)/m(i+1) \leq p(i)/(t_i(r)-1)$. Since at most two levels in Group 2 can contain γ_r -squares in the same γ_r -interval, and only the last Group 2 level has no γ_r -square packed in a later Group 2 level, the cumulative shortfall for all Group 2 levels is at most

$$s(n) + 2s(n) \sum_{i=1}^{\infty} \frac{1}{t_i(r)-1} \leq s(n)[1 + 2\gamma_r] < 4.4s(n).$$

Suppose level i is in Group 3. Since $x(i)$ contains no γ_r -squares,

$$\frac{W_r(s(j))}{s(j)} \geq \frac{m(i)+1}{m(i)} \quad \text{for } s(j) \in x(i).$$

Also, as before,

$$\sum_{s(j) \in x(i)} s(j) \geq 1 - \frac{1}{m(i+1)}.$$

Thus,

$$A_r(i) \geq \frac{m(i)+1}{m(i)} \left[1 - \frac{1}{m(i+1)} \right] p(i) \geq p(i) - \frac{m(i)-m(i+1)+1}{m(i)m(i+1)} p(i).$$

Thus, the shortfall for level i is at most

$$\frac{m(i)-m(i+1)+1}{m(i)m(i+1)} s(n)$$

and the cumulative shortfall for all levels in Group 3 is at most

$$s(n) \sum_{i=1}^{K-1} \frac{m(i) - m(i+1) + 1}{m(i)m(i+1)} \leq s(n) \left[\frac{m(1) - m(K-1)}{m(1)m(K-1)} + \sum_{i=1}^{\infty} \frac{1}{i^2} \right] \\ \leq s(n) \left[1 + \frac{\pi^2}{6} \right] < 3s(n).$$

Combining the shortfall for Groups 1–3 and the shortfall of at most $p(K) - s(n)W_r(s(n)) = s(n) - s(n)/r$ for level K , we have a total shortfall over all levels of at most $8.4s(n) - s(n)/r$ as desired.

To show tightness, consider the following packing of a list L . Given r, k , and suitably small ϵ , let N be divisible by each $t_i(r) - 1, 1 \leq i \leq k$, and by r . Pack $r - 1$ columns of $N(r + 1)$ squares of size $1/(r + 1) + \epsilon$ for a total height of $N + N(r + 1)\epsilon$. Next to these columns, pack one column each of $Nt_i(r)$ squares of size $1/t_i(r) + \epsilon$, for a total height of $N + Nt_i(r)\epsilon, 1 \leq i \leq k$. Thus, for this list $L, \text{OPT}(L) \leq N + O(\epsilon)$. The NFI rule, on the other hand, packs $Nt_i(r)/(t_i(r) - 1)$ levels of squares of size $1/t_i(r) + \epsilon$ for a total height of at least $N/(t_i(r) - 1)$ for $1 \leq i \leq k$. Then it packs $N(r + 1)(r - 1)/r$ levels of squares of size $1/(r + 1) + \epsilon$. Thus,

$$\text{NFI}(L) \geq N \sum_{i=1}^k \frac{1}{t_i(r) - 1} + \frac{N(r - 1)}{r}.$$

By appropriate choice of N, k , and ϵ , we can make $\text{NFI}(L)/\text{OPT}(L)$ as close as desired to γ_r^* . Since $\text{BUI}(L) \geq \text{NFI}(L) - s(n), \gamma_r^*$ is also a tight asymptotic bound for BUI. \square

6. Concluding remarks. Our approach to square-packing algorithms has been to devise and analyze easily implemented approximation algorithms having very simple structures. At some sacrifice in simplicity other algorithms can be designed which can be expected to produce better packings. For example, a first-fit-decreasing rule could be applied iteratively, as described in [6] for the one-dimensional case. Unfortunately, because of the added complication, the prospects of tight asymptotic bounds appear to be considerably worse.

It has been noted that the guaranteed efficiency of the NFI algorithm improves as the height increases relative to the width. Therefore, a natural algorithm, NFI* say, for packing squares into an arbitrary rectangle would be the following:

- (1) Rotate the rectangle until the larger dimension is vertical.
- (2) Pack the squares using the NFI algorithm.

We conjecture that a worst case example for NFI* is given by the squares-into-a-square problem of Fig. 2, where the ratio $N_{\text{NFI}^*}/N_{\text{OPT}}$ is given by $\frac{5}{7}$.

Finally, packing rectangles into a rectangle is a natural generalization of our problem deserving further study. Unfortunately, effective algorithms with a simplicity comparable to NFI and BUI do not appear possible. Specifically, level-oriented or bottom-up algorithms with lists ordered by either dimension do not have finite worst case bounds.

REFERENCES

[1] B. S. BAKER, D. J. BROWN AND H. P. KATSEFF, *A 5/4 algorithm for two-dimensional packing*, J. Algorithms, 2 (1981), pp. 348–368.
 [2] B. S. BAKER AND E. G. COFFMAN, JR., *A tight asymptotic bound for next-fit decreasing bin-packing*, this Journal, 2 (1981), pp. 147–152.

- [3] B. S. BAKER, E. G. COFFMAN, JR. AND R. L. RIVEST, *Orthogonal packings in two dimensions*, SIAM J. Comput., 9 (1980), pp. 845–855.
- [4] F. R. K. CHUNG, M. R. GAREY AND D. S. JOHNSON, *On packing two-dimensional bins*, this Journal, 3 (1982), pp. 66–76.
- [5] E. G. COFFMAN, JR., M. R. GAREY, D. S. JOHNSON AND R. E. TARJAN, *Performance bounds for level-oriented two-dimensional packing algorithms*, SIAM J. Comput., 9 (1980), pp. 808–826.
- [6] E. G. COFFMAN, JR. AND J. Y. LEUNG, *Combinatorial analysis of an efficient algorithm for processor and storage allocation*, SIAM J. Comput., 8 (1979), pp. 202–217.
- [7] E. G. COFFMAN, JR., J. Y. LEUNG AND D. W. TING, *Bin packing: maximizing the number of pieces packed*, Acta Informatica, 9 (1978), pp. 263–271.
- [8] P. ERDŐS AND R. L. GRAHAM, *On packing squares with equal squares*, J. Combin. Theory Ser. A, 19 (1975), pp. 119–123.
- [9] M. GARDNER, *Some packing problems that cannot be solved by sitting on the suitcase*, Scientific American, Oct. 1979, pp. 18–26.
- [10] M. R. GAREY AND D. S. JOHNSON, *Approximation algorithms for bin-packing problems: A survey*, in Analysis and Design of Algorithms in Combinatorial Optimization, G. Ausiello and M. Lucertini, eds., Springer-Verlag, New York, 1981, pp. 147–172.
- [11] D. J. KLEITMAN AND M. K. KRIEGER, *An optimal bound for two dimensional bin packing*, Proc. 16th Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Long Beach, CA, 1975, pp. 163–168.
- [12] D. K. D. B. SLEATOR, *A 2.5 times optimal algorithm for bin packing in two dimensions*, Information Processing Lett., 10 (1980), pp. 37–40.

A COMBINATORIAL CONSTRUCTION OF PERFECT CODES*

K. T. PHELPS†

Abstract. A combinatorial construction for perfect binary single error correcting codes is presented. Several results are derived from this construction. In particular, we establish that there are a large number of nonisomorphic perfect codes of length 15.

Key words. perfect codes, Steiner triple systems

AMS subject classification. 94B25, 05B30.

1. Introduction. A binary code of length n is a subset of V^n , a vector space of dimension n over $GF(2)$. Alternately one can consider a binary code as a collection of subsets of an n -set, since for every subset of an n -set there is a corresponding (characteristic) vector in V^n . Many of the more interesting codes have the property that the nonempty subsets of minimal size (i.e., vectors of minimal weight) form a t -design. There are various combinatorial constructions for t -designs and it is natural to investigate similar constructions for the corresponding codes. This is precisely the motivation behind the combinatorial construction for perfect binary codes presented here.

A perfect binary single error correcting code of length n , which we will henceforth refer to as a perfect 1-code of length n , exists when $n = 2^m - 1$, for $m \geq 3$. The linear perfect 1-codes are unique—they are simply the well-known Hamming codes. Non-linear perfect 1-codes of length n have been constructed by Vasil'ev [9] for all (admissible) n . More recently Bauer, Ganter and Hergert [1] presented some algebraic techniques for constructing nonlinear codes which enabled them to construct nonlinear perfect 1-codes of length 15 which are "non-Vasil'ev," i.e., nonequivalent to the Vasil'ev codes.

Two codes $C, D \subset V^n$ are isomorphic if there is a permutation, π , of the coordinates which maps the vectors of one code into the other (e.g. $D = \{\pi(x) | x \in C\}$). If $D = C + a = \{x + a | x \in C\}$ then D is a translation (or coset or affine subspace as the case may be) of C . Two codes are equivalent if they are isomorphic or if one code is isomorphic to the translate of another. Unless otherwise stated, we will always assume that the code has the zero vector. Establishing that two codes (designs, graphs, etc.) are nonisomorphic is in general a problematical issue. Using results and techniques from block designs we are able to establish a (probably weak) lower bound on the number of nonisomorphic perfect 1-codes of length 15.

The problem of nonequivalence of codes is more difficult and is only briefly discussed here. A thorough computational study would be needed to establish a reasonable lower bound—on the number of nonequivalent perfect 1-codes of length 15.

For any perfect 1-code of length n , the words of weight 3 (i.e., the minimal nonempty subsets) form a Steiner triple system of order n , (briefly STS(n)). Any perfect 1-code of length n can be extended to a code of length $n + 1$ by adding an overall parity check bit. This is equivalent to adding a new element, ∞ , to every subset of odd cardinality in the code. In this extended perfect 1-code, every word has even weight and the words of minimal weight (i.e., weight 4) form a Steiner quadruple

* Received by the editors June 11, 1982, and in revised form October 25, 1982.

† School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

system of order $n + 1$ (briefly SQS ($n + 1$)). Thus for a STS (n) to be contained in a perfect 1-code of length n it is necessary for it to be a derived triple system—i.e., that it can be extended to an SQS ($n + 1$). Which Steiner triple systems (of order n) are contained in a perfect 1-code (of length n)? (i.e., which STS (n) are “perfect”?). This is a difficult question even when $n = 15$.

There are only 80 nonisomorphic STS (15) ([4], cf. [2], [3]) of which 43 are known to be derived (Phelps [8]). We will show that at least 23 of these derived triple systems are “perfect,” i.e. belong to perfect 1-codes of length 15. If we consider the extended perfect 1-codes of length 16 and ask a similar question for SQS (16), we can say more. There are at least 31,021 nonisomorphic SQS (16) [6] and each of these belongs to some extended perfect 1-code of length 16. Our combinatorial construction produces many different extended 1-codes of length 16. However, completely determining the isomorphism classes of these codes is tedious, time consuming, and costly, so we have not attempted such a classification. We remark that for two perfect 1-codes of length n to be isomorphic it is necessary that the words of weight 3 (STS (n)) and weight 4 be isomorphic.

MacWilliams and Sloane [5, p. 180, problem 6.6] give the following research problem: Find all perfect nonlinear single-error-correcting codes over $GF(q)$. The results of this paper suggest that answering this question even for $q = 2$ will be impossible.

2. Construction. Given a perfect 1-code of length n we construct a perfect 1-code of length $2n + 1$ containing the given code as a “subcode.” Let $C \subset V^n$ be a perfect 1-code of length n , $n = 2^m - 1$ and $C = C_0, C_1, C_2, \dots, C_n$ be a partition of the V^n with $|C| = |C_i| = 2^{n-m}$ and such that the minimum distance between any 2 words in C_i is 3. For any code C_i , let C_i^* denote the extended code of length $n + 1$ constructed by adding an overall parity check bit. The vectors in C_i^* each have even weight and the minimum distance for any C_i^* is now 4.

Let C, B be two perfect 1-codes of length n and $\{C = C_0, C_1, \dots, C_n\}$ and $\{B = B_0, B_1, \dots, B_n\}$ be any two partitions of V^n having the properties described in the preceding paragraph. Let B_i^*, C_i^* denote the extended codes and let α be any permutation of $\{0, 1, \dots, n\}$. Define $E^* \subset V^{2n+2}$ as follows:

$$(b, d) \in E^* \text{ if and only if } b \in C_i^*, d \in B_j^* \text{ and } \alpha(i) = j.$$

THEOREM 2.1. *The code, E^* , constructed above is an extended perfect 1-code of length $2n + 2$.*

Proof. E^* has $(n + 1)(2^{n-m})^2$ codewords of length $2(n + 1) = 2^{m+1}$, which is the correct number of codewords. The distance between any 2 codewords is 4. If $x, y \in E, x = (b, d), y = (b', d')$, then the distance between x and y is $d(x, y) = d(b, b') + d(d, d')$. If $b = b'$, and $b \in C_i^*$ then $d, d' \in B_j^*$ where $\alpha(i) = j$. Thus by definition of the B_j^* , $d(d, d') \geq 4$ and $d(x, y) \geq 4$. Similarly if $d = d'$, then $d(b, b') \geq 4$. Suppose $b \neq b'$ and $d \neq d'$; then since each of b, b', d, d' has even weight $d(b, b') \geq 2$ and $d(d, d') \geq 2$ and thus $d(x, y) \geq 4$.

If we assume that $\alpha(0) = 0$, then the 0-vector will be in E^* and both C^* and B^* will be subcodes of E^* . Puncturing E^* , i.e., deleting a coordinate, gives a perfect 1-code, E , of length $2n + 1$. (Considering E^* as a collection of subsets this is equivalent to deleting the element i from all subsets of E^* to which it belongs.) By choosing the appropriate coordinate we will have C as a subcode of E .

LEMMA 2.2. *For every perfect 1-code $C \subset V^n$, there exists a partition $C = C_0, C_1, \dots, C_n$ of V with $|C| = |C_i|$ for $i = 0, 1, \dots, n$, and the minimum distance for C_i is 3.*

Proof. Let $x_1, x_2, \dots, x_n \in V^n$ be the n vectors of weight 1. Choose the C_i to be translates of C , that is, $C_i = C + x_i$, $i = 1, 2, \dots, n$. Since C is perfect, every vector, $y \notin C$, has distance 1 from some codeword and hence $y \in C + x_i$ for some i .

We see that for every perfect 1-code C we can construct at least one partition of V^n having the required properties. For a given code, C , are there other, nonisomorphic, partitions? Given a code C and partition $C = C_0, C_1, \dots, C_n$ consider its extension $C^* = C_0^*, C_1^*, \dots, C_n^*$. Let us consider the codewords as subsets, and let $F_i \subset C_i^*$ be the set of 2-element subsets of C_i^* , $i = 1, 2, \dots, n$. Since each C_i^* has minimal distance 4 and every 2-subset is in some C^* , we conclude that F_1, F_2, \dots, F_n is a 1-factorization of the complete graph, K_{n+1} . If the partition C_0, C_1, \dots, C_n was formed by taking translates of C , then the 1-factorization is a Steiner 1-factorization and is in effect constructed from the Steiner triple system contained in C . There are of course many other 1-factorizations and it would be interesting to know which of these arise from such partitions of V^n . Since there are on the order of $n^{n^2/2}$ nonisomorphic 1-factorizations of K_n , this suggests that the number of nonisomorphic partitions of V^n for a given code, C , could be large indeed. While we cannot prove this, we do show how one can construct a new partition from a given one.

Let $C^* = C_0^*, C_1^*, \dots, C_n^*$ be such a partition of V . For $i, j \geq 1$ we form a "graph" on the codewords of C_i^*, C_j^* , where the codewords are the vertices and two codewords x, y are adjacent if and only if their distance $d(x, y) = 2$. This gives us a bipartite graph. If it is not connected, then for any component $G_1 \cup G_2$ where $G_1 \subset C_i^*$ and $G_2 \subset C_j^*$ we can "switch" the codewords so that $(C_i^* \setminus G_1) \cup G_2$ and $(C_j^* \setminus G_2) \cup G_1$ are now classes. Replacing C_i^* and C_j^* by these new classes gives us a different partition of V^n which still has the required properties. This "switching" process is a common approach used in the construction of nonisomorphic designs, Latin squares and other combinatorial configurations. (For example, compare the extended partitions I, II in the following section).

3. Perfect 1-codes of length 15. In this section, we apply the ideas of the previous section to the construction of perfect 1-codes of length 15. This involves finding nonisomorphic partitions of V^7 . Fortunately the perfect 1-code of length 7 is unique so we only need to construct nonisomorphic partitions for this code. As we remarked previously, each such partition of V induces a 1-factorization of K_8 . As is well-known, there are exactly 6 nonisomorphic 1-factorizations of K_8 (cf. Wallis [10], Brouwer [2]). It is an easy matter to test whether two such 1-factorizations are isomorphic. Using various approaches we construct 6 nonisomorphic (and nonequivalent) partitions (listed below) which contain each of the nonisomorphic 1-factorizations I, II, III, IV, V, VI respectively (cf. Brouwer's listing [2]). Applying Theorem 2.1 to these 6 nonisomorphic partitions of V_7 gives us at least 31,021 different extended perfect 1-codes of length 16. If we consider the effect of this construction on the words of weight 4—i.e., the Steiner quadruple systems, we see that it is nothing more than the well-known doubling construction for such designs (cf. Phelps [7]). Puncturing these extended codes will produce perfect 1-codes of length 15. The words of weight 3 will be one of 23 nonisomorphic STS (15) (#1–22, 61 in Bussemaker and Seidel's listing [3]). By an appropriate choice of the partitions and the permutation α one can insure that each of these STS (15) are contained in at least one perfect 1-code. (Cf. Brouwer [2] p. 11.)

COROLLARY 3.1. *There are at least 31,021 perfect 1-codes of length 15.*

Proof. Lindner and Rosa [6] constructed 31,021 nonisomorphic SQS (16). Each one of these will be contained in one of the perfect extended codes constructed above.

TABLE 1

	$C^* = \{0, 1, 2, 3\}$	$\{0, 2, 5, 7\}$	$\{1, 2, 5, 6\}$	$\{2, 3, 6, 7\}$
	$\{0, 1, 4, 5\}$	$\{0, 3, 4, 7\}$	$\{1, 3, 4, 6\}$	$\{4, 5, 6, 7\}$
	$\{0, 1, 6, 7\}$	$\{0, 3, 5, 6\}$	$\{1, 3, 5, 7\}$	
	$\{0, 2, 4, 6\}$	$\{1, 2, 4, 7\}$	$\{2, 3, 4, 5\}$	
	$\emptyset, \{0, 1, \dots, 7\}$			

I.	C_1^*	C_2^*	C_3^*	C_4^*	C_5^*	C_6^*	C_7^*
	0, 1	0, 2	0, 3	0, 4	0, 5	0, 6	0, 7
	2, 3	1, 3	1, 2	1, 5	1, 4	1, 7	1, 6
	4, 5	4, 6	4, 7	2, 6	2, 7	2, 4	2, 5
	6, 7	5, 7	5, 6	3, 7	3, 6	3, 5	3, 4
	0, 2, 4, 7	0, 1, 4, 7	0, 1, 4, 6	0, 1, 2, 7	0, 1, 2, 6	0, 1, 2, 5	0, 1, 2, 4
	0, 2, 5, 6	0, 1, 5, 6	0, 1, 5, 7	0, 1, 3, 6	0, 1, 3, 7	0, 1, 3, 4	0, 1, 3, 5
	0, 3, 4, 6	0, 3, 4, 5	0, 2, 4, 5	0, 2, 3, 5	0, 2, 3, 4	0, 2, 3, 7	0, 2, 3, 6
	0, 3, 5, 7	0, 3, 6, 7	0, 2, 6, 7	0, 5, 6, 7	0, 4, 6, 7	0, 4, 5, 7	0, 4, 5, 6
	(plus complements of these codewords)						

II.	C_1^*	C_2^*	C_3^*	C_4^*	C_5^*	C_6^*	C_7^*
	0, 1	0, 2	0, 3	0, 4	0, 5	0, 6	0, 7
	2, 3	1, 3	1, 2	1, 5	1, 4	1, 7	1, 6
	4, 5	4, 6	4, 7	2, 6	2, 7	<u>2, 5</u>	<u>2, 4</u>
	6, 7	5, 7	5, 6	3, 7	3, 6	<u>3, 4</u>	<u>3, 5</u>
	0, 2, 4, 7	0, 1, 4, 7	0, 1, 4, 6	0, 1, 2, 7	0, 1, 2, 6	<u>0, 1, 2, 4</u>	<u>0, 1, 2, 5</u>
	0, 2, 5, 6	0, 1, 5, 6	0, 1, 5, 7	0, 1, 3, 6	0, 1, 3, 7	<u>0, 1, 3, 5</u>	<u>0, 1, 3, 4</u>
	0, 3, 4, 6	0, 3, 4, 5	0, 2, 4, 5	0, 2, 3, 5	0, 2, 3, 4	0, 2, 3, 7	0, 2, 3, 6
	0, 3, 5, 7	0, 3, 6, 7	0, 2, 6, 7	0, 5, 6, 7	0, 4, 6, 7	0, 4, 5, 7	0, 4, 5, 6
	(plus the complements of these codewords)						

III.	(isomorphic)	C_1^*	C_2^*	C_3^*	C_4^*	C_5^*	C_6^*	C_7^*
		0, 1	0, 2	0, 3	0, 4	0, 5	0, 6	0, 7
		2, 3	1, 3	1, 2	1, 6	1, 7	1, 4	1, 5
		4, 5	4, 6	4, 7	2, 7	2, 6	2, 5	2, 4
		6, 7	5, 7	5, 6	3, 5	3, 4	3, 7	3, 6
		0, 2, 4, 7	0, 1, 4, 7	0, 1, 4, 6	0, 1, 2, 5	0, 1, 2, 4	0, 1, 2, 7	0, 1, 3, 4
		0, 3, 4, 6	0, 1, 5, 6	0, 1, 5, 7	0, 1, 3, 7	0, 1, 3, 6	0, 1, 3, 5	0, 1, 2, 6
		0, 2, 5, 6	0, 3, 4, 5	0, 2, 4, 5	0, 2, 3, 6	0, 2, 3, 7	0, 2, 3, 4	0, 2, 3, 5
		0, 3, 5, 7	0, 3, 6, 7	0, 2, 6, 7	0, 5, 6, 7	0, 4, 6, 7	0, 4, 5, 7	0, 4, 5, 6
	(plus the complements of these codewords)							

IV.	C_1^*	C_2^*	C_3^*	C_4^*	C_5^*	C_6^*	C_7^*
	0, 1	0, 2	0, 3	0, 4	0, 5	0, 6	0, 7
	2, 3	1, 3	2, 1	1, 6	1, 4	1, 7	1, 5
	4, 5	4, 6	4, 7	2, 7	2, 6	2, 5	2, 4
	6, 7	5, 7	5, 6	3, 5	3, 7	3, 4	3, 6
	0, 2, 4, 7	0, 1, 4, 7	0, 2, 4, 5	0, 1, 2, 5	0, 1, 2, 7	0, 1, 2, 4	0, 1, 2, 6
	0, 2, 5, 6	0, 1, 5, 6	0, 2, 6, 7	0, 1, 3, 7	0, 1, 3, 6	0, 1, 3, 5	0, 1, 3, 4
	0, 3, 4, 6	0, 3, 4, 5	0, 1, 4, 6	0, 2, 3, 6	0, 2, 3, 4	0, 2, 3, 7	0, 2, 3, 5
	0, 3, 5, 7	0, 3, 6, 7	0, 1, 5, 7	0, 5, 6, 7	0, 4, 6, 7	0, 4, 5, 7	0, 4, 5, 6
	(plus complements)						

TABLE 1—continued

V.	$C^* = C_0^* = 0, 1, 2, 3$ 0, 1, 4, 6		0, 1, 5, 7 0, 2, 4, 5	0, 2, 6, 7 0, 3, 4, 7	0, 3, 5, 6 \emptyset		
	(plus complements)						
	C_1^*	C_2^*	C_3^*	C_4^*	C_5^*	C_6^*	C_7^*
	0, 1	0, 2	0, 3	0, 4	0, 5	0, 6	0, 7
	2, 3	1, 3	1, 4	1, 7	1, 6	1, 2	1, 5
	4, 5	4, 6	2, 7	2, 5	2, 4	3, 5	2, 6
	6, 7	5, 7	5, 6	3, 6	3, 7	4, 7	3, 4
	0, 2, 4, 7	0, 1, 4, 7	0, 1, 2, 5	0, 1, 2, 6	0, 1, 2, 7	0, 1, 3, 7	0, 1, 2, 4
	0, 2, 5, 6	0, 1, 5, 6	0, 1, 6, 7	0, 1, 3, 5	0, 1, 3, 4	0, 1, 4, 5	0, 1, 3, 6
	0, 3, 4, 6	0, 3, 4, 5	0, 2, 4, 6	0, 2, 3, 7	0, 2, 3, 6	0, 2, 3, 4	0, 2, 3, 5
	0, 3, 5, 7	0, 3, 6, 7	0, 4, 5, 7	0, 5, 6, 7	0, 4, 6, 7	0, 2, 5, 7	0, 4, 5, 6
	(plus complements)						
VI.	$C^* = C_0^* = 0, 2, 3, 5$ 0, 2, 6, 7		0, 1, 5, 6 0, 3, 4, 6	0, 4, 5, 7 0, 1, 3, 7	0, 1, 2, 4 \emptyset		
	(plus complements)						
	C_1^*	C_2^*	C_3^*	C_4^*	C_5^*	C_6^*	C_7^*
	0, 1	0, 2	0, 3	0, 4	0, 5	0, 6	0, 7
	2, 7	1, 3	1, 5	1, 7	1, 2	1, 4	1, 6
	3, 6	4, 7	2, 4	2, 6	3, 7	2, 3	2, 5
	4, 5	5, 6	6, 7	3, 5	4, 6	5, 7	3, 4
	0, 2, 5, 6	0, 1, 4, 6	0, 1, 4, 7	0, 1, 2, 5	0, 1, 6, 7	0, 1, 2, 7	0, 1, 2, 3
	0, 2, 3, 4	0, 1, 5, 7	0, 1, 2, 6	0, 1, 3, 6	0, 1, 3, 4	0, 1, 3, 5	0, 1, 4, 5
	0, 3, 5, 7	0, 3, 4, 5	0, 2, 5, 7	0, 5, 6, 7	0, 2, 3, 6	0, 2, 4, 5	0, 2, 4, 6
	0, 4, 6, 7	0, 3, 6, 7	0, 4, 5, 6	0, 2, 3, 7	0, 2, 4, 7	0, 3, 4, 7	0, 3, 5, 6
	(plus complements)						

Hence there must be at least this many perfect 1-codes of length 15. We remark that there are other (possibly nonisomorphic) partitions than those listed below.

In Table 1 below the codewords are listed as subsets. Whenever a codeword is in some C^* so is its complement. Hence to save space we do not include complements.

4. Equivalence. Although there are a large number of nonisomorphic extended perfect 1-codes of length 16, the number of nonequivalent codes can be realized by the previous combinatorial construction could be drastically smaller. However, it appears that here should be several thousand at least.

In support of this statement we point out that the six partitions, I–VI, listed above are all nonequivalent, not just nonisomorphic, even though each of the sets C_i^* is equivalent to the (linear) extended Hamming code of length 8. We use the “distance graphs” mentioned earlier to establish this claim.

Given a partition $C_0^*, C_1^*, \dots, C_n^*$, where each C_i^* is an extended perfect 1-code, define a graph G_{ij} on $C_i^* \cup C_j^*$ for each $i \neq j, i, j = 0, 1, \dots, n$. The codewords are the vertices and two codewords are adjacent if and only if their Hamming distance is two. Since distance is invariant under equivalence, equivalence partitions must have isomorphic collections of graphs G_{ij} .

In the case at hand, each G_{ij} is a 4-regular bipartite graph on 32 vertices. The subgraph F_{ij} induced by the codewords of weight two in $C_i^* \cup C_j^*$ ($i, j \geq 1$) is 2-regular and will be either an 8-cycle or two disjoint 4-cycles. If it is an 8-cycle then it is easy to see that the containing graph G_{ij} must consist of one component. If F_{ij} consists of

two 4-cycles then G_{ij} can have two components. Since the subgraphs F_{ij} characterize the nonisomorphic 1-factorizations of K_8 and have been well studied (cf. Brouwer [2, pp. 5–10]) we can use this information to quickly determine that the sets of graphs G_{ij} , for the different partitions, are not isomorphic and hence the partitions are not equivalent.

Establishing the existence of nonequivalent partitions, unfortunately, is only a first step to establishing a lower bound on the number of nonequivalent perfect 1-codes of length 16. However it does lend support to the conjecture that there are at least several thousand such codes.

5. Conclusion. Finding all partitions of V^7 associated with the Hamming codes is a computationally tractable problem. From these reasonable lower bounds on the number of nonequivalent perfect codes of length 15 could be computed. For the vector space over $GF(2)$ of dimension 15 (i.e., the next case), the problem will not be tractable.

Recent work (unpublished, F. Hergert) has established that some previously known perfect 1-codes of length 15 (cf. Vasil'ev [9], Bauer, Ganter, and Hergert [1]) cannot be realized by the combinatorial construction presented above. It would be interesting to see how the algebraic techniques of Bauer, Ganter and Hergert relate to the combinatorial techniques discussed above.

REFERENCES

- [1] H. BAUER, B. GANTER AND F. HERGERT, *Algebraic techniques for nonlinear codes*, preprint Nr 609, Technische Hochschule Darmstadt, Fachbereich Mathematik.
- [2] A. E. BROUWER, *The linear spaces on 15 points*, *Ars Combinatoria*, 12 (1981), pp. 3–35.
- [3] F. C. BUSSEMAKER AND J. J. SEIDEL, *Symmetric Hadamard matrices of order 36*, Report 70 WSK-02, Technological University Eindhoven, the Netherlands, 1970.
- [4] F. N. COLE, L. D. CUMMINGS AND H. S. WHITE, *The complete enumeration of trial systems in 15 elements*, *Proc. Nat. Acad. Sci. USA*, 3 (1917), pp. 197–199.
- [5] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1978.
- [6] C. C. LINDNER AND A. ROSA, *There are at least 31,021 nonisomorphic Steiner quadruple systems of order 16*, *Utilitas Math.*, 10 (1976), pp. 61–64.
- [7] K. T. PHELPS, *Some sufficient conditions for a Steiner triple system to be a derived triple system*, *J. Combin. Theory A*, 20 (1976), pp. 393–397.
- [8] K. T. PHELPS, *A survey of derived triple systems*, *Ann. Discr. Math.*, 7 (1980), pp. 105–114.
- [9] J. L. VASIL'EV, *On nongroup closepacked codes*, *Probl. Kibernet.*, 8 (1962), pp. 337–339. (In Russian.)
- [10] W. D. WALLIS, *Some results on root square isomorphism*, Res. Rep. 86, Univ. Newcastle, 1973.

ESTIMATION OF SPARSE JACOBIAN MATRICES*

GARRY N. NEWSAM† AND JOHN D. RAMSDELL‡

Abstract. When finding a numerical solution to a system of nonlinear equations, one often estimates the Jacobian by finite differences. Curtis, Powell and Reid [J. Inst. Math. Applics., 13 (1974), pp. 117-119] presented an algorithm that reduces the number of function evaluations required to estimate the Jacobian by taking advantage of sparsity. We show that the problem of finding the best of the Curtis, Powell and Reid type algorithms is NP-complete, and then propose two procedures for estimating the Jacobian that may use fewer function evaluations.

Key words. sparse nonlinear equations, sparse Jacobian estimation, graph coloring

1. Notation.

\mathbf{Z}_n is the integers modulo n .

\mathbf{x} is a vector in \mathbf{R}^n with components x_0, x_1, \dots, x_{n-1} .

X is an $n \times m$ matrix.

X_j is the j th column of X .

X_{ij} is the i th component of the column vector X_j .

\mathbf{x}^k is one of a collection of vectors in \mathbf{R}^n .

$g(\mathbf{x})$ is a function $g: \mathbf{R}^n \rightarrow \mathbf{R}$.

$\mathbf{f}(\mathbf{x})$ is a function $\mathbf{f}: \mathbf{R}^n \rightarrow \mathbf{R}^n$.

$\partial/\partial x_i$ will be abbreviated by ∂_i .

$\partial g(\mathbf{x})$ is the row vector whose components are $\partial_i g(\mathbf{x})$.

$J(\mathbf{x})$ is the Jacobian matrix, i.e. $J_{ij}(\mathbf{x}) \equiv \partial_j f_i(\mathbf{x})$.

2. Introduction. When solving a system of n nonlinear equations

$$(1) \quad \mathbf{f}(\mathbf{x}) = 0,$$

many algorithms require the estimation of the Jacobian matrix. The most straightforward estimate, requiring $n + 1$ function evaluations and n vector differences, is

$$J_j(\mathbf{x}) \equiv \partial_j \mathbf{f}(\mathbf{x}) \cong \frac{\mathbf{f}(\mathbf{x} + hI_j) - \mathbf{f}(\mathbf{x})}{h}$$

where I is the identity matrix and h is a small parameter.

In many large systems $J(\mathbf{x})$ is sparse; if a particular element $J_{ij}(\mathbf{x})$ is known to be zero, then obviously $\mathbf{f}_i(\mathbf{x} + hI_j)$ need not be calculated. However in many applications the components of \mathbf{f} cannot be calculated independently, e.g. in a program in which the evaluation of all components of \mathbf{f} is done using a single call to a procedure. For such a system it is no longer obvious how to take advantage of sparsity to reduce the number of function evaluations needed to approximate $J(\mathbf{x})$.

Nevertheless Curtis, Powell and Reid, henceforth abbreviated to CPR, have shown in [3] how to approximate the Jacobian using fewer than $n + 1$ function

* Received by the editors October 23, 1981, and in revised form September 16, 1982.

† Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138. The research of this author was supported in part by National Science Foundation grant no. MCS 80-05863 and Harvard University.

‡ Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138. The research of this author was supported in part by Fire Center of the National Bureau of Standards grant no. G7-9011 and Harvard University.

evaluations. They realized that if the sparsity of $J(\mathbf{x})$ is such that for two columns $J_j(\mathbf{x})$ and $J_k(\mathbf{x})$, $J_{ij}(\mathbf{x}) = 0$ or $J_{ik}(\mathbf{x}) = 0$ for every i , then

$$J_{ij}(\mathbf{x}) \text{ or } J_{ik}(\mathbf{x}) = J_{ij}(\mathbf{x}) + J_{ik}(\mathbf{x}) \cong \frac{\mathbf{f}_i(\mathbf{x} + h\mathbf{I}_j + h\mathbf{I}_k) - \mathbf{f}(\mathbf{x})}{h}.$$

Thus the evaluation of \mathbf{f} at \mathbf{x} and $\mathbf{x} + h\mathbf{I}_j + h\mathbf{I}_k$ gives one difference from which two columns of $J(\mathbf{x})$ may be estimated. In [3], CPR use the above principle to construct $m \leq n$ vectors X_k from the given zero-nonzero structure of $J(\mathbf{x})$. X has the property that every column of $J(\mathbf{x})$ may be estimated from some difference $\mathbf{f}(\mathbf{x} + X_k) - \mathbf{f}(\mathbf{x})$, and, where possible, several columns are estimated simultaneously from the same difference.

CPR left unanswered the question of whether the $m + 1$ function evaluations used to estimate $J(\mathbf{x})$ were minimal. We therefore address the general problem of estimating a sparse Jacobian in a minimum number of function evaluations under the constraint that the components of \mathbf{f} cannot be evaluated independently. The thrust of the paper is the use of precise definitions of admissible estimates to solve this problem. In the next section we formalize the CPR principle and define admissible estimates based on the principle. We give a complete account of the complexity of finding admissible CPR estimates using a minimum number of function evaluations; we show in general that such estimates are hard to construct, but give an example in which they can be found with ease. In § 4 we propose a broader definition of admissible estimations and show that under this definition, the Jacobian can be estimated with $m + 1$ function evaluations, where m is the maximum of the number of nonzero elements in any row of the Jacobian. This estimate requires the same number or fewer function evaluations than the CPR estimate, but in general to find this approximation an additional n small linear systems must be solved. By comparison in CPR estimates the elements $J_{ij}(\mathbf{x})$ may be read off directly from a matrix of differences. Finally, in the Appendix, we consider an extension of the CPR principle.

For clarity $J(\mathbf{x})$ has been presented as a square matrix, but a trivial padding construction extends the results below to the estimation of rectangular $J(\mathbf{x})$ (i.e. the system in (1) is either under or over determined). We leave such constructions to the reader and where it is convenient to work with rectangular Jacobians we shall do so, with the implicit assumption that such a padding will be used to render the Jacobian square.

It follows that the algorithms described will efficiently estimate $J(\mathbf{x})$ when $\mathbf{f}(\mathbf{x})$ may be partitioned into two or more subvectors $\mathbf{f}^i(\mathbf{x})$ such that $\mathbf{f}^i(\mathbf{x})$ can be calculated independently of $\mathbf{f}^j(\mathbf{x})$ for $i \neq j$, but the components of $\mathbf{f}^i(\mathbf{x})$ must be calculated together. Such a partition of $\mathbf{f}(\mathbf{x})$ induces a corresponding partition of $J(\mathbf{x})$ into rectangular submatrices, e.g.

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \mathbf{f}^1(\mathbf{x}) \\ \mathbf{f}^2(\mathbf{x}) \end{bmatrix} \Rightarrow J(\mathbf{x}) = \begin{bmatrix} J^1(\mathbf{x}) \\ J^2(\mathbf{x}) \end{bmatrix}.$$

Efficient estimation of $J(\mathbf{x})$ now reduces to the independent problems of efficient estimation of each $J^i(\mathbf{x})$ from differences of the form $\mathbf{f}^i(\mathbf{x} + X_k^i) - \mathbf{f}^i(\mathbf{x})$.

3. The CPR principle and Jacobian estimation. CPR reduce the number of function evaluations required to estimate the Jacobian by taking advantage of the zero-nonzero structure of the Jacobian matrix. Actually CPR take advantage of the known-unknown structure of the Jacobian matrix. Without loss of generality,

the known elements can be replaced by zeros by defining $M(\mathbf{x})$ and $\bar{J}(\mathbf{x})$ as

$$M_{ij}(\mathbf{x}) = \begin{cases} J_{ij}(\mathbf{x}) & \text{if } J_{ij}(\mathbf{x}) \text{ is known,} \\ 0 & \text{otherwise,} \end{cases}$$

and $\bar{J}(\mathbf{x}) \equiv J(\mathbf{x}) - M(\mathbf{x})$. Then $\bar{J}(\mathbf{x})$ has the required zero-nonzero structure for the application of the CPR principle. The CPR approach is best described using the concept of isolated variables in (1).

DEFINITION 1. \mathbf{x}_i is isolated from \mathbf{x}_j iff $\forall \mathbf{x} \forall k \in \mathbf{Z}_n \partial_i \mathbf{f}_k(\mathbf{x}) \partial_j \mathbf{f}_k(\mathbf{x}) = 0$.

This concept allows a formal statement of the CPR principle outlined in the introduction and an associated definition of an m -CPR-estimable Jacobian.

DEFINITION 2. *The CPR principle.* Let $\{C_k\}_{k=0}^{m-1}$ be a partition of \mathbf{Z}_n . For all $i \in C_k$ the columns $J_i(\mathbf{x})$ may be estimated from one difference

$$\mathbf{f}(\mathbf{x} + X_k) - \mathbf{f}(\mathbf{x}) \text{ with } X_k = h \sum_{i \in C_k} I_i$$

if $\forall i, j \in C_k, i = j$ or \mathbf{x}_i is isolated from \mathbf{x}_j .

DEFINITION 3. The Jacobian $J(\mathbf{x})$ is m -CPR-estimable iff there exists an $n \times m$ matrix X , as in Definition 2, such that every column of $J(\mathbf{x})$ may be estimated from one column of the matrix $B(\mathbf{x})$ where

$$B: \mathbf{R}^n \rightarrow \mathbf{R}^{n \times m}, \quad B_k(\mathbf{x}) = \mathbf{f}(\mathbf{x} + X_k) - \mathbf{f}(\mathbf{x}),$$

and columns estimated by the same difference correspond to isolated variables.

To establish the complexity of deciding whether an arbitrary Jacobian is m -CPR-estimable, we first show that the problem is equivalent to deciding if a special graph is m -colorable, thus showing that the problem is no harder than graph coloring. We next show that if we could decide whether any Jacobian is m -CPR-estimable then we could decide if an arbitrary graph is m -colorable. This shows graph coloring is no harder than finding a CPR estimation. Therefore they have the same complexity and finding an m -CPR-estimation with the smallest m is equivalent to finding a minimum graph coloring, an NP-complete problem [4], [5].

3.1. Graph coloring. A graph $G = (V, E)$ is a set V of vertices and a set E of edges which are unordered pairs of vertices. An edge e is denoted by (u, v) and connects vertex u to vertex v . An m -coloring of the graph is an assignment of one of m colors to each vertex such that no two vertices in the graph have the same color if they are connected by an edge. A minimum coloring is a coloring of the graph such that there exists no coloring of the graph using a smaller number of colors.

3.2. The variable isolation graph. To reduce CPR estimation to graph coloring we construct for every Jacobian $J(\mathbf{x})$ an associated graph G such that an m -CPR-estimation of $J(\mathbf{x})$ is equivalent to an m -coloring of G .

DEFINITION 4. The variable isolation graph $G = (V, E)$ associated with (1) is

$$V = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{n-1}\}$$

$$E = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ is not isolated from } \mathbf{x}_j\}.$$

THEOREM 1. *The variable isolation graph is m -colorable iff $J(\mathbf{x})$ is m -CPR-estimable.*

Proof. Given an m -CPR-estimation, we can color the variable isolation graph with m colors by assigning the same color to vertices corresponding to columns

evaluated together. Variables of the same color will not be connected because they must be isolated for the associated columns to be estimated simultaneously. Given an m -coloring of the variable isolation graph, one constructs an m -CPR-estimation of the Jacobian by estimating columns associated with vertices of the same color in the same vector difference.

3.3. The incidence matrix. We now show that every graph G is the variable isolation graph of some Jacobian $J(\mathbf{x})$.

DEFINITION 5. The incidence matrix $D \in \mathbf{R}^{|E| \times |V|}$ associated with an arbitrary graph $G = (V, E)$ has entries $D_{ij} = D_{ik} = 1$ iff $e_i = (v_j, v_k)$ with the remaining elements being zero.

THEOREM 2. G is m -colorable iff the incidence matrix D is m -CPR-estimable.

Proof. The incidence matrix D of G is the Jacobian of the function $\mathbf{f} : \mathbf{R}^{|V|} \rightarrow \mathbf{R}^{|E|}$ defined by

$$\mathbf{f}_i(\mathbf{x}) = \mathbf{x}_j + \mathbf{x}_k \text{ iff } e_i = (v_j, v_k).$$

G is the variable isolation graph of \mathbf{f} and from Theorem 1 G is m -colorable iff \mathbf{f} 's Jacobian is m -CPR-estimable.

We have shown that an arbitrary $n \times n$ Jacobian is m -CPR-estimable iff the associated isolation graph on n vertices is m -colorable, and that an arbitrary graph on n vertices is m -colorable iff the associated incidence matrix of dimension at most $n^2/2 \times n^2/2$ is m -CPR-estimable. This shows that the complexities of graph coloring and CPR estimation are the same.

3.4. Minimum CPR algorithms. A minimum CPR algorithm is an algorithm which takes the structure of an arbitrary Jacobian $J(\mathbf{x})$ as input and outputs the smallest m such that $J(\mathbf{x})$ is m -CPR-estimable together with an X that achieves this estimate. Since deciding if an arbitrary Jacobian is m -CPR-estimable is NP-complete, one cannot construct an efficient minimum CPR algorithm for arbitrary Jacobians in the same sense that one cannot construct an efficient algorithm for minimum graph coloring of arbitrary graphs. Therefore one must use some approximate algorithm [3], or, by Theorem 1, an approximate algorithm for minimum graph coloring [2].

For particular problems with regular structure exact minima may be easily found. A practical example arises in the solution of nonlinear partial differential equations

$$(Pu)(\mathbf{s}, \mathbf{t}) = 0$$

on square domains. If the differential operator is replaced by a finite difference operator using function values on an $m \times m$ grid of points $(\mathbf{s}_i, \mathbf{t}_j)$ (giving a new operator \bar{P}) the problem is now a system of $n = m^2$ nonlinear equations in m^2 variables;

$$\mathbf{x}_{ij} = u(\mathbf{s}_i, \mathbf{t}_j), \quad \mathbf{f}_{ij}(\mathbf{x}) = (\bar{P}u)(\mathbf{s}_i, \mathbf{t}_j)$$

where the double index ij represents the single index $i + mj$.

If the finite difference scheme gives a 5 point stencil, two stencils centered at \mathbf{x}_{ij} and \mathbf{x}_{kl} overlap iff $\|(i, j) - (k, l)\|_1 \leq 2$. Overlapping stencils indicate nonisolated variables. The variable isolation graph for a 5 point stencil is

$$V = \{\mathbf{x}_{ij} \mid i, j \in \mathbf{Z}_m\},$$

$$E = \{(\mathbf{x}_{ij}, \mathbf{x}_{kl}) \mid \|(i, j) - (k, l)\|_1 \leq 2\}.$$

A minimum coloring of this graph is given by

$$\text{color}(\mathbf{x}_{ij}) \equiv (i + 3j) \pmod{5}.$$

The matrix X with 5 columns is given by

$$X \in \mathbf{R}^{n \times 5}, \quad X_k = h \sum_{p \in C_k} I_p,$$

$$C_k = \{i + mj \mid \text{color}(\mathbf{x}_{ij}) = k\}.$$

If the finite difference scheme leads to a 9 point stencil, then stencils overlap iff $\|(i, j) - (k, l)\|_\infty \leq 2$ and a minimum coloring is given by

$$\text{color}(\mathbf{x}_{ij}) \equiv (i + 3j) \pmod{9}.$$

4. More general Jacobian estimation. In the previous section we addressed the question of whether a Jacobian is m -estimable under the restriction that estimates be made using the CPR principle only. The principle gives an intuitive but limited insight into possibilities of efficient estimation. Since each difference is processed independently of all others, information about an element that appears in more than one difference is not pooled. Furthermore differences are restricted to essentially the standard coordinate directions. The following simple example illustrates these limitations and indicates how more efficient estimates are possible.

Example. If $J(\mathbf{x})$ has the zero-nonzero structure

$$J(\mathbf{x}) = \begin{bmatrix} \times & \times & 0 & \times \\ \times & \times & \times & 0 \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{bmatrix},$$

then the associated variable isolation graph is the complete graph on four vertices. Therefore a minimum coloring requires four colors and a CPR estimate requires five function evaluations. However with the choice of

$$X = h \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

and the matrix B defined in Definition 3,

$$B \cong JX = h \begin{bmatrix} \partial_0 \mathbf{f}_0 + \partial_3 \mathbf{f}_0 & \partial_1 \mathbf{f}_0 & \partial_0 \mathbf{f}_0 \\ \partial_0 \mathbf{f}_1 & \partial_1 \mathbf{f}_1 & \partial_0 \mathbf{f}_1 + \partial_2 \mathbf{f}_1 \\ \partial_3 \mathbf{f}_2 & \partial_1 \mathbf{f}_2 & \partial_2 \mathbf{f}_2 \\ \partial_3 \mathbf{f}_3 & 0 & \partial_2 \mathbf{f}_3 \end{bmatrix}$$

and $J(\mathbf{x})$ may be read off from B using two subtractions. This estimate requires only four function evaluations.

This section presents a more general definition of a Jacobian estimate, based on local approximations by affine functions, that includes estimates such as that in the example. With this estimate and a careful choice of matrix X , a Jacobian with known zero-nonzero structure may be estimated using $m + 1$ function evaluations, where m is the greatest number of nonzero elements in any row of $J(\mathbf{x})$. Moreover this is the minimum possible number with which $J(\mathbf{x})$ may be estimated. The matrix X is chosen so that the estimate \bar{J} is the unique matrix with the same sparsity as $J(\mathbf{x})$ that minimizes $\|X^T \bar{J}^T - B^T\|_F$, where $\|\cdot\|_F$ is the Frobenius norm and B is defined in Definition 3.

4.1. A general gradient estimate. The i th row of the Jacobian $J(\mathbf{x})$ is the gradient of the i th component function of the vector $\mathbf{f}(\mathbf{x})$. Therefore a definition of a Jacobian estimate may be decomposed to give a definition of a gradient estimate, and conversely a Jacobian estimate definition may be constructed from a gradient estimate definition. For a scalar function $g(\mathbf{x})$ with gradient $\partial g(\mathbf{x})$, reasonable conditions that an estimate \mathbf{z} to $\partial g(\mathbf{x})^T$ at \mathbf{x} , constructed from the function values $g(\mathbf{x}^k)$ at the points \mathbf{x}^k , $k \in \mathbf{Z}_{m+1}$, should satisfy are

(i) \mathbf{z} is unique.

(ii) \mathbf{z} is a linear function of the data $g(\mathbf{x}^k)$; i.e. if \mathbf{z} and \mathbf{z}' are estimates to $\partial g(\mathbf{x})$ and $\partial g'(\mathbf{x})$ constructed from data over the same points \mathbf{x}^k then $\beta\mathbf{z} + \gamma\mathbf{z}'$ will be the estimate to $\partial(\beta g + \gamma g')(\mathbf{x})$ constructed from data on the points \mathbf{x}^k .

(iii) \mathbf{z}^T is the gradient of an affine function which is a best approximation to $g(\mathbf{x})$ on the points \mathbf{x}^k in some norm.

An affine function $h(\mathbf{x})$ with gradient \mathbf{z}^T may be written as

$$h(\mathbf{x}) = \alpha + \mathbf{z}^T (\mathbf{x} - \mathbf{x}^m)$$

where $\alpha \in \mathbf{R}$ is undetermined. Condition (iii) requires that for each estimate $\bar{\mathbf{z}}$ there exists a scalar $\bar{\alpha}$ such that the pair $(\bar{\alpha}, \bar{\mathbf{z}})$ minimize $\|\mathbf{r}(\alpha, \mathbf{z})\|$ where

$$\mathbf{r} \in \mathbf{R}^{m+1}, \quad \mathbf{r}_k = g(\mathbf{x}^k) - \alpha - \mathbf{z}^T (\mathbf{x}^k - \mathbf{x}^m), \quad k \in \mathbf{Z}_{m+1}.$$

After definition of the vectors

$$(2) \quad \begin{aligned} \mathbf{c} &\in \mathbf{R}^{m+1}, & \mathbf{c}_k &= g(\mathbf{x}^k), \\ \mathbf{e} &\in \mathbf{R}^m, & \mathbf{e}_k &= 1, \\ \mathbf{y} &\in \mathbf{R}^{n+1}, & \mathbf{y} &= \begin{bmatrix} \mathbf{z} \\ \alpha \end{bmatrix}, \end{aligned}$$

and the matrices

$$(3) \quad \begin{aligned} X &\in \mathbf{R}^{n \times m}, & X_k &= \mathbf{x}^k - \mathbf{x}^m, \\ A &\in \mathbf{R}^{(m+1) \times (n+1)}, & A &= \begin{bmatrix} X^T & \mathbf{e} \\ 0^T & 1 \end{bmatrix}, \end{aligned}$$

the condition is equivalent to requiring that $\bar{\mathbf{y}} = \begin{bmatrix} \bar{\mathbf{z}} \\ \bar{\alpha} \end{bmatrix}$ minimize $\|A\bar{\mathbf{y}} - \mathbf{c}\|$. The only p for which the solution to this minimization problem with a norm $\|\cdot\| \equiv \|\cdot\|_p$ depends linearly on the data \mathbf{c} is $p = 2$. Therefore from here on we assume $\|\cdot\| \equiv \|\cdot\|_2$. Conditions (i-iii) now give the following definition of a gradient estimate.

DEFINITION 6. $\bar{\mathbf{z}}$ is the best affine estimate of $\partial g(\mathbf{x})^T$ at \mathbf{x} with respect to the points \mathbf{x}^k and function values $g(\mathbf{x}^k)$ iff there exists a scalar $\bar{\alpha}$ such that the vector $\bar{\mathbf{y}} = \begin{bmatrix} \bar{\mathbf{z}} \\ \bar{\alpha} \end{bmatrix}$ is the unique vector that minimizes $\|A\bar{\mathbf{y}} - \mathbf{c}\|$, where A and \mathbf{c} are defined in (2) and (3).

Before estimates using a minimum number of function evaluations may be constructed, necessary and sufficient conditions for their existence are needed. There is always at least one vector that minimizes $\|A\bar{\mathbf{y}} - \mathbf{c}\|$; estimates are distinguished by the requirement that they be the unique minimizer. This uniqueness criterion, together with the form of A , immediately gives the following results on existence of estimates.

PROPOSITION 1. A gradient $\partial g(\mathbf{x})$ is m -estimable from the $m + 1$ function values $g(\mathbf{x}^k)$ iff $m \geq n$ and the matrix X of (3) has full rank.

Definition 6 does not require that $\mathbf{x} \in \{\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m\}$. If it is then we shall assume $\mathbf{x} = \mathbf{x}^m$ and the gradient approximation so formed corresponds to a one-sided difference

approximation to a derivative. If \mathbf{x} is in the interior of the convex hull of $\{\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m\}$ then the estimate corresponds to a centered difference.

4.2. Estimates using a priori knowledge. The estimate of Definition 6 used no knowledge of $g(\mathbf{x})$ apart from the function values $g(\mathbf{x}^k)$. In the presence of a priori knowledge of $g(\mathbf{x})$, such as the zero-nonzero structure of $\partial g(\mathbf{x})$, we change the definition to take advantage of such knowledge, giving an estimate that is the solution of a constrained least squares minimization.

The knowledge that p elements of $\partial g(\mathbf{x})$ are zero is a special case of a priori knowledge that $\partial g(\mathbf{x})$ satisfies p linearly independent constraints, i.e.

$$(4) \quad W \partial g^T = d, \quad W \in \mathbf{R}^{p \times n}, \quad d \in \mathbf{R}^p.$$

With such predetermined conditions on $\partial g(\mathbf{x})$ the definition of a gradient estimate becomes:

DEFINITION 7. $\bar{\mathbf{z}}$ is the best affine estimate for $\partial g(\mathbf{x})^T$ at \mathbf{x} with respect to the points \mathbf{x}^k , function values $g(\mathbf{x}^k)$ and constraints of (4) iff there exists a scalar $\bar{\alpha}$ such that $\bar{\mathbf{z}}$ satisfies $W\bar{\mathbf{z}} = d$ and the vector $\bar{\mathbf{y}} = [\bar{\alpha}]$ is the unique vector that minimizes $\|A\mathbf{y} - \mathbf{c}\|$.

This definition implicitly rates a priori knowledge higher than knowledge of function values to give a constrained linear least squares problem. The next proposition gives necessary and sufficient conditions for the existence of a unique solution to such a problem, deduced from the solution characterization given in [1]; the proof is left to the reader.

PROPOSITION 2. A gradient $\partial g(\mathbf{x})$ satisfying (4) is m -estimable from the $m + 1$ function values $g(\mathbf{x}^k)$ iff $m \geq n - p$ and the matrix $[X^W]$ has full rank.

In the special case of sparsity Proposition 2 reduces to

PROPOSITION 3. If $\partial_j g(\mathbf{x}) = 0, j \in C, |C| = p, \partial g(\mathbf{x})$ is m -estimable from the $m + 1$ function values $g(\mathbf{x}^k)$ iff $m \geq n - p$ and the matrix \bar{X}^T formed by deletion of the p columns $X_j^T, j \in C$ from X^T has full rank.

In gradient estimation at a point \mathbf{x} $g(\mathbf{x})$ is usually already available so for efficiency \mathbf{x} is included among the points \mathbf{x}^k , i.e. $\mathbf{x} \equiv \mathbf{x}^m$. Since the gradient is to be estimated at \mathbf{x} , the affine approximation is usually required to be exact at \mathbf{x} , i.e. $\bar{\alpha} \equiv g(\mathbf{x}^m) \equiv g(\mathbf{x})$. Under these restrictions the minimizations in the above definitions reduce to minimization of $\|X^T \mathbf{z} - \mathbf{b}\|$ where $\mathbf{b}_k = g(\mathbf{x}^k) - g(\mathbf{x})$. However the conditions under which these restricted estimates exist are still given by Propositions 1–3.

4.3. A Jacobian estimate using a priori knowledge. We now generate a definition of an estimate of a sparse Jacobian $J(\mathbf{x})$ from the above definition of a gradient estimate by use of the following construction principle.

Principle. \bar{J} is an estimate of $J(\mathbf{x})$ iff the i th row of \bar{J} is an estimate of $\partial \mathbf{f}_i(\mathbf{x})$.

DEFINITION 8. The matrix \bar{J} is the best affine estimate of $J(\mathbf{x})$ at \mathbf{x} with respect to the $m + 1$ points \mathbf{x}^k and function values $\mathbf{f}(\mathbf{x}^k)$ iff there exists a vector $\bar{\alpha} \in \mathbf{R}^n$ such that for each $j \in \mathbf{Z}_n, \bar{J}_j^T$ has the same sparsity as $J_j^T(\mathbf{x})$ and the vector $\bar{Y}_j = [\bar{\alpha}_j^T]$ is the unique vector that minimizes $\|A Y_j - C_j^T\|$, where

$$Y \in \mathbf{R}^{(m+1) \times (n+1)}, \quad C \in \mathbf{R}^{n \times (m+1)}, \quad C_k = \mathbf{f}(\mathbf{x}^k).$$

If the restrictions outlined in the last paragraph of § 4.2 are applied to each component function, Definition 8 takes on the form:

DEFINITION 9. \bar{J} is the best affine estimate if $J(\mathbf{x})$ at \mathbf{x} with respect to the $m + 1$ points \mathbf{x}^k and function values $\mathbf{f}(\mathbf{x}^k)$ iff \bar{J} is the unique matrix having the same sparsity as $J(\mathbf{x})$ that minimizes $\|X^T J^T - B^T\|_F$, where B is the matrix of Definition 3.

The reduction to the Frobenius norm is made possible by the independence of the least squares problems of estimating each row of \bar{J} .

Proposition 3 may be expanded using the construction principle to provide necessary and sufficient conditions for the existence of the estimates of Definition 8.

PROPOSITION 4. $J(\mathbf{x})$ is m -estimable from the $m + 1$ function values $\mathbf{f}(\mathbf{x}^k)$ iff $m \geq \max_{j \in \mathbf{Z}_n} (n - p_j)$, where p_j is the number of zero elements in the j th row of $J(\mathbf{x})$, and every matrix \bar{X}^T formed by deletion of columns in X^T corresponding to zeros in the j th row of $J(\mathbf{x})$ has full rank.

We now prove the main result of this section; that it is always possible to construct a matrix X so that a Jacobian with given sparsity may be estimated in the minimum number of function evaluations required by Proposition 4.

THEOREM 3. $J(\mathbf{x})$ is m -estimable iff $m \geq \max_{j \in \mathbf{Z}_n} (n - p_j)$.

Proof. For the proof it suffices to construct a matrix X^T of size $m \times n$ such that deletion of any $P = \min_{j \in \mathbf{Z}_n} p_j$ or more columns leaves a submatrix \bar{X}^T with full rank. One such matrix is the rectangular Vandermonde matrix

$$X_{ij}^T = \lambda_j^i \quad i \in \mathbf{Z}_m, \quad j \in \mathbf{Z}_n$$

where $\lambda_j \neq \lambda_k$ if $j \neq k$. Suppose P or more columns are deleted, leaving only columns X_{j_k} , $k \in \mathbf{Z}_K$, $K \leq n - P = m$ forming the $m \times K$ matrix \bar{X}^T . Then the upper $K \times K$ submatrix $\bar{X}^{T'}$ of \bar{X}^T is a square Vandermonde matrix

$$\bar{X}_{ik}^{T'} = \lambda_{j_k}^i \quad i, k \in \mathbf{Z}_K$$

with

$$\det(\bar{X}^{T'}) = \prod_{1 \leq i < k \leq K} (\lambda_{j_k} - \lambda_{j_i}) \neq 0.$$

Therefore $\bar{X}^{T'}$ is nonsingular and \bar{X}^T has full rank.

4.4. The choice of matrix X . Theorem 3 gives a minimum size for the linear equations that must be solved in efficient Jacobian estimation. It remains to choose the matrix X so that these systems may be solved with minimum effort and maximum accuracy. Two issues that arise are scaling and conditioning.

In the examples above X involved a small parameter h . It is convenient to separate the small parameter from the direction of the differences. This is done by writing X as $D_1 X_1 D_2$ where the D_i are diagonal matrices and X_1 is normalized so that each new row has Euclidean norm 1. Then D_{1kk} represents scaling done on each variable \mathbf{x}_k and D_{2kk} represents scaling done on each directional difference X_{1k} , e.g. $D_{2kk} = h \|X_{1k}\|^{-1}$. Further scaling on B may be necessary, but this does not affect X . From here on we shall not distinguish between X and X_1 .

The Vandermonde matrices used in the proof of Theorem 3 are notoriously ill-conditioned. However we know of no class of matrices with provably good condition numbers and the requisite rank properties that might serve in their place¹. The next proposition indicates that for a fixed choice of λ_j an improvement in conditioning, as measured by the determinant, is gained with

$$(5) \quad X_{ij}^T = c_j T_i(\lambda_j)$$

where T_i is a i th Chebyshev polynomial and c_j is the appropriate normalizing constant.

¹ When \mathbf{f} is complex analytic some improvement results with the use of the roots of unity as elements of the Vandermonde matrix.

PROPOSITION 5. If $\lambda_j \in [-1, 1]$, $j \in \mathbf{Z}_p$ and

$$W_{ij} = c_j T_i(\lambda_j), \quad X_{ij} = d_j \lambda_j^i, \quad i, j \in \mathbf{Z}_p$$

where c_j and d_j are chosen so that

$$\sum_{i \in \mathbf{Z}_p} W_{ij}^2 = \sum_{i \in \mathbf{Z}_p} X_{ij}^2 = 1, \quad j \in \mathbf{Z}_p$$

then $\det(W)$ and $\det(X)$ satisfy

$$(6) \quad \begin{aligned} 1 &\geq |\det(W)| \geq |\det(X)|, \\ 2(2^{p-3}p)^{p/2} |\det(X)| &\geq |\det(W)| \geq 2\left(\frac{2^{p-3}}{p}\right)^{p/2} |\det(X)|. \end{aligned}$$

Proof. By Hadamard's inequality for determinants

$$\det^2(W) \leq \prod_{j \in \mathbf{Z}_p} \left(\sum_{i \in \mathbf{Z}_p} W_{ij}^2 \right) \leq 1;$$

likewise $|\det(X)| \leq 1$. Since determinants are homogeneous of degree 1 under multiplication of rows or columns by scalars and invariant under addition of rows to other rows

$$\det(W) = \det(c_j T_i(\lambda_j)) = \left(\prod_{j \in \mathbf{Z}_p} c_j \right) \det\left(2^{\alpha_i} \lambda_j^i + \sum_{k=1}^i \beta_{ik} \lambda_j^{i-k}\right)$$

where $\alpha_0 = 0$ and $\alpha_i = i - 1$ for $i \geq 1$

$$(7) \quad \begin{aligned} &= \left(\prod_{i \in \mathbf{Z}_p} c_i 2^{\alpha_i} \right) \det(\lambda_j^i) \\ &\Rightarrow \det(W) = \left(\prod_{i \in \mathbf{Z}_p} \frac{c_i 2^{\alpha_i}}{d_i} \right) \det(X). \end{aligned}$$

Since $|T_i(\lambda)| \leq 1$ for $\lambda \in [-1, 1]$, and $\lambda_j^0 = 1$,

$$\begin{aligned} c_j^2 &= \left(\sum_{i \in \mathbf{Z}_p} T_i^2(\lambda_j) \right)^{-1} \geq p^{-1}, \\ d_j^2 &= \left(\sum_{i \in \mathbf{Z}_p} \lambda_j^{2i} \right)^{-1} \leq 1 \\ &\Rightarrow |\det(W)| \geq 2 \left[\frac{2^{p-3}}{p} \right]^{p/2} |\det(X)|. \end{aligned}$$

The upper bound follows similarly from (7), since $T_0(\lambda) \equiv 1$ implies $c_j \leq 1$ and $\lambda \in [-1, 1]$ implies $|\lambda_j^i| \leq 1$ so that $d_j \geq p^{-1/2}$.

The bounds of Proposition 5 show that W is better conditioned than X , but also that W may still be quite ill-conditioned. For instance if the p points λ_i are uniformly distributed over $[-1, 1]$,

$$\det(X) = \prod_{i=1}^{p-1} \prod_{j=0}^{i-1} (\lambda_i - \lambda_j) = \prod_{i=1}^{p-1} \prod_{j=1}^i \frac{2j}{p-1} = \left(\frac{2}{p-1}\right)^{p(p-1)/2} \prod_{i=1}^{p-1} i!$$

Since

$$\sum_{i=2}^{p-1} \ln i \leq \int_1^p \ln x \, dx = p \ln p - p + 1,$$

we may show that

$$\det (X) \leq e^{\alpha p^2 + O(p \ln p)} \alpha = \frac{\ln 2}{2} - \frac{3}{4} \cong -.4,$$

$$\det (W) \leq e^{\beta p^2 + O(p \ln p)} \beta = \ln 2 - \frac{3}{4} \cong -.05.$$

Because β is less than zero $\det (W)$ will rapidly approach zero as p grows large.

Nevertheless numerical experiments suggest this more general Jacobian estimation may have practical value. As a test, five 100×100 random matrices were generated, each with approximately 90% of its elements zero. For each matrix the maximum number, m , of nonzero elements in any row was found and the $m \times 100$ matrix $X_{ij}^T = c_j \cos ((2j + 1)i\pi/2n)$ of (5) was formed. Trials indicated that this choice of the points λ_j performed better than an even distribution. Finally the condition numbers of each of the one hundred $m \times (100 - p_i)$ submatrices were calculated using singular value decomposition. Table 1 gives the distribution of these condition numbers.

TABLE 1

	Number of submatrices in each interval			
Condition number	1-10 ³	10 ³ -10 ⁵	10 ⁵ -10 ⁷	10 ⁷ -10 ¹⁰
Number of instances	471	25	3	1

The results show that almost all submatrices are well conditioned. The few large condition numbers correspond to rows with a large number of adjacent or nearly adjacent nonzero elements; such condition numbers can be reduced by simply permuting the rows of X so that in each \bar{X}^T , the differences $|\lambda_{jk} - \lambda_{jl}|, k \neq l$ are reasonably large.

Some reduction in the cost of using the more general Jacobian estimation is possible by noting that the problem of determining the points λ_j may be reduced from a specification of n variables to one of k variables, if the isolation graph is k -colorable.

PROPOSITION 6. *Given a coloring of the variable isolation graph of $J(\mathbf{x})$, rows in X corresponding to variables with the same color may be identical.*

Proof. It suffices to show that two such columns X_i^T and X_j^T corresponding to variables \mathbf{x}_i and \mathbf{x}_j with the same color never appear in the same submatrix \bar{X}^T . If they did, then for some column $\bar{J}_k^T, \bar{J}_{ik}^T$ and \bar{J}_{jk}^T (therefore $J_{ik}^T(\mathbf{x})$ and $J_{jk}^T(\mathbf{x})$) would be nonzero. But this implies columns i and j of $J(\mathbf{x})$ are not isolated, a contradiction.

Proposition 6 implies that we need only construct an $m \times k$ matrix X^T . If k and m differ only slightly then a computationally efficient and tolerably ill-conditioned choice for X^T would be the $m \times m$ identity matrix with $k - m$ appended columns of the form $c_j \lambda_j^i$. For any submatrix \bar{X}^T of size $m \times m$ formed from X^T expansion of $\det (\bar{X}^T)$ about columns of the identity gives

$$\det (\bar{X}^T) = \det \begin{vmatrix} \lambda_{j_0}^{i_0} & \cdots & \lambda_{j_p}^{i_0} \\ \vdots & & \vdots \\ \lambda_{j_0}^{i_{p-1}} & \cdots & \lambda_{j_p}^{i_{p-1}} \end{vmatrix}$$

where $p \leq k - m$. Remes [7] has shown that if $i_k \neq i_l$ when $k \neq l$ then $\{e^{i_0 x}, e^{i_1 x}, \dots, e^{i_{p-1} x}\}$ satisfies the Haar condition on any finite interval. Therefore the transformation $\lambda_{jk} = e^{x_{jk}}$ establishes that if $\lambda_{jk} \in (0, 1]$ then $\det (\bar{X}^T) \neq 0$ and X satisfies Proposition 4.

5. Concluding remarks. We have attempted to provide a clear idea of efficient Jacobian estimation by careful definition of admissible estimates. In doing so we have shown that the finding of most efficient estimates may be reduced to well studied problems in graph theory or to constrained least squares problems. Furthermore the approaches outlined here for Jacobians whose sole known property is their zero-nonzero structure may be modified to produce characterizations of efficient estimates for Jacobians with other constraints, such as symmetry as in [6].

The best estimate of the two estimates presented above and the third described in the Appendix is obviously problem dependent, but some general comments may be made. The computation of Jacobian estimates in an iterative solution to (1) divides into fixed and repeated costs. Fixed costs include the graph colorings used by all three estimates; they are computations that may be done independently of the values at any iteration. A further fixed cost for the estimates of § 4 would be the QR factorizations of matrices appearing in the linear least squares problems set up there. Repeated costs are the computations necessary at each iteration and consist almost exclusively of function evaluations. For CPR estimates and the EI estimates defined in the Appendix the remaining repeated costs are trivial, consisting only of reading off $J(\mathbf{x})$ from B . Additional costs for the other estimate consist of at most $2n m \times m$ matrix multiplications, assuming the QR factorizations are already available, so are also comparatively small.

For systems requiring a large number of iterations to solution fixed costs are negligible. If a well conditioned matrix X can be obtained either from (5) or by the method outlined after Proposition 6, then the estimates of § 4 appear to be the best choice for such systems. Otherwise EI estimates would be preferable to CPR estimates, although the fixed cost of an EI estimate is the cost of an approximate solution of a much larger graph coloring problem than for a CPR estimate.

In a report that came to the authors' attention after this paper was submitted for publication, Coleman and Moré [2] present a detailed study of CPR estimation. Using a different proof, they demonstrate the equivalence of CPR estimation and graph coloring, give some fast algorithms for generating approximate minimum colorings and consider their performance on a wide range of problems. In the data appearing there, only for randomly generated matrices does the number k of function evaluations differ from m by more than 50%, the usual difference being considerably less. These results suggest that for many practical problems the most efficient estimation algorithm would be construction of a k -CPR-estimation for a minimal k with the option of using this to construct an estimate of § 4, based on the matrix X described after Proposition 6.

6. Appendix. The element isolation principle and Jacobian estimation. Efficient CPR estimates of the Jacobian are constructed column by column. Two or more columns are estimated from one difference, i.e. one column of the matrix B of Definition 3, if and only if the columns correspond to isolated variables. In this section the construction of estimates element by element is considered. With an extended concept of isolation we show that two or more elements may be estimated from one difference iff they are isolated. We link element isolation with graph theory to determine the complexity of construction of estimations requiring a minimum number of function evaluations. The discussion is terse as the analysis and results are very similar to those on efficient CPR estimates.

DEFINITION 10. J_{ij} is isolated from J_{pq} iff

$$J_{iq} = J_{pj} = 0 \quad \text{or} \quad j = q.$$

The corresponding principle for efficient Jacobian estimation is the following.

DEFINITION 11. *The EI (Element Isolation) principle.* Let C_k be a set of double subscripts. The $|C_k|$ elements $J_{ij}(\mathbf{x})$, $ij \in C_k$, may be estimated in one difference

$$\mathbf{f}(\mathbf{x} + X_k) - \mathbf{f}(\mathbf{x}) \text{ where } X_k = h \sum_{i \exists j \ ij \in C_k} I_i$$

if $\forall ij, pq \in C_k J_{ij}$ is isolated from J_{pq} . Concepts of m -EI-estimability and the element isolation graph are defined analogous to those of m -CPR-estimability and the variable isolation graph.

It is obvious that any coloring of the variable isolation graph induces a coloring of the element isolation graph, so estimation of a Jacobian by the EI principle requires at most as many function evaluations as estimation by the CPR principle. The following example, due to Eisenstadt [2], shows that a minimum EI estimate may use strictly fewer function evaluations:

$$(8) \quad J(\mathbf{x}) = \begin{bmatrix} \times & 0 & 0 & \times & 0 & 0 \\ 0 & \times & 0 & 0 & \times & 0 \\ 0 & 0 & \times & 0 & 0 & \times \\ \times & \times & \times & 0 & 0 & 0 \\ \times & 0 & 0 & 0 & \times & \times \\ 0 & \times & 0 & \times & 0 & \times \\ 0 & 0 & \times & \times & \times & 0 \end{bmatrix}.$$

The variable isolation graph of the Jacobian in (8) is the complete graph on 6 vertices so 7 function evaluations are needed for a minimum CPR estimate. This generalization of CPR estimation allows estimates of nonzero elements in one column of the Jacobian to be obtained from more than one difference. The choice of sets C_k

$$\begin{aligned} C_0 &= \{00, 11, 22\}, \\ C_1 &= \{03, 14, 25\}, \\ C_2 &= \{30, 40, 53, 63\}, \\ C_3 &= \{31, 51, 44, 64\}, \\ C_4 &= \{32, 62, 45, 55\}, \end{aligned} \quad X = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

shows that an EI estimate is possible with only 6 function evaluations.

We now show that deciding if a Jacobian is m -EI-estimable is still an NP-complete problem.

THEOREM 4. *The element isolation graph is m -colorable iff $J(\mathbf{x})$ is m -EI-estimable.*

Proof. The proof is that of Theorem 1 with the words “element” and “ m -EI-estimable” replacing “variable” and “ m -CPR-estimable”.

A proof of the direct converse of Theorem 4, that every graph is the element isolation graph of some Jacobian, is not possible. The graph in Fig. 1 is a counter example; enumeration of all possible Jacobians with only four nonzero elements shows that it cannot be an element isolation graph. Therefore we construct a graph G_m that is a polynomial extension of G such that G_m is m -colorable iff G is m -colorable and show that G_m is the element isolation of a particular Jacobian $J(\mathbf{x})$.

DEFINITION 12. Given a graph $G = (V, E)$, the expansion graph $G_m(\vec{V}, \vec{E})$ of G

is defined by

$$\begin{aligned}
 S_i &= \{j \mid (v_i, v_j) \in E\}, \\
 \bar{V} &= \bar{V}_1 \cup \bar{V}_2 \cup \bar{V}_3, \\
 \bar{V}_1 &= \{\bar{v}_{ij} \mid i \in \mathbf{Z}_{|V|}, j \in S_i\}, \\
 \bar{V}_2 &= \{v_{i|V|} \mid i \in \mathbf{Z}_{|V|}\}, \\
 \bar{V}_3 &= \{\bar{v}_{i(k+|V|+1)} \mid i \in \mathbf{Z}_{|V|}, k \in \mathbf{Z}_{m-1}\}, \\
 \bar{E} &= \bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_3, \\
 \bar{E}_1 &= \{(\bar{v}_{ij}, \bar{v}_{pq}) \mid j = p \in S_i \text{ or } q = i \in S_p\}, \\
 \bar{E}_2 &= \{(\bar{v}_{ij}, \bar{v}_{i(k+|V|+1)}) \mid j \in S_i \cup \{|V|\}, k \in \mathbf{Z}_{m-1}\}, \\
 \bar{E}_3 &= \{(\bar{v}_{i(j+|V|+1)}, \bar{v}_{i(k+|V|+1)}) \mid j \neq k \text{ and } j, k \in \mathbf{Z}_{m-1}\}.
 \end{aligned}$$

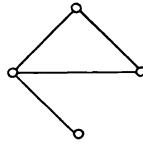


FIG. 1

Fig. 2 shows the expansion G_3 of the graph G in Fig. 1. In words, each vertex $v_i \in V$ is replaced by a cluster of vertices $\bar{v}_{ij} \in \bar{V}_1 \cup \bar{V}_2$. Each edge $(v_i, v_p) \in E$ is replaced by a collection of edges $(\bar{v}_{ij}, \bar{v}_{pq}) \in \bar{E}_1$ such that every vertex in the i th cluster is connected to a common vertex in the p th cluster and one vertex in the i th cluster is connected to every vertex in the p th cluster. \bar{V}_3 and \bar{E}_3 indicate that associated with each cluster is a clique of size $m - 1$, \bar{E}_2 connects every vertex in the cluster to every vertex in this clique and \bar{V}_2 contains a special vertex in the cluster that serves as a link between cluster and clique.

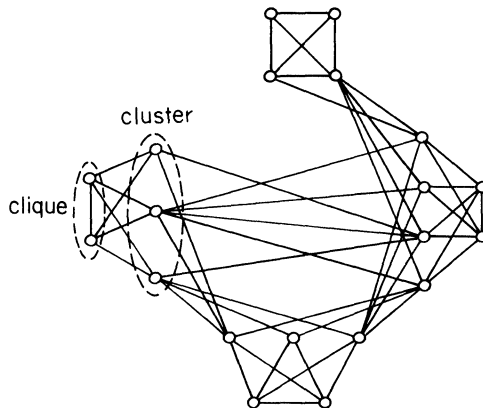


FIG. 2

THEOREM 5. *A graph G is m -colorable iff the expansion graph G_m is m -colorable.*

Proof. If G has an m -coloring then assigning to each vertex in the i th cluster the color of v_i and coloring the associated clique in the remaining $m - 1$ colors gives an

- [2] T. F. COLEMAN AND J. J. MORÉ, *Estimation of sparse Jacobian matrices and graph coloring problems*, Report ANL-81-39, Argonne National Laboratory, Argonne, IL, 1981.
- [3] A. R. CURTIS, M. J. D. POWELL AND J. K. REID, *On the estimation of sparse Jacobian matrices*, J. Inst. Math. Appl., 13 (1974), pp. 117–119.
- [4] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.
- [5] 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. 85–103.
- [6] M. J. D. POWELL AND PH. L. TOINT, *On the estimation of sparse Hessian matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 1060–1074.
- [7] E. YA. REMES, *General Computational Methods of Tchebychev Approximation*, AECT 4491, Kiev (In Russian). (See also p. 78 of E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.)

ERDŐS-KO-RADO THEOREM—22 YEARS LATER*

M. DEZA† AND P. FRANKL†

Abstract. In 1961 Erdős, Ko, and Rado proved that, if a family \mathcal{F} of k -subsets of an n -set is such that any 2 sets have at least l elements in common, then for n large enough $|\mathcal{F}| \leq \binom{n-l}{k-l}$. This result had great impact on combinatorics. Here we give a survey of known and of some new generalizations and analogues of this theorem. We consider mostly problems which were not included or were touched very briefly in the survey papers [17], [46], [51], [61].

1. The Erdős-Ko-Rado theorem. Let $X = \{x_1, \dots, x_n\}$ be a finite set of cardinality $|X| = n$. 2^X denotes the power set of X , while $\binom{X}{k}$ stands for the set of all k -subsets of X .

THEOREM 1.1 (Erdős, Ko, Rado [15]). *Let n, k, l be integers, with $n > k > l > 0$ and suppose \mathcal{F} is a family of k -subsets of X , i.e. $\mathcal{F} \subseteq \binom{X}{k}$. Suppose further that any two members of \mathcal{F} intersect in at least l elements. Then, for $n > n_0(k, l)$,*

- a) $|\mathcal{F}| \leq \binom{n-l}{k-l}$,
- b) $|\mathcal{F}| = \binom{n-l}{k-l}$ iff for some $X_0 \in \binom{X}{l}$ we have $\mathcal{F} = \{F \in \binom{X}{k} : X_0 \subset F\}$.

Remark 1.2. Actually, in [15] the theorem was formulated for antichains with $|F| \leq k$, i.e. $\mathcal{F} \subseteq (\binom{X}{l} \cup \dots \cup \binom{X}{k})$ and there are no $F, F' \in \mathcal{F}$ such that $F \subset F'$ holds. However, this version is an easy consequence of Theorem 1.1.

We say that a family, $\mathcal{F} \subseteq 2^X$ is l -intersecting if for any $F, F' \in \mathcal{F}$, $|F \cap F'| \geq l$ holds. We define now a special class of l -intersecting families.

Let X_i be a subset of X of cardinality $l + 2i$. Define $\mathcal{F}_i = \{F \in \binom{X}{k} : |F \cap X_i| \geq l + i\}$. It is easy to see that for $F, F' \in \mathcal{F}_i$ one has $|F \cap F'| \geq l$, i.e. \mathcal{F}_i is l -intersecting. Note that

$$|\mathcal{F}_0| = \binom{n-l}{k-l},$$

$$|\mathcal{F}_0| = \max_{0 \leq i \leq k-l} |\mathcal{F}_i| \text{ iff } n \geq (k-l+1)(l+1).$$

Thus Theorem 1.1 does not hold for $n < (k-l+1)(l+1)$.

THEOREM 1.3 [27]. *Suppose $l \geq 15$, $\mathcal{F} \subseteq \binom{X}{k}$ and \mathcal{F} is l -intersecting.*

- a) *If $n > (k-l+1)(l+1)$, then*
 $|\mathcal{F}| \leq \binom{n-l}{k-l}$, and equality holds iff \mathcal{F} is of the form \mathcal{F}_0 .
- b) *If $n = (k-l+1)(l+1)$, then*
 $|\mathcal{F}| \leq \binom{n-l}{k-l}$, and equality holds iff \mathcal{F} is of the form \mathcal{F}_0 or \mathcal{F}_1 .
- c) *There exists an absolute constant $c, c < 1$, such that for*

$$c(k-l+1)(l+1) \leq n < (k-l+1)(l+1)$$

we have

$$|\mathcal{F}| \leq (l+2)\binom{n-l-2}{k-l-1} + \binom{n-l-2}{k-l-2}, \text{ and equality holds iff } \mathcal{F} \text{ is of the form } \mathcal{F}_1.$$

Remark 1.4. In the case $l = 1$, Erdős, Ko, and Rado proved that $|\mathcal{F}| \leq \binom{n-1}{k-1}$ iff $n \geq 2k$. For $2 \leq l \leq 14$, in [27] $|\mathcal{F}| \leq \binom{n-l}{k-l}$ is established for $n \geq 2(k-l+1)(l+1)$. The original bound of Erdős, Ko, and Rado was $n \geq l + (k-l)\binom{l}{1}^3$. Hsieh [44] improved this to $n \geq l + (k-l+1)(l+1)(k-l)$. In the case $l = 1, n = 2k$ each maximal (i.e. non-extendable) family has maximum size $\binom{2k-1}{k-1}$.

* Received by the editors January 29, 1982, and in revised form July 15, 1982.

† Centre National de la Recherche Scientifique, 15 Quai Anatole France, 75007 Paris, France.

The general case would be settled by the following

Conjecture 1.5 [27]. Suppose $\mathcal{F} \subset \binom{X}{k}$ and \mathcal{F} is l -intersecting. Then

$$|\mathcal{F}| \leq \max_{0 \leq i \leq k-l} |\mathcal{F}_i|.$$

The case $l = 2, n = 2k, k$ even of the above conjecture was already conjectured by Erdős, Ko, and Rado.

The original proof of Theorem 1.1 and that of Theorem 1.3 use the so-called exchange operation.

DEFINITION 1.6. Let \mathcal{F} be a family of subsets of the ordered set $X = \{x_1, \dots, x_n\}, 1 \leq i < j \leq n$. The exchange operation $E_{i,j}$ is defined by

$$E_{i,j}(F) = \begin{cases} \{x_i\} \cup F - \{x_j\} & \text{if } x_i \notin F, x_j \in F, (\{x_i\} \cup F - \{x_j\}) \notin \mathcal{F}, \\ F & \text{otherwise,} \end{cases}$$

$$E_{i,j}(\mathcal{F}) = \{E_{i,j}(F) : F \in \mathcal{F}\}.$$

This operation is called also compression or pushing. The main importance of this operation lies in the following easy lemma.

PROPOSITION 1.7. If $\mathcal{F} \subset 2^X$ is l -intersecting, then $E_{i,j}(\mathcal{F})$ is l -intersecting as well.

Iterating the exchange operation $E_{i,j}$ for every pair $1 \leq i < j \leq n$, at last we obtain a family \mathcal{F}^* which is stable i.e. $E_{i,j}(\mathcal{F}^*) = \mathcal{F}^*$ for every $1 \leq i < j \leq n$.

For the case $l = 1$ of Theorem 1.1, Katona [45] gave a nice, simple proof, Daykin [6] showed that this case is a consequence of the Kruskal–Katona theorem (cf. [52], [48], [5]).

Remark 1.8. Let us define the graph $G(n, k, l)$ whose vertex set is $\binom{X}{k}$ and in which (F, F') is an edge iff $|F \cap F'| < l$. Then Theorem 1.1 gives that for $n > n_0(k, l)$ the independence number $\alpha(G(n, k, l)) = \binom{n-l}{k-l}$. Using the linear programming bound for association schemes of Delsarte [7], in [59] Schrijver strengthened the Erdős–Ko–Rado theorem by proving that the capacity (cf. Shannon [60]) of $G(n, k, l)$ is equal to $\binom{n-l}{k-l}$ for $n > n'(k, l)$. Brouwer and Frankl [3] showed that it holds for $n > k^2/2$. It would be interesting to know whether the same holds already for $n \geq (k-l+1)(l+1)$.

There is an exciting conjecture of Chvátal that is connected with the case $l = 1$. To state it we need a definition.

A family \mathcal{F} is said to be a *simplicial complex* (ideal, hereditary system, down-set) if $G \subset F \in \mathcal{F}$ implies $G \in \mathcal{F}$.

For $x \in X$ let $\mathcal{F}(x)$ denote $\{F \in \mathcal{F} : x \in F\}$.

Conjecture 1.9 (Chvátal [4]). Suppose \mathcal{F} is a simplicial complex, $\mathcal{G} \subset \mathcal{F} \subset 2^X$, and any two elements of \mathcal{G} have nonempty intersection. Then $|\mathcal{G}| \leq \max_{x \in X} |\mathcal{F}(x)|$.

Remark 1.10. Chvátal [4] proved that the conjecture is true for stable families (with respect to the exchange operation). He also showed that the corresponding statement fails for $l > 1$. Kleitman [51] proposes an interesting new approach to this conjecture.

2. Close relatives: stability of the extremal families, degree conditions, nonuniform case. One of the most natural questions to ask with respect to Theorem 1.1 is: what happens if we exclude the optimal families, i.e. if we suppose there is no l -set contained in all the members of the family. This problem was solved for the case $|F \cap F'| \geq 1$ i.e. $l = 1$ by Hilton and Milner [43] and for the general case in [25]:

THEOREM 2.1. Suppose $\mathcal{F} \subset \binom{X}{k}$, \mathcal{F} is l -intersecting, but $|\bigcap_{F \in \mathcal{F}} F| < l$; moreover $|\mathcal{F}|$ is maximal with respect to these constraints. Then for $n > n_0(k, l)$ ($n_0(k, 1) = 2k$):

a) $k > 2l + 1$ or $k = 3, l = 1$. There exist $D_1, D_2 \subset X, D_1 \cap D_2 = \emptyset, |D_1| = l, |D_2| =$

$k - l + 1$ such that

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : D_1 \subset F, F \cap D_2 \neq \emptyset \right\} \cup \left\{ F \in \binom{X}{k} ; D_2 \subseteq F, |F \cap D_1| \geq l - 1 \right\}.$$

b) $k \leq 2l + 1$. There is a $D \in \binom{X}{l+2}$ such that

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : |F \cap D| \geq l + 1 \right\}.$$

Another natural question was asked by Erdős, Rothschild, and Szemerédi (cf. Erdős [16]). Let c be an absolute constant, $0 < c < 1$, and suppose along with the conditions of Theorem 1.1, that every element $x \in X$ is contained in at most $c|\mathcal{F}|$ sets, i.e. the *degree* of every point is $\leq c|\mathcal{F}|$. What is the maximum of $|\mathcal{F}|$ and which are the extremal systems as functions of c ? This question was completely answered for several choices of c ($c \geq \frac{2}{3}$ or $\frac{1}{2} > c \geq \frac{3}{7}$ in [25], $\frac{1}{2} \leq c < \frac{2}{3}$ by Füredi [35]):

THEOREM 2.2. *Suppose $\mathcal{F} \subset \binom{X}{k}$, $F \cap F' \neq \emptyset$ and for every $x \in X$, $|\mathcal{F}(x)| \leq c|\mathcal{F}|$. Moreover, suppose $|\mathcal{F}|$ is maximal subject to these constraints. Then for $n > n_0(k, c)$, $c > \frac{3}{7}$, one of the following possibilities holds:*

a) $\frac{2}{3} < c < 1$. There exists $D \in \binom{X}{3}$ such that

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : |F \cap D| \geq 2 \right\}.$$

b) $c = \frac{2}{3}$. There exists $D \in \binom{X}{3}$ such that

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : |F \cap D| = 2 \right\}.$$

c) $\frac{3}{5} < c < \frac{2}{3}$. There exists an $\mathcal{A} \subset \binom{X}{3}$, $|\mathcal{A}| = 10$, $A \cap A' \neq \emptyset$ for all $A, A' \in \mathcal{A}$, and $|\cup_{A \in \mathcal{A}} A| \leq 6$ (there are 6 nonisomorph possibilities), such that

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : \exists A \in \mathcal{A}, A \subseteq F \right\}.$$

d) $\frac{1}{2} < c \leq \frac{3}{5}$. Let \mathcal{A}_0 be the family among those in c) which has $|X_0 = \cup A| = 6$ and every point of X_0 has degree 5.

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : \exists A \in \mathcal{A}_0, A \subseteq F \right\}.$$

e) $c = \frac{1}{2}$. The family \mathcal{A}_0 and X_0 is as in d),

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : (F \cap X) \in \mathcal{A}_0 \right\}.$$

f) $\frac{3}{7} < c < \frac{1}{2}$. There exists $V \in \binom{X}{7}$, $\mathcal{P} \subset \binom{V}{3}$ such that $\mathcal{P} = \{P_1, P_2, \dots, P_7\}$ is a projective plane of order 2 on V (i.e. $|P_i| = 3$, $|P_i \cap P_j| = 1$, $1 \leq i \neq j \leq 7$) such that

$$\mathcal{F} = \left\{ F \in \binom{X}{k} : P \in \mathcal{P}, P \subset F \right\}.$$

Füredi [35] proved the following general theorem, too.

THEOREM 2.3. *With the conditions of Theorem 2.2, there exists a sequence $1 = c_1 > c_2 > \dots > c_s > \dots$ of positive constants tending to 0 as $s \rightarrow \infty$ such that for $c_i \leq c < c_{i-1}$ $|\mathcal{F}| = \Omega(n^{k-i})$ holds. Moreover, if a projective plane of order $i - 1$ exists, then*

$c_i = i/(i^2 - i + 1)$. (Note that $|\mathcal{F}| = \Omega(n^\alpha)$ means that for fixed k and c there exist constants d_0, d_1 such that for every n we have $d_0 n^\alpha \leq |\mathcal{F}| \leq d_1 n^\alpha$.)

If we do not make restrictions on $|F|$ for $F \in \mathcal{F}$, we have

THEOREM 2.4 (Katona [47], Kleitman [49]). *Suppose $\mathcal{F} \subset 2^X$, $|F \cap F'| \geq l \geq 2$ for $F, F' \in \mathcal{F}$, and $|\mathcal{F}|$ is maximal. Then*

a) $n + l = 2t$.

$$\mathcal{F} = \{F \subset X : |F| \geq t\}.$$

b) $n + l = 2t + 1$. There exists an $x \in X$ such that

$$\mathcal{F} = \{F \subset X : |F \cap (X - \{x\})| \geq t\}.$$

Remark 2.5. In the case $l = 1$, one trivially has $|\mathcal{F}| \leq 2^{n-1}$ (if $F \in \mathcal{F}$ then $(X - F) \notin \mathcal{F}$) and, as in the case $n = 2k$ of Theorem 1.1, every nonextendable family \mathcal{F} has $|\mathcal{F}| = 2^{n-1}$.

Let us also mention

THEOREM 2.6 (Milner [56]). *Suppose $\mathcal{F} \subset 2^X$ is an l -intersecting antichain. Then*

$$|\mathcal{F}| \leq \binom{n}{\lfloor \frac{n+l+1}{2} \rfloor}. \text{ Moreover, for } n+l \text{ even equality holds iff } \mathcal{F} = \binom{X}{\frac{n+l}{2}}.$$

Remark 2.7. It is shown in [24] that for $n + l = 2t + 1$ there is only one more possibility for equality, namely

$$\mathcal{F} = \left\{ F \in \binom{X}{t+1} : Y \not\subset F \right\} \cup \left\{ F \in \binom{X}{t} : Y \subset F \right\}, \text{ where } Y \in \binom{X}{l} \text{ is fixed.}$$

THEOREM 2.8 (Hilton [42]). *Suppose $\mathcal{F} \subset 2^X$, $g \leq |F| \leq h \ \forall F \in \mathcal{F}$, where $0 < h \leq n$, $g \leq \min(h, n - h)$. Moreover, \mathcal{F} is l -intersecting. Then*

$$|\mathcal{F}| \leq \sum_{g \leq i \leq h} \binom{n-1}{i-1}.$$

THEOREM 2.9 (Green, Katona, Kleitman [36]). *Suppose $\mathcal{F} \subset 2^X$ is an antichain with $|F| \leq n/2$ and $F \cap F' \neq \emptyset$ for $F, F' \in \mathcal{F}$. Then*

$$\sum_{F \in \mathcal{F}} \binom{n-1}{|F|-1}^{-1} \leq 1.$$

3. Families with prescribed cardinalities for pairwise intersection. In this section we consider the following general problem.

Let $L = \{l_1, l_2, \dots, l_s\}$ be a set of integers with $0 \leq l_1 < \dots < l_s < k$. A family $\mathcal{F} \subset \binom{X}{k}$ is called an (n, k, L) -system if for all $F \neq F' \in \mathcal{F}$ we have $|F \cap F'| \in L$. What is the maximum cardinality of an (n, k, l) -system, which are the optimal families?

THEOREM 3.1 (Deza, Erdős, Frankl [10]). *Suppose \mathcal{F} is an (n, k, L) -system. Then*

a) For $n > n_0(k, L)$

$$|\mathcal{F}| \leq \prod_{1 \leq i \leq s} \frac{n - l_i}{k - l_i}.$$

b) There exists a constant $c = c(k, L)$ such that

$$|\mathcal{F}| \geq c \prod_{2 \leq i \leq s} \frac{n - l_i}{k - l_i}$$

implies $(l_2 - l_1)|(l_3 - l_2)| \cdots |(k - l_s)$ and $\exists X_0 \in \binom{X}{l_1}$ such that $X_0 \subseteq F, \forall F \in \mathcal{F}$.

Remark 3.2. Theorem 1.1 is the case $L = \{l, l + 1, \dots, k - 1\}$. The case of equality in a) corresponds to perfect matroid designs, i.e. matroids in which any two flats of the same rank have the same size (for survey see Deza and Singhi [11]).

Remark 3.3. The case $s = 1$ of the above theorem was earlier proved in [8], where for $c = c(k, L)$ the best possible bound $k^2 - k + 2$ is established. The case $s = 2$ was settled by Deza, Erdős and Singhi [9].

The following theorem combines results of Frankl and Wilson [33], and Frankl and Rosenberg [32]:

THEOREM 3.4. *Suppose \mathcal{F} is an (n, k, L) -system, $r \geq 2$, integer, such that $k \not\equiv l_i \pmod r$ for $i = 1, \dots, s$. Suppose the l_i 's lie in altogether t different congruence class mod r .*

a) *If r is a prime, then*

$$|\mathcal{F}| \leq \binom{n}{t}.$$

b) *If r is a prime power, then*

$$|\mathcal{F}| \leq \binom{n}{r-1}.$$

c) *If $t = 1$, then*

$$|\mathcal{F}| \leq n.$$

Remark 3.5. If we choose r to be a prime with $r > k$, then $k \not\equiv l_i \pmod r$ and $t = s$, thus we obtain $|\mathcal{F}| \leq \binom{n}{s}$, which was proved by Ray-Chaudhuri and Wilson [58]. The case c) improves earlier results by Babai and Frankl [1] and Deza and Rosenberg [14].

Remark 3.6. Theorem 3.4 has important applications. Here we sketch one of them. Let G_n be the graph whose vertices are the points in E^n , the Euclidean space of dimension n , and the edges those pairs of points whose Euclidean distance is 1. Suppose this graph has chromatic number m , and let $E^n = B_1 \cup B_2 \cup \dots \cup B_m$ be a corresponding coloration. Let p be a prime (we shall fix it later) and let X denote the set of points $x = (x_1, \dots, x_n)$ in E^n which satisfy $x_i = 0$ for $n - 2p + 1$ and $x_i = 1/\sqrt{2p}$ for $2p - 1$ values of i . For $x \in X$ define $F(x) = \{i : x_i = 1/\sqrt{2p}\}$. Then $F(x)$ is a $(2p - 1)$ -subset of $\{1, \dots, n\}$, and x, x' are at distance 1 if and only if $|F(x) \cap F(x')| = p - 1$ holds. Applying Theorem 3.4 with $k = 2p - 1, r = p, L = \{0, 1, \dots, p - 2, p, \dots, 2p - 2\}$ we obtain $|X \cap B_i| \leq \binom{n}{p-1}$ for $i = 1, \dots, m$. Thus $m \geq \binom{n}{2p-1} / \binom{n}{p-1}$ holds. Choosing $p \sim (2 - \sqrt{2})n/4$ we infer $m \geq (1.2)^n$ i.e. the chromatic number of G_n is growing exponentially.

Both Theorem 3.1 and Theorem 3.4 deal with general L ; they can be improved for particular choices of k and L . Let us denote by $m(n, k, L)$ $\max |\mathcal{F}|$, \mathcal{F} is an (n, k, L) -system.

In [30] and [19] the correct order of magnitude of $m(n, k, L)$ (i.e. upper and lower bounds which are only a constant factor apart) is determined for $k \leq 7$ and $k = 8, 9$ and L is arbitrary.

In [19] the case $|L| = 3$ is considered. Necessary and sufficient conditions are given for $m(n, k, L) = O(n)$ and $m(n, k, L) \geq O(n^2)$.

As a curiosity let us mention that $m(n, 12543, \{0, 112, 1233\}) = O(n)$ iff there is no projective plane of order 10.

A conjecture of Erdős and Sós (cf. Erdős [16]) was proved in [26]:

THEOREM 3.7. For $n \geq n_0(k)$, $k \geq 4$,

$$m(n, k, \{0, 2, 3, \dots, k-1\}) = \binom{n-2}{k-2}$$

and \mathcal{F} attains this bound iff for some $x, y \in X$ we have $\mathcal{F} = \{F \in \binom{X}{k} : \{x, y\} \subset F\}$.

Erdős and Frankl have the following general conjecture.

Conjecture 3.8. Suppose $n > n_0(k, l)$, $k > l \geq 0$.

a) If $k > 2l + 1$, then

$$m(n, k, \{0, 1, \dots, k\} - \{l\}) = \binom{n-l-1}{k-l-1},$$

and for $\mathcal{F} \subset \binom{X}{k}$ equality holds if and only if for some $Y \in \binom{X}{l+1}$ we have $\mathcal{F} = \{F \in \binom{X}{k} : Y \subseteq F\}$.

b) If $k \leq 2l + 1$, then

$$m(n, k, \{0, 1, \dots, k\} - \{l\}) \leq \binom{n}{l} \binom{2k-l-1}{k} / \binom{2k-l-1}{l}$$

and equality holds for $\mathcal{F} \subset \binom{X}{k}$ iff there exists an $(l, 2k-l-1, n)$ -Steiner system, ζ , and $\mathcal{F} = \{F \in \binom{X}{k} : \text{there exists an } S \in \zeta \text{ with } F \subseteq S\}$.

(A (t, s, n) -Steiner system is a family of s -subsets of an n -set, such that each t -subset is contained in exactly one member of the family.)

Using Theorem 3.4b, it was proved in [33] that the inequality of Conjecture 3.8b is true if $k \leq 2l + 1$, and $k - l$ is a prime power.

Combining a result of [29] and part b) of Theorem 3.4, we have

THEOREM 3.9. If $k \geq 3l + 2$ or if $k > 2l + 1$ and $k - l$ is a prime power, then

$$m(n, k, \{0, 1, \dots, k-1\} - \{l\}) = \left(1 + o(1) \binom{n-l-1}{k-l-1}\right).$$

The determination of the correct order of magnitude of $m(n, k, L)$ seems to be hopeless. We cannot even decide whether there exists some integer k which is not of the form $2^a - 1$ or 3^b but $m(n, k, \{0, 1, 3\}) > cn^3$ for some positive constant c (any such k should satisfy $k \equiv 1$ or $3 \pmod{6}$).

4. More generalizations of the Erdős–Ko–Rado theorem for systems of finite sets.

THEOREM 4.1 (Erdős [18]). Let $s \geq 2$ be an integer, $\mathcal{F} \subset \binom{X}{k}$, and suppose \mathcal{F} does not contain F_1, \dots, F_s such that $F_i \cap F_j = \emptyset$ for all $i \leq i < j \leq k$. Then for $n > n_0(k, s)$

$$|\mathcal{F}| \leq \binom{n}{k} - \binom{n-s+1}{k},$$

and equality holds iff for some $Y \in \binom{X}{s-1}$: $\mathcal{F} = \{F \in \binom{X}{k} : F \cap Y \neq \emptyset\}$.

Remark 4.2. The case $s = 2$ corresponds to the case $l = 1$ of Theorem 1.1. As to the bound $n_0(k, s)$ we conjecture $n_0(k, r) < ckr$ but only $n_0(k, s) \leq 2k^3s$ (Bollobas, Daykin, Erdős [2]) and $n_0(k, s) \leq c'ks^2$ ([19]) are known. In [31] under much more general conditions a weaker estimate $|\mathcal{F}| < ken^{k-1}$ is proved.

Remark 4.3. For the case $l > 1$ of Theorem 1.1., Hajnal and Rothschild [41] gave the corresponding generalization. In Deza, Erdős, Frankl [10] asymptotic bounds, were obtained for the more general case: if among any s members F_1, \dots, F_s of \mathcal{F} there are two with $|F_i \cap F_j| \in L$.

Kleitman [50] considered the problem of maximizing $|\mathcal{F}|$ such that $\mathcal{F} \subset 2^X$, \mathcal{F} contains no s pairwise disjoint sets, i.e. the nonuniform case. He obtained best possible bounds for $n \equiv 0, -1 \pmod s$.

There are several generalizations of Theorem 1.1 to multiple intersections. We list some of them below.

THEOREM 4.4 [23]. *Suppose $\mathcal{F} \subset \binom{X}{k}$, any t members of \mathcal{F} have nonempty intersection and $n \geq (t/(t-1))k$. Then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1}.$$

Remark 4.5. The restriction $n \geq (t/(t-1))k$ is best possible, since for $k > ((t-1)/t)n$ any t sets of cardinality k have nonempty intersection. It would be very interesting to obtain best possible bounds also for the case $l > 1$, analogously with Theorem 1.3. The case \mathcal{F} is an antichain was solved for $n > 1.000$, $t = 3$ in [23] and for $t \geq 4$ by Gronau [38], who also settled most of the remaining cases for $t = 3$ (cf. Gronau [39], [40]).

Let $X_i \in \binom{X}{i+n}$ and define $\mathcal{F}_i^t = \{F \in \binom{X}{k} : |F \cap X_i| \geq l + (t-1)i\}$.

Conjecture 4.6. Suppose $\mathcal{F} \subset \binom{X}{k}$, $\forall F_1, \dots, F_t \in \mathcal{F}$, $|F_1 \cap \dots \cap F_t| \geq l$ holds. Then

$$|\mathcal{F}| \leq \max_{0 \leq i \leq k} |\mathcal{F}_i^t|.$$

This conjecture would generalize Conjecture 1.5. With the above notation, define $\mathcal{A}_i^l = \{A \subset X : |A \cap X_i| \geq l + (t-1)i\}$. Then $\mathcal{A}_i^l \subset 2^X$ and for every $A_1, \dots, A_t \in \mathcal{A}_i^l$, of course, $|A_1 \cap \dots \cap A_t| \geq l$ holds.

Conjecture 4.7 [28]. Suppose $\mathcal{A} \subset 2^X$ and $\forall A_1, \dots, A_t \in \mathcal{A}$, $|A_1 \cap \dots \cap A_t| \geq l$ holds. Then

$$|\mathcal{A}| \leq \max_{0 \leq i \leq (n-l)/t} |\mathcal{A}_i^l|.$$

In [28], this conjecture is proved for $l \leq t^2/150$. Theorem 2.4 shows that it holds for $t = 2$. In [20], it is proved for $l = 2$, $t = 3$ ($|\mathcal{A}| \leq |\mathcal{A}_0^l| = 2^{n-2}$).

Most of the theorems could have been formulated for unions instead of intersections—it suffices to take the complement of the sets. However, if we make restrictions on both unions and intersections at the same time, interesting new problems arise. The following result was conjectured by Katona in [46].

THEOREM 4.8 [22]. *Let $n > t \geq 1$ and suppose $\mathcal{F} \subset 2^X$ such that $\forall F_1, F_2 \in \mathcal{F}$, $F_1 \cap F_2 \neq \emptyset$, $|F_1 \cup F_2| \leq n - l$.*

a) *If $n - 1 + l = 2t$, then*

$$|\mathcal{F}| \leq \sum_{i=t}^{n-1} \binom{n-1}{i}.$$

b) *If $n - 1 + l = 2t + 1$, then*

$$|\mathcal{F}| \leq 2 \sum_{i=t}^{n-2} \binom{n-2}{i}.$$

Here we give a new and short proof of Theorem 4.8 using a result of Chvátal (mentioned in Remark 1.10) and Theorem 2.4 (Katona, Kleitman).

Let \mathcal{F} be as in the theorem. Define $\mathcal{F}_* = \{G : \exists F \in \mathcal{F}, G \subseteq F\}$. Then $\mathcal{F} \subset \mathcal{F}_*$ and \mathcal{F}_* is an ideal, the ideal generated by \mathcal{F} . Moreover for $F, F' \in \mathcal{F}_*$, $|F \cup F'| \leq n - l$ holds.

Let us apply repeatedly the exchange operation (cf. Definition 1.6) to \mathcal{F}_* . At last we obtain a family which is stable under this operation. Let it be \mathcal{A}_* and let $\mathcal{A} \subset \mathcal{A}_*$ the family corresponding to \mathcal{F} . Of course $|\mathcal{A}| = |\mathcal{F}|$, \mathcal{A} satisfies the assumptions of the theorem, and for all $A, A' \in \mathcal{A}_*$, $|A \cup A'| \leq n - l$ holds.

As \mathcal{A}_* is hereditary, stable, and any two elements of \mathcal{A} have nonempty intersection, we may apply the above cited result of Chvátal: there exists $x \in X$, such that $|\mathcal{A}_*(x) = \{A \in \mathcal{A}_*: x \in A\}| \geq |\mathcal{A}|$.

Set $\tilde{\mathcal{A}} = \{A - \{x\}: A \in \mathcal{A}_*(x)\}$, $\mathcal{B} = \{(X - \{x\}) - A: A \in \tilde{\mathcal{A}}\}$.

For $A, A' \in \tilde{\mathcal{A}}$, $|A \cup A'| \leq n - 1 - l$. Thus for $B, B' \in \mathcal{B}$, $|B \cap B'| \geq l$ holds. We may apply Theorem 2.4 to $\mathcal{B} \subset 2^{X - \{x\}}$. The statement of Theorem 4.8 follows from $|\mathcal{A}| \leq |\mathcal{A}_*(x)| = |\tilde{\mathcal{A}}| = |\mathcal{B}|$.

For the general case we have

Conjecture 4.9 [21]. Let $l, r \geq 2$ be integers. Suppose $\mathcal{F} \subset 2^X$ satisfies $|F \cup F'| \leq n - l$, $|F \cap F'| \geq r$ for all $F, F' \in \mathcal{F}$. Assume, moreover, that $|\mathcal{F}|$ is maximal. Then there exist integers $s, t \geq 0$ and disjoint sets $A, B \in X^2$ such that $|A| = 2s + l$, $|B| = 2t + r$ and

$$\mathcal{F} = \{F \in 2^X: |F \cap A| \leq s, |F \cap B| \geq t + r\}.$$

Let us note that Winkler has recently stated the same conjecture (cf. [61]).

5. Algebraic generalizations. The subsets of $X = \{x_1, \dots, x_n\}$ can be represented by 0-1 sequences, their characteristic functions:

Let A be a subset of X , and define $f: \{1, 2, \dots, n\} \rightarrow \{0, 1\}$ by $f_A(i) = 1$ iff $x_i \in A$. Then $|A \cap A'|$ is the number of nonzero positions in which the two sequences agree.

For fixed integers $k, s, l \geq 1$, it is natural to ask what is the maximum number of functions $f: \{1, 2, \dots, n\} \rightarrow \{0, 1, \dots, s\}$ such that for any two functions f, f'

- (*) $|\{i: f(i) \neq 0\}| = k$,
- (**) $|\{i: f(i) = f'(i) \neq 0\}| \geq l$.

We shall denote by $T_{n,s}$ the set of all functions $f: \{1, \dots, n\} \rightarrow \{0, \dots, s\}$.

The Erdős-Ko-Rado theorem gives the answer for $s = 1$, $n \geq n_0(k, l)$.

Let us define the *pushing-up operation* in the i th position P_i for a family \mathcal{F} of functions by

$$P_i(f)(j) = f(j) \quad \text{for } j \neq i,$$

$$P_i(f)(i) = \begin{cases} s & \text{if } f(i) \neq 0 \text{ and the function defined by setting } f(i) = s \text{ is not in } \mathcal{F}, \\ f(i) & \text{otherwise.} \end{cases}$$

Saying it with words, we replace f in $f(i) \neq 0$ by a function which differs from it only in the i th position, where its value is s , if this new function was not yet in the system.

We set, of course, $P_i(\mathcal{F}) = \{P_i(f): f \in \mathcal{F}\}$.

It is easy to check that the number of nonzero positions of f is the same as that of $P_i(f)$. If \mathcal{F} satisfies (*) and (**), then so does $P_i(\mathcal{F})$. Repeated application of these operations, for $1 \leq i \leq n$ yields a family $\tilde{\mathcal{F}}$ which is stable under the application of P_i i.e. $P_i(\tilde{\mathcal{F}}) = \tilde{\mathcal{F}}$ for $1 \leq i \leq n$.

For $f \in \tilde{\mathcal{F}}$ define $B(f) = \{i: f(i) = s\}$, $\mathcal{B} = \{B(f): f \in \tilde{\mathcal{F}}\}$.

PROPOSITION 5.1. *If $\tilde{\mathcal{F}}$ satisfies (*) and (**), then*

- a) $|B(f)| \leq k$ for $f \in \tilde{\mathcal{F}}$,
- b) $|B(f) \cap B(f')| \geq l$ for $f, f' \in \tilde{\mathcal{F}}$.

Proof. Let us define the function g by $g(i) = f(i)$ if $f(i) = 0$ or $f'(i) = s$ and $g(i) = s$ otherwise. As $\tilde{\mathcal{F}}$ is stable, $g \in \tilde{\mathcal{F}}$. The two functions g and f' agree in a nonzero position j if $g(j) = f'(j) = s$, thus $f(j) = s$, which means $j \in (B(f) \cap B(f'))$, proving b), a) is trivial. \square

Let h be any function in $T_{n,s}$ such that h has k nonzero values and $B(h) = B(f)$ for some $f \in \tilde{\mathcal{F}}$. Then, in view of Proposition 5.1 h has at least l nonzero common values with every $f' \in \tilde{\mathcal{F}}$. Thus if $|\mathcal{F}|$ was maximal then $h \in \tilde{\mathcal{F}}$. This means that $\tilde{\mathcal{F}}$ can be defined via $\mathcal{B} = \{B(f) : f \in \tilde{\mathcal{F}}\}$ in the following way

$$\tilde{\mathcal{F}} = \{f \in T_{n,s} : B(f) \in \mathcal{B}\}.$$

On the other hand, let \mathcal{B} be a subset of $2^{\{1, \dots, n\}}$, and suppose \mathcal{B} is l -intersecting with $|B| \leq k$ for $B \in \mathcal{B}$. Let $\mathcal{F}(\mathcal{B}) = \{f \in T_{n,s} : B(f) \in \mathcal{B}; |\{i : f(i) \neq 0\}| = k\}$.

Then $\mathcal{F}(\mathcal{B})$ satisfies (*) and (**). Moreover set $b_r = |\{B : B \in \mathcal{B}, |B| = r\}|$. Then it follows that

$$(***) \quad |\mathcal{F}(\mathcal{B})| = \sum_{0 \leq r \leq k} b_r \binom{n-r}{k-r} (s-1)^{k-r}.$$

THEOREM 5.2. *If $l = 1$ and \mathcal{F} satisfies (*) and (**), then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1} s^{k-1} = |\{f \in T_{n,s} : |\{i : f(i) \neq 0\}| = k, f(1) = s\}|.$$

Proof. In view of the above preparations, we may assume $\mathcal{F} = \mathcal{F}(\mathcal{B})$ for a $\mathcal{B} \subset 2^{\{1, 2, \dots, n\}}$ with $|B| \leq k$ and $|B \cap B'| \geq 1$ for all $B, B' \in \mathcal{B}$. If $k > n/2$ and j is some integer with $n/2 < j \leq k$. Then $b_j + b_{n-j} \leq \binom{n}{j}$ (since from any set and its complement at most 1 is in \mathcal{B}). Also $b_{n-j} \leq \binom{n-1}{n-j-1}$ (from Theorem 1.1). Together this yields:

$$\begin{aligned} & b_j \binom{n-j}{k-j} (s-1)^{k-j} + b_{n-j} \binom{n-(n-j)}{k-(n-j)} (s-1)^{k-(n-j)} \\ & \leq \binom{n-1}{j-1} \binom{n-j}{k-j} (s-1)^{k-j} + \binom{n-1}{n-j-1} \binom{n-(n-j)}{k-(n-j)} (s-1)^{k-(n-j)}. \end{aligned}$$

Using these inequalities and $b_r \leq \binom{n-1}{r-1}$ for $r \leq n/2$, we obtain

$$|\mathcal{F}| \leq \sum_{0 \leq j \leq k} \binom{n-1}{j-1} \binom{n-j}{k-j} (s-1)^{k-j} = \binom{n-1}{k-1} s^{k-1}.$$

Remark 5.3. This theorem was first stated by Meyer [54]. However his proof was incomplete. Hence, we included this proof, which was given by the second author in 1976, but was never published. One can also prove the uniqueness of the extremal systems unless $s = 1, k = n$. The case $l > 1$ is more complicated. The above proof yields

THEOREM 5.4. *Suppose $n > n_0(k, l)$, and \mathcal{F} satisfies (*) and (**). Then, $|\mathcal{F}| \leq \binom{n-l}{k-l} s^{k-l}$, and equality holds iff for l pairs of integers $(i_t, j_t), 1 \leq i_1 < \dots < i_l \leq n, 1 \leq j_t \leq s, 1 \leq t \leq l$ we have*

$$\mathcal{F} = \{f \in T_{n,s} : f(i_t) = j_t \text{ for } 1 \leq t \leq l, f \text{ satisfies } (*)\}.$$

Frankl and Füredi [34] proved

THEOREM 5.5. *Suppose $k = n$, \mathcal{F} satisfies (*) and (**), and \mathcal{F} has maximal cardinality. Then for $l \geq 15$*

$$|\mathcal{F}| = s^{n-l} \text{ iff } l + 1 \leq s.$$

Remark 5.6. The proof gives an interesting application of Theorem 1.3. As a matter of fact Theorem 1.3 implies via (***) that $|\mathcal{F}|/s^n \leq (1 + o(1))s^{-l}$ from which $|\mathcal{F}| \leq s^{n-l}$ is deduced quite easily for every n , i.e. we deduce an exact result from an asymptotic one. Recently Moon [59] gave a nice proof for Theorem 5.5. However her proof works only in the case $s \geq l + 2$.

Remark 5.7. It is relatively easy to determine the maximal number of $(1, 2, \dots, s)$ sequences of length n such that they agree in at least one out of any r consecutive positions. This maximum is attained by the sequences which have 1 in the r 'th, $2r$ th, \dots , $[n/r] \cdot r$ th positions.

An analogous problem of coding type was considered by Miczo [55].

A special type of functions are permutations. Let us denote by $R(n, \geq l)$ the maximum number of permutations of $\{1, 2, \dots, n\}$ such that any two permutations $\pi(i), \rho(i)$ agree in at least l positions (i.e. $\pi\rho^{-1}$ has at least l fixed-points). Let S_n denote the symmetric group on n elements. For $\pi \in S_n$ define $F(\pi) = \{i \in \{1, \dots, n\} : \pi(i) = 1\}$. Define further $P(k, s) = \{\pi \in S_n : |F(\pi) \cap \{1, \dots, k\}| \geq s\}$. We use the notation $\pi P = \{\pi\rho : \rho \in P\}$.

THEOREM 5.8 [13]. *Suppose $n > n_0(n-l)$ and P is a set of permutations such that any two members of P agree in at least l positions. Assume moreover, that $|P|$ is maximal.*

a) *If $n+l = 2t$. Then there exists a $\pi \in S_n$ such that*

$$P = \pi P(n, t).$$

b) *If $n+l = 2t+1$. Then there exists a $\pi \in S_n$ such that*

$$P = \pi P(n-1, t).$$

For the case $n > n_0(l)$, we made in [13] the following conjecture.

Conjecture 5.9. *If $n > n_0(l)$, then*

$$R(n, \geq l) = (n-l)!$$

Remark 5.10. In [13] we proved that the above conjecture is true if in S_n there exists a sharply l -transitive set Q of permutations, i.e. for any two ordered l -subsets of S_n there is exactly one permutation in Q , mapping the first on the second. In particular for $l=1, n$ arbitrary; $l=2, n$ a prime power; $l=3, n$ a prime power plus one. At the time being we can only prove the following result.

THEOREM 5.11. *Suppose $P \subset S_n$ and any three elements of P have at least l positions in common. Then for $n \geq n_0(l)$ we have $|P| \leq (n-l)!$*

Of course, this theorem yields the same bound if we replace 3 by $t, t \geq 3$.

The determination of $R(n, \geq l)$ would be settled by the following conjecture.

Conjecture 5.12 [12]. *There exists a family \mathcal{F} of subsets of $\{1, 2, \dots, n\}$ such that $|F \cap F'| \geq l$ for $F, F' \in \mathcal{F}$ and*

$$R(n, \geq l) = |P = \{\pi \in S_n : F(\pi) \in \mathcal{F}\}|.$$

Taking into consideration that every permutation is a function $\pi: (1, 2, \dots, n) \rightarrow \{1, \dots, n\}$ Theorem 5.5 yields, for $l \geq 15, R(n, \geq l) \leq n^{n-l} \sim (n-l)! e^{n-l} \sqrt{2\pi n}$.

S_n can be made to a metric space by defining $d(\pi, \rho) = n - |F(\pi\rho^{-1})|$. With this terminology we are concerned with anticodes i.e. sets with given maximal distance d . Theorem 5.8 states that for $n > n_0(d)$ any maximal sized anticode is a sphere or near sphere. Note the analogy between Theorem 5.8 and Theorem 2.4.

Let us denote by $R(n, L)$ the maximum cardinality of a set of permutations $P \subseteq S_n$ such that for every $\pi \neq \rho \in P$ we have $|F(\pi\rho^{-1})| \in L$ ($L = \{l_1, \dots, l_s\} \subseteq \{0, \dots, n-2\}$). Kyota [49] gave a very elegant argument showing that if P is a group, then $|P| \leq \prod_{i=1}^s (n-l_i)$. For $L = \{0, 1, \dots, l-1\}$ equality corresponds to sharply l -transitive permutation groups.

Intersection problems can be considered for vector spaces as well. Hsieh [44] proved the following

THEOREM 5.13. *Let \mathcal{F} be a family of k -dimensional subspaces of $V(n, q)$, an n -dimensional vector space over the finite field of q elements. Suppose for all $A, B \in \mathcal{F}$ we have $\dim(A \cap B) \geq l > 0$. Assume $k \leq (n - 1)/2$, or $k < (n - 1)/2$ in the case $q = 2$, $l > 1$. Then*

$$|\mathcal{F}| \leq \binom{n-l}{k-l}_q = \text{the number of } (k-l)\text{-dimensional subspaces of } V(n-l, q).$$

Greene and Kleitman [37] showed, employing a method of Katona [45], that Theorem 5.13 remains true for $l = 1$ and $n = 2k$. Hsieh's proof is long and involves a lot of calculation. Here we sketch how the case $l = 1$ can be deduced quickly using the special case $n = 2k$.

We apply induction on n , starting from $2k$. Let n be the smallest value for which the statement is not proved yet. Let v_1, v_2, \dots, v_n form a basis for $V(n, q)$. Set $V_i = \langle v_1, \dots, v_i \rangle$, the subspace generated by v_1, \dots, v_i . For $u \in V_i, v \in V_n - V_i$, we define an exchange operation.

Let $A \in \mathcal{F}$, such that $u \notin A, v \in A$, then choose an arbitrary k -subspace of $\langle A, v \rangle$ containing u but not v , and which is not yet in \mathcal{F} (if such a subspace exists), replace A by this subspace (simultaneously for all such $A \in \mathcal{F}$). We keep on applying this operation for all possible pairs (u, v) until we obtain a stable set $\tilde{\mathcal{F}}$ (i.e. $\tilde{\mathcal{F}}$ is invariant under the exchange operation). We claim that $\tilde{\mathcal{F}}$ has the property $A \cap B \neq \{0\}$, and even that any $A, B \in \tilde{\mathcal{F}}$ have nontrivial intersection in V_{n-1} . In fact, suppose the contrary, and let $A_0 = A \cap V_{n-1}, B_0 = B \cap V_{n-1}, 0 \neq w \in A \cap B$. As $\dim A_0 + \dim B_0 = 2k - 2 < n - 1$ there exists a k -dimensional subspace B' of V_{n-1} such that $B_0 \subset B', A_0 \cap B' = \{0\}$. Let $u \in B' - B_0$. Then the application of the exchange operation u, v could exchange B for B' . But $\tilde{\mathcal{F}}$ is stable, thus $B' \in \tilde{\mathcal{F}}$. Moreover, $B' \cap A = \{0\}$, a contradiction. Now the result follows by induction.

REFERENCES

[1] L. BABAI AND P. FRANKL, *Note on set-intersections*, J. Combin. Theory A, 28 (1980), pp. 103-105.
 [2] B. BOLLOBAS, D. E. DAYKIN AND P. ERDÖS, *On the number of independent edges in a hypergraph*, Quart. J. Math. Oxford, (2) 27 (1976), pp. 25-32.
 [3] A. BROUWER AND P. FRANKL, unpublished.
 [4] V. CHVÁTAL, *Intersection families of edges in hypergraphs having the hereditary property*, Hypergraph Seminar, Springer-Verlag, Berlin, 1974, pp. 61-66.
 [5] G. CLEMENTS AND B. LINDSTRÖM, *A generalization of a combinatorial theorem of Macaulay*, J. Combin. Theory, 7 (1969), pp. 230-238.
 [6] D. E. DAYKIN, *Erdős-Ko-Rado from Kruskal-Katona*, J. Combin. Theory A, 17 (1974), p. 252.
 [7] P. DELSARTE, *An algebraic approach to the association schemes in coding theory*, Philips Res. Reports Suppl., 10, 1973.
 [8] M. DEZA, *Solution d'un problème de Erdős and Lovasz*, J. Combin. Theory B, 16 (1974), pp. 166-167.
 [9] M. DEZA, P. ERDÖS AND N. SINGHI, *Combinatorial problems on subsets and their intersections*, Adv. in Math. Suppl. Stud., 1 (1978), pp. 259-265.
 [10] M. DEZA, P. ERDÖS AND P. FRANKL, *On intersection properties of the system of finite sets*, Proc. London Math. Soc., (3) 36 (1978), pp. 369-384.
 [11] M. DEZA AND N. SINGHI, *Some properties of perfect matroid-designs*, Ann. Disc. Math., 6 (1980), pp. 57-76.
 [12] M. DEZA, *Pavage généralisé parfait comme généralisation de matroïde-configuration et de simple t -configurations*, Proc. de Colloque Int. CNRS sur le Combinatoire, Orsay, 1976, pp. 97-100.
 [13] M. DEZA AND P. FRANKL, *Maximum number of permutations with given minimal or maximal distance*, J. Combin. Theory A, 22 (1977), pp. 352-760.
 [14] M. DEZA AND I. G. ROSENBERG, *Cardinalités de sommets et d'arêtes d'hypergraphes satisfaisant certaines conditions sur l'intersection d'arêtes*, Cahiers CERO 20, 3-4 (1978), pp. 279-286.

- [15] P. ERDŐS, C. KO AND R. RADO, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford, (2) 12 (1961), pp. 313–320.
- [16] P. ERDŐS, *Problems and results in combinatorial analysis*, Théorie Combinatoire, Colloq. int. Roma 1973, vol. 2, Acad. Naz. Lincei, Roma, 1976, pp. 3–17.
- [17] P. ERDŐS AND D. J. KLEITMAN, *Extremal problems among subsets of a set*, Disc. Math., 8 (1974), pp. 281–294.
- [18] P. ERDŐS, *A problem on independent r -tuples*, Ann. Univ. Sci. Budapest, 8 (1965), pp. 93–95.
- [19] P. FRANKL, unpublished.
- [20] ———, *An intersection theorem for finite sets*, Bull. Austral. Math. Soc., 15 (1976), pp. 73–79.
- [21] ———, Ph.D. thesis, Budapest, 1975.
- [22] ———, *The proof of a conjecture of Katona*, J. Combin. Theory A, 19 (1975), pp. 208–213.
- [23] ———, *On Sperner families satisfying an additional condition*, J. Combin. Theory A, 20 (1976), pp. 1–11.
- [24] ———, *Generalizations of theorems of Katona and Milner*, Acta Math. Acad. Sci. Hung., 27 (1976), pp. 359–363.
- [25] ———, *On intersecting families of finite sets*, J. Combin. Theory A, 24 (1978), pp. 146–161.
- [26] ———, *On families of finite sets no two of which intersect in a singleton*, Bull. Austral. Math. Soc., 17 (1977), pp. 125–134.
- [27] ———, *The Erdős–Ko–Rado theorem is true for $n = ckt$* , Proc. Fifth Hung. Comb. Coll., North-Holland, Amsterdam, 1978, pp. 365–375.
- [28] ———, *Families of finite sets satisfying a union condition*, Disc. Math., 26 (1979), pp. 111–118.
- [29] ———, *Extremal problems and coverings of the space*, Europ. J. Combin., 1 (1980), pp. 101–106.
- [30] ———, *Families of finite sets with prescribed cardinalities for pairwise intersections*, Acta Math. Acad. Hung., 35 (1980), pp. 351–360.
- [31] ———, *A general intersection theorem for finite sets*, Ann. Disc. Math., 9 (1980), pp. 43–49.
- [32] P. FRANKL AND I. G. ROSENBERG, *A finite set intersection theorem*, Europ. J. Combin., 2 (1981), pp. 127–129.
- [33] P. FRANKL AND R. M. WILSON, *Intersection theorems with geometric consequences*, Combinatorica, 1 (1981), pp. 357–368.
- [34] P. FRANKL AND Z. FÜREDI, *The Erdős–Ko–Rado theorem for integer sequences*, this Journal, 1 (1980), pp. 376–781.
- [35] Z. FÜREDI, *Erdős–Ko–Rado type theorems with upper bounds on the maximum degree*, Colloquia Math. Soc. J. Bolyai 25, Szeged, 1978, pp. 177–207.
- [36] C. GREENE, G. O. H. KATONA AND D. J. KLEITMAN, *Extensions of the Erdős–Ko–Rado theorem*, Stud. Appl. Math., 55 (1976), pp. 1–8.
- [37] C. GREENE AND D. J. KLEITMAN, *Proof techniques in the theory of finite sets*, in MAA Studies in Math., 17, Mathematical Association of America, 1978, pp. 12–79.
- [38] H. D. GRONAU, *On Sperner families in which no k sets have an empty intersection*, J. Combin. Theory A, 28 (1980), pp. 54–63.
- [39] ———, *On Sperner families in which no 3 sets have an empty intersection*, Acta Cybernetica, 4 (1978), pp. 213–220.
- [40] ———, *On Sperner families in which no k sets have an empty intersection*, J. Combin. Theory A, 30 (1981), pp. 298–316.
- [41] A. HAJNAL AND B. ROTHSCHILD, *A generalization of the Erdős–Ko–Rado theorem on finite set systems*, J. Combin. Theory A, 15 (1973), pp. 359–362.
- [42] A. J. W. HILTON, *Analogs of a theorem of Erdős, Ko and Rado on a family of finite sets*, Quart. J. Math. Oxford, (2) 25 (1974), pp. 19–28.
- [43] A. J. W. HILTON AND E. C. MILNER, *Some intersection theorems for systems of finite sets*, Quart. J. Math. Oxford, (2) 18 (1967), pp. 369–384.
- [44] W. N. HSIEH, *Intersection theorems for systems of finite vector spaces*, Disc. Math., 12 (1975), pp. 1–16.
- [45] G. O. H. KATONA, *A simple proof of the Erdős–Chao Ko–Rado theorem*, J. Combin. Theory B, 13 (1972), pp. 183–184.
- [46] ———, *Extremal problems for hypergraphs*, Math. Centre Tracts 56, Amsterdam, 1974, pp. 13–42.
- [47] ———, *Intersection theorems for finite sets*, Acta Math. Acad. Sci. Hungar., 15 (1964), pp. 329–337.
- [48] ———, *A theorem on finite sets*, in Theory of Graphs, Proc. Colloq. Tihany, Budapest, 1968, pp. 187–207.
- [49] D. J. KLEITMAN, *On a combinatorial conjecture of Erdős*, J. Combin. Theory, 1 (1966), pp. 209–214.
- [50] ———, *Maximum number of subsets of a finite set no k of which are pairwise disjoint*, J. Combin. Theory, 5 (1968), pp. 157–163.
- [51] ———, *Extremal hypergraph problems*, Proc. Seventh British Combinatorial Conference, London Math. Society, Lecture Notes, 38, 1979, pp. 44–65.

- [52] J. B. KRUSKAL, *The number of simplices in a complex*, in *Mathematical Optimization Techniques*, Univ. California Press, Berkeley, 1963, pp. 251–278.
- [53] M. KYOTA, *An inequality for finite permutation groups*, *J. Combin. Theory A*, 27 (1979), p. 119.
- [54] J.-C. MEYER, *Quelques problèmes concernant les cliques des hypergraphes k -complets et q -parti h -complets*, *Hypergraph Seminar*, Springer-Verlag, Berlin, 1974, pp. 127–139.
- [55] A. MICZO, *A bound of lightweight sequences with application to definite decoding*, *IEEE Trans. Inform. Theory*, IT-20 (1974), pp. 535–538.
- [56] E. C. MILNER, *A combinatorial theorem on system of sets*, *J. London Math. Soc.*, 43 (1968), pp. 204–206.
- [57] A. MOON, *An analogue of the Erdős-Ko-Rado theorem for the Hamming schemes $H(n, q)$* , *J. Combin. Theory A*, 32 (1982), pp. 386–390.
- [58] D. RASY-CHAUDHURI AND R. M. WILSON, *On t -designs*, *Osaka J. Math.*, 12 (1975), pp. 1–16.
- [59] A. SCHRIJVER, *Association schemes and the Shannon capacity, Eberlein polynomials and the Erdős-Ko-Rado theorem*, *Colloquia Math. Soc. J. Bolyai* 25, Szeged, 1978, pp. 671–680.
- [60] C. E. SHANNON, *The zero-error capacity of a noisy channel*, *IRE Trans. Inform. Theory*, 3 (1956), pp. 3–15.
- [61] D. B. WEST, *Extremal problems in partially ordered sets*, *Proc. Colloquium on Ordered Sets*, Banff, 1981, D. Reidel, Dordrecht, 1982.

ON THE MATRIX EQUATION $AX + X^*A^* = C^\dagger$

P. LANCASTER[‡] AND P. ROZSA[§]

Abstract. The structure of the solution set of the matrix equation $AX + X^*A^* = C$ is described when all (possibly rectangular) matrices involved are complex, and when they are all real.

1. Introduction. Let $\mathbb{C}^{m \times n}$ and $\mathbb{R}^{m \times n}$ denote the set of all $m \times n$ matrices with complex and real entries, respectively. If $A \in \mathbb{C}^{m \times n}$ then a function $G_A: \mathbb{C}^{n \times m} \rightarrow \mathbb{C}^{m \times m}$ is defined by

$$(1) \quad G_A(X) = AX + X^*A^*,$$

where the asterisk denotes conjugate transpose. O. Taussky and H. Wielandt in [3] investigated properties of this function, as well as the *linear* function \hat{G}_A defined by

$$(2) \quad \hat{G}_A(X) = AX + X^T A^T,$$

where the T denotes the transpose, but only when matrix A is square. Our main concern in this note is with solutions of the equation

$$(3) \quad G_A(X) = AX + X^*A^* = C,$$

where, of course, $C^* = C$. We do not use the major results of Taussky and Wielandt but will take advantage of some of their observations. We return to the function \hat{G}_A in § 3. The first observation is that G_A is linear on the space $\mathbb{C}^{n \times m}$ over the *real* field and with this understanding the solution set of (3) can, as usual, be described in terms of a particular solution and the kernel of G_A , (written $\text{Ker } G_A$).

Now the equation $AX + X^*A^* = 0$ is equivalent to the relation $AX = S$ where S is an arbitrary $m \times m$ skew-hermitian matrix. Consequently, *when A is $m \times m$ and invertible, $X \in \text{Ker } G_A$ if and only if $X = A^{-1}S$ for some skew-hermitian S , and the dimension of $\text{Ker } G_A$ is just m^2 , the number of real parameters in S .*

In contrast, one sees immediately that the dimension of $\text{Ker } \hat{G}_A$ (when A is $m \times m$, real, and invertible) is $\frac{1}{2}m(m-1)$.

Equations of the form (3) arise in a number of applications. For example, the solutions of a time-invariant Hamiltonian system of the general form $H\dot{x}(t) = iKx(t)$ with $\det H \neq 0$, $H^* = H$, $K^* = K$, are determined by the eigenvalues and eigenvectors of the H -selfadjoint matrix $A = H^{-1}K$. It is shown in § 4 that examination of the relationship between the eigenvectors of A associated with eigenvalues in the open upper and lower halves of the complex plane leads to an equation of the form (3).

2. The general case. Proceed now to the case of a singular matrix $A \in \mathbb{C}^{m \times n}$ and first perform a *rank factorization* of A . Thus, if the rank of A is $r \leq \min(m, n)$ then there is a $U \in \mathbb{C}^{m \times r}$ and $V \in \mathbb{C}^{r \times n}$, both of full rank, for which

$$(4) \quad A = UV^*.$$

In addition, it may be assumed without loss of generality that the rows of V^* are orthonormal, i.e. that $V^*V = I$. (Note also that r is the rank of A in the usual sense, *not* as a transformation on \mathbb{C}^r over \mathbb{R} .) Then equation (3) implies that

$$(5) \quad G_U(Y) = UY + Y^*U^* = C$$

[†] Received by the editors August 2, 1982, and in revised form November 10, 1982. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26-29, 1982.

[‡] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada, T2N 1N4.

[§] Technical University, Műegyetem rkp. 3, Budapest, XI, Hungary.

where

$$(6) \quad V^*X = Y.$$

It may also be assumed that U has the partitioned form

$$(7) \quad U = \begin{bmatrix} U_r \\ U_{m-r} \end{bmatrix}$$

where U_r is invertible and $U_{m-r} \in \mathbb{C}^{(m-r) \times r}$.

All solutions of (3) (when such exist) will be described by first obtaining the solution set of (5) and combining these with the general solution of (6). We introduce some more notation: for any matrix M , M^{-*} denotes $(M^{-1})^*$. Let the hermitian matrix C of (3) be partitioned in the form

$$(8) \quad C = \begin{bmatrix} C_1 & C_2^* \\ C_2 & C_3 \end{bmatrix},$$

where C_1 is $r \times r$, C_3 is $m-r \times m-r$, and let

$$(9) \quad W = [-U_{m-r}U_r^{-1} \quad I_{m-r}],$$

an $(m-r) \times m$ matrix. Note that W is independent of the choice of factors U and V in (4), and $WU = WA = 0$ so that $WCW^* = 0$ is an obvious necessary condition for the existence of a solution of (3).

LEMMA 1. (a) *There exist solutions Y of (5) if and only if $WCW^* = 0$.*

(b) $\dim(\text{Ker } G_U) = r^2$.

(c) *When $WCW^* = 0$ the general solution of (5) has the form*

$$(10) \quad Y = U_r^{-1} [\frac{1}{2}C_1 + S \quad C_2^* - (\frac{1}{2}C_1 - S)U_r^{-*}U_{m-r}^*],$$

where S is an arbitrary skew-hermitian $r \times r$ matrix.

Proof. With the partition $Y = [Y_r \quad Y_{m-r}]$ equation (5) is equivalent to the three equations

$$(11) \quad \begin{aligned} U_r Y_r + Y_r^* U_r^* &= C_1, \\ U_{m-r} Y_r + Y_{m-r}^* U_r^* &= C_2, \\ U_{m-r} Y_{m-r} + Y_{m-r}^* U_{m-r}^* &= C_3. \end{aligned}$$

Since U_r is nonsingular the first of these equations has a particular solution given by $U_r Y_r = \frac{1}{2}C_1$ and (as noted above) the general solution is $Y_r = U_r^{-1} (\frac{1}{2}C_1 + S)$ where S is an arbitrary skew-hermitian $r \times r$ matrix. The second of equations (11) now associates a unique Y_{m-r} (and hence Y) with each choice of S and leads to the formula (10). Since S has r^2 undetermined real parameters, part (b) of the lemma is established, (given the existence of solutions).

Now it is clear that the consistency of (5) is equivalent to the third equation in (11). Using the representation for Y_{m-r} implicit in (10) some calculation shows that the third equation of (11) is equivalent to $WCW^* = 0$ and the lemma is established. \square

LEMMA 2. *The general solution of (6) is given by*

$$(12) \quad X = VY + (I_n - VV^*)R$$

where R is an arbitrary $n \times m$ matrix.

Proof. See [1, p. 40], for example. \square

We remark that, in (12), $I_n - VV^*$ is the orthogonal projector onto $\text{Ker } V^* = \text{Ker } A$ along $\text{Im } V$ and, consequently, the last term of the equation is just an arbitrary

$n \times m$ matrix all of whose columns lie in $\text{Ker } A$. Introduce the fixed $n \times (n - r)$ matrix N whose columns form a basis for $\text{Ker } A$, and an alternate representation of the general solution (12) is

$$(13) \quad X = VY + NR\hat{R}$$

where \hat{R} is an arbitrary $(n - r) \times m$ matrix. But now it is clear that for each Y there is a $2m(n - r)$ real parameter family of solutions X for (6).

Combining this statement with Lemma 1 suggests the following:

THEOREM 3. *Let $A = UV^*$ be a rank factorization of $A \in \mathbb{C}^{m \times n}$ with $V^*V = I$. Then*

$$(14) \quad \text{Ker } G_A = V \text{Ker } G_U \oplus \mathcal{N}$$

where

$$\mathcal{N} = \{M \in \mathbb{C}^{n \times m} \mid M = NR\hat{R} \text{ and } \hat{R} \in \mathbb{C}^{(n-r) \times m}\}.$$

Furthermore,

$$(15) \quad \dim(\text{Ker } G_A) = r^2 + 2m(n - r).$$

Proof. We have seen that $\dim \mathcal{N} = 2m(n - r)$ and, from Lemma 1, $\dim(\text{Ker } G_U) = r^2$. Since V has full rank we also have $\dim(V \text{Ker } G_U) = r^2$. Now (13) shows that $\text{Ker } G_A = (V \text{Ker } G_U) + \mathcal{N}$ so to establish (14) and (15) it is only necessary to see that the sum is direct. But this is clear since $X \in V \text{Ker } G_U$ implies the columns of X are in $\text{Im } V$ and $X \in \mathcal{N}$ implies the columns of X are in $\text{Ker } A = \text{Ker } V^*$, while $\text{Im } V$ and $\text{Ker } V^*$ are orthogonal complements. \square

As an immediate consequence of the theorem and the first lemma we have:

COROLLARY 4. *For any $A \in \mathbb{C}^{m \times n}$ there exist solutions of the equation $AX + X^*A^* = C$ if and only if $WCW^* = 0$, and all solutions are described by combining (10) with (12) or (13).*

Example. Consider the matrix $A = \text{diag}[i, 1, 0]$. We have $m = n = 3$ and $r = 2$ so that (15) gives $\dim(\text{Ker } G_A) = 10$. It is instructive to verify this directly.

According to [3, Thm. 2] the transformation G_A (over \mathbb{R}) has eigenvalues $0, 1 + i, i, 1 - i, 2, 1, -i, 1, 0$ and another 9 zeros. Thus, the algebraic multiplicity of the zero eigenvalue is 11 and so these eigenvalues have nine associated linear elementary divisors and one elementary divisor of second degree. \square

Note that in (3) A can be replaced by $e^{i\theta}A$ where $\theta \in \mathbb{R}$, and then V by $e^{-i\theta}V$, and Theorem 3 will still hold. In particular, taking $\theta = \pi/2$ the resulting equation is $iAX - iX^*A^* = C$ or

$$(16) \quad AX - X^*A^* = \hat{C}$$

when \hat{C} is skew-hermitian. Thus, in effect, analysis of (16) is included in the results above.

3. The cases of $AX \pm X^T A^T = C$. Consider now the function $\hat{G}_A: F^{n \times m} \rightarrow F^{m \times m}$ defined by

$$(17) \quad \hat{G}_A(X) = AX + X^T A^T$$

where A is a fixed $m \times n$ matrix with elements from a field F (not of characteristic 2) and T denotes the conjugate transpose. The distinction between \hat{G}_A and G_A is that \hat{G}_A is a linear transformation on $F^{n \times m}$ over F . However the arguments of the previous section can be repeated with only minor variations. For example, the matrix S of formula (10) becomes an arbitrary $r \times r$ skew-symmetric matrix over F , which therefore

contains $\frac{1}{2}r(r-1)$ parameters and, in part (b) of Lemma 1, one obtains $\dim(\text{Ker } \hat{G}_U) = \frac{1}{2}r(r-1)$.

THEOREM 5. *Let $A = UV^T$ be a rank factorization of $A \in F^{m \times n}$ with $V^T V = I$. Then $\text{Ker } \hat{G}_A = V \text{Ker } \hat{G}_U \oplus N$ where*

$$(18) \quad \mathcal{N} = \{M \in F^{n \times m} \mid M = N\hat{R} \text{ and } \hat{R} \in F^{(n-r) \times m}\}.$$

Furthermore,

$$\dim(\text{Ker } \hat{G}_A) = \frac{1}{2}r(r-1) + m(n-r),$$

and there exist solution(s) of $G_A(X) = C$ (where $C^T = C$) if and only if $WCW^T = 0$ and all solutions are described by

$$Y = U_r^{-1} [\frac{1}{2}C_1 + S \quad C_2^T - (\frac{1}{2}C_1 - S)U_r^{-T}U_{m-r}^T]$$

where S is an arbitrary $r \times r$ skew-symmetric matrix over F and, as in (13), $X = VY + N\hat{R}$.

Example. Let $F = \mathbb{R}$ and $A = \text{diag}[1, -1, 0]$. Then $m = n = 3$ and $r = 2$, so that $\dim(\text{Ker } \hat{G}_A) = 4$, which is easily verified directly. According to Theorem 2* of [3] the eigenvalues of \hat{G}_A are 2, -2, 1, -1, and zero with multiplicity five. It follows that the zero eigenvalue of \hat{G}_A has one nonlinear elementary divisor.

Finally, consider the function $J_A: F^{n \times m} \rightarrow F^{m \times m}$ defined by

$$J_A(X) = AX - X^T A^T$$

for a fixed $A \in F^{m \times n}$. The trick used at the end of § 2 will no longer work for us. However, repetition of the same line of argument leads to:

THEOREM 6. *Let $A = UV^T$ be a rank factorization of $A \in F^{m \times n}$ with $V^T V = I$. Then $\text{Ker } J_A = V \text{Ker } J_U \oplus \mathcal{N}$ where \mathcal{N} is defined by (18). Furthermore,*

$$\dim(\text{Ker } J_A) = \frac{1}{2}r(r+1) + m(n-r)$$

and there exist solutions of $J_A(X) = C$ (where $C^T = -C$) if and only if $WCW^T = 0$ and all solutions are described by

$$Y = U_r^{-1} [\frac{1}{2}C_1 + S \quad C_2^T - (\frac{1}{2}C_1 - S)U_r^{-T}U_{m-r}^T],$$

where S is an arbitrary $r \times r$ symmetric matrix over F and, as in (13),

$$X = VY + N\hat{R}.$$

4. Eigenvectors of H -selfadjoint matrices. In this last section a problem will be outlined which was really the starting point of the authors' investigations, but serves here to illustrate one situation in which (3) arises with a rectangular coefficient matrix A . Let $A, H \in \mathbb{C}^{n \times n}$, $H^* = H$, $\det H \neq 0$, and $HA = A^*H$. In this situation A is said to be H -selfadjoint (and is, of course, hermitian if $H = I$).

It is clear that in this case A and A^* are similar so that all nonreal eigenvalues appear in conjugate pairs and each eigenvalue of such a pair has the same partial multiplicities. However, there seems to be no simple connection between the right eigenvectors associated with $\lambda \in \sigma(A)$, ($\lambda \neq \bar{\lambda}$) and those of $\bar{\lambda}$. We consider the situation in which complete sets of right Jordan chains are known for: (a) the real eigenvalues of A , and are written as the columns of a matrix V_r where $AV_r = V_r J_r$ and J_r is a Jordan matrix with real eigenvalues; and (b) the eigenvalues of A in the open upper half plane, and are written as the columns of matrix V_1 where $AV_1 = V_1 J_c$ and J_c is a Jordan matrix.

Now $J = \text{diag}[J_r, J_c, \bar{J}_c]$ is a Jordan form for A and there is a matrix V_2 such that

$$A[V_r V_1 V_2] = [V_r V_1 V_2]J$$

and $V = [V_r V_1 V_2]$ is nonsingular. The problem is to express V_2 as simply as possible in terms of V_r , and V_1 .

First recall a canonical reduction for A and H ([2, Thm. S5.1], for example). There is a $T = [T_r T_1 T_2]$, the partition being the same as that of V , such that

$$(19) \quad T^* H T = P_J \quad \text{and} \quad T^{-1} A T = J,$$

where

$$P_J = \begin{bmatrix} P_r & 0 & 0 \\ 0 & 0 & P_c \\ 0 & P_c & 0 \end{bmatrix}.$$

P_r is a block diagonal partitioned matrix like J_r , but with each diagonal block of J_r replaced by a matrix like

$$\pm \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

(and adjusted for size), and P_c bears the same relation to J_c but with all positive signs.

It is easily seen that $P_J^* = P_J$ and $P_J^2 = P_J$ and then that the first of equations (19) implies

$$(20) \quad T_1 P_c T_2^* + T_2 P_c T_1^* = H^{-1} - T_r P_r T_r^*.$$

Furthermore, T_r and T_1 are related to V_r , V_1 by transformations of the form

$$T_r = V_r Y_r, \quad T_1 = V_1 Y_1,$$

where Y_r , Y_1 are nonsingular matrices commuting with J_r , J_c , respectively (see [2, Thm. S2.2], for example).

Assume first that a Y_r is found so that (ref. (19)) $T_r^* H T_r = Y_r^* (V_r^* H V_r) Y_r = P_r$. Then (20) gives $V_1 Y_1 P_c T_2^* + T_2 P_c Y_1^* V_1^* = H^{-1} - (V_r Y_r) P_r (V_r Y_r)^*$. But it can be verified that $Y_1 P_c = P_c Y_1^T$ so that $P_c Y_1^* = \bar{Y}_1 P_c$ and $(V_1 P_c)(T_2 \bar{Y}_1)^* + (T_2 \bar{Y}_1) \times (V_1 P_c)^* = H^{-1} - (V_r Y_r) P_r (V_r Y_r)^*$. But now $V_2 = T_2 \bar{Y}_1$ defines another complete set of Jordan chains for the eigenvalues of A in the lower half plane and $(V_1 P_c) V_2^* + V_2 (V_1 P_c)^* = H^{-1} - (V_r Y_r) P_r (V_r Y_r)^*$, an equation of type (3). Furthermore, the condition $WCW^* = 0$ of Corollary 4 is guaranteed by the orthogonality properties of the root subspaces of A . Note that if A has no real eigenvalues the problem simplifies considerably; the last term of the equation goes away and the problems of finding P_r and Y_r are avoided.

REFERENCES

[1] A. BEN ISRAEL AND T. E. GREVILLE, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.
 [2] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.
 [3] O. TAUSSKY AND H. WIELANDT, *On the matrix function AX + X'A'*, Arch. Rational Mech. Anal., 9 (1962), pp. 93-96.

SIGNS OF TERMS IN A COMBINATORIAL RECURSION*

STEPHEN M. TANNY† AND MICHAEL ZUKER‡

Abstract. The sequence $\{a_n\}_{n=0}^\infty$ defined by the recursion

$$a_{n+l+1} = a_{n+l} - a_n, \quad n \geq 0,$$

with $a_0 = a_1 = \dots = a_l = 1$ can be written explicitly as

$$a_n = \sum_{j \geq 0} (-1)^j \binom{n-jl}{j},$$

where the usual conventions regarding binomial coefficients are assumed to hold. For $l = 1$ the sequence comprises alternating pairs of 1's and -1's separated by a 0, while in the case $l = 2$ it has been shown via fundamental units that $a_n = 0$ precisely for $n = 3$ and $n = 12$.

In the present paper we show that for any l the sequence $\{a_n\}$ can be decomposed into essentially equal-sized "blocks" of nonnegative and nonpositive integers containing at most one zero at the end. More precisely, we show that for fixed l the length of the blocks eventually alternates between two consecutive integers; further, we derive an asymptotic formula for the block length as a function of l .

1. Introduction. The sequence $\{a_n\}_{n=0}^\infty$ defined by the recursion

$$(1.1) \quad a_{n+l+1} = a_{n+l} - a_n, \quad n \geq 0,$$

with $a_0 = a_1 = \dots = a_l = 1$ can be written explicitly as

$$(1.2) \quad a_n = \sum_{j \geq 0} (-1)^j \binom{n-jl}{j},$$

where the usual conventions regarding binomial coefficients are assumed to hold. For $l = 1$ the sequence comprises alternating pairs of 1's and -1's separated by a 0. The case $l = 2$ has recently received some attention [1], [3], [4]. In particular, Bernstein [1] has shown that in this case $a_n = 0$ precisely for $n = 3$ and $n = 12$. Bernstein's algebraic approach via fundamental units leads to two remarkable combinatorial identities [1, p. 269] which have subsequently been extended by Carlitz [3], [4] using strictly combinatorial methods.

For $l > 2$ nothing appears to be known in general concerning the number of zeros in the sequence $\{a_n\}$. We have accumulated considerable empirical evidence which suggests that apart from the trivial zero occurring at $n = l + 1$, at most one other zero occurs. These initial investigations suggested certain further interesting properties regarding $\{a_n\}$ which are the subject of the present paper. In particular the sequence $\{a_n\}$ can be decomposed into essentially equal-sized "blocks" of nonnegative and nonpositive integers containing at most one zero at the end. More precisely, we show that for fixed l the length of the blocks eventually alternates between two consecutive integers; further, we derive an asymptotic formula for the block length as a function of l .

2. Basic formulation. The characteristic polynomial $f(z)$ of the recursion (1.1) is given by

$$(2.1) \quad f(z) = z^{l+1} - z^l + 1,$$

* Received by the editors March 5, 1982.

† Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. The research of this author was supported in part by the National Research Council under grant A9256.

‡ National Research Council of Canada, Ottawa, Ontario, Canada K1A 0R6.

where we assume $l \geq 2$. Since $f'(z) = z^{l-1}[(l+1)z - l]$, f and f' have no common roots and thus f has $l+1$ distinct (complex) roots z_1, z_2, \dots, z_{l+1} . The following facts are easily shown:

Fact 1. $a_n = \sum_{j=1}^{l+1} b_j z_j^n$ where the b_j are determined from the initial conditions of the recursion (1.1). The b_j satisfy the matrix equation $Z\mathbf{b} = \mathbf{1}$ where $Z = (Z_{ij})$ is an $(l+1) \times (l+1)$ matrix with $Z_{ij} = z_j^{i-1}$, \mathbf{b} is the column vector of the $b_j, j = 1, 2, \dots, l+1$ and $\mathbf{1}$ is an $(l+1)$ -vector of ones. The matrix Z is of Vandermonde type [2, p. 284] so that

$$b_j = \frac{\prod_{r \neq j} (1 - z_r)}{\prod_{r \neq j} (z_j - z_r)} \neq 0 \quad \text{for all } j.$$

Fact 2. Writing the $(l+1)$ roots $z_j = r_j e^{i\theta_j}$ with $r_j > 0$ and $0 < \theta_1 \leq \theta_2 \leq \dots \leq \theta_l < 2\pi$, we have

$$(2.2) \quad r_j^{l+1} \cos(l+1)\theta_j = r_j^l \cos l\theta_j - 1,$$

$$(2.3) \quad r_j \sin(l+1)\theta_j = \sin l\theta_j,$$

$$(2.4) \quad r_j^{2l} ((r_j \cos \theta_j - 1)^2 + r_j^2 \sin^2 \theta_j) = 1.$$

Fact 3. From (2.3) it is immediate that the θ_j 's are distinct. Rewriting (2.4) we obtain $\cos \theta_j = \frac{1}{2}r_j + 1/2r_j - 1/2r_j^{2l+1}$, which can be used to show that for $0 < \theta_1 < \dots < \theta_l < \pi$ we have $r_1 > r_2 > \dots > r_l > 0$, where $l = [(l+1)/2]$.

Fact 4. From the preceding we can write $a_n = Cr_1^n \cos(n\theta_1 + \phi) + O(r_2^n)$ where C is real and nonzero and ϕ is a phase angle, and both depend upon b_1 .

3. Block decomposition. For an arbitrary sequence $\{s_n\}$ define a *block* of length $n+1$ as a segment $\{s_m, s_{m+1}, \dots, s_{m+n}\}$ such that

- (i) $s_m \neq 0$;
- (ii) s_{m+1}, \dots, s_{m+n} all have the same sign as s_m or are zero;
- (iii) $s_{m-1} = 0$ or $\text{sgn } s_{m-1} = -\text{sgn } s_m$;
- (iv) $s_{m+n+1} \neq 0$ and $\text{sgn } s_{m+n+1} = -\text{sgn } s_m$.

Obviously any real sequence can be decomposed into blocks in a unique way. It is easy to verify that the recursion (1.1) with the given initial conditions has a block decomposition in which all blocks have length at least $l+1$. In fact, we now show that the blocks eventually all have length either $[\pi/\theta_1]$ or $[\pi/\theta_1] + 1$, where θ_1 is defined in § 2 as the smallest argument associated with a root of (2.1).

Recall that

$$(3.1) \quad a_n = Cr_1^n \cos(n\theta_1 + \phi) + O(r_2^n).$$

Since $r_1 > r_2$, the first term dominates in determining the sign of a_n , at least for n large enough and $n\theta_1 + \phi$ not too close to $-\pi/2$ or $\pi/2$ modulo π . To make this precise, suppose π/θ_1 is not an integer. Then we can choose d small enough so that $[(\pi + d)/\theta_1] = [(\pi - d)/\theta_1]$. Thus, if $n\theta_1 + \phi \in (-\pi/2 + d/2, \pi/2 - d/2)$ modulo π then $\cos[n\theta_1 + \phi] \geq \sin d/2 > 0$. If we choose n so large that $|O(r_2^n)| < (\frac{1}{2} \sin d/2)(r_1)^n$, then it follows that the sign of a_n is dominated by the first term in (3.1), so long as $n\theta_1 + \phi \in (-\pi/2 + d/2, \pi/2 - d/2)$ (modulo π). For convenience, in what follows we drop the qualifier “(modulo π)”.

Since at most $[(\pi + d)/\theta_1] + 1$ consecutive values of n can occur so that $n\theta_1 + \phi$ remains inside the interval $(-\pi/2 - d/2, \pi/2 + d/2)$, it follows that the lengths of blocks containing no negative terms in the block decomposition of $\{a_n\}$ are eventually no more than $[(\pi + d)/\theta_1] + 1 = [\pi/\theta_1] + 1$. On the other hand, there are at least

$[(\pi - d)/\theta_1]$ consecutive values of n such that $n\theta_1 + \phi$ is in the interval $(-\pi/2 + d/2, \pi/2 - d/2)$ so that the block lengths are eventually at least $[(\pi - d)/\theta_1] = [\pi/\theta_1]$. Thus, for n sufficiently large, all the blocks of positive a_n have lengths either $[\pi/\theta_1]$ or $[\pi/\theta_1] + 1$. It is clear that the analogous argument holds for blocks containing negative a_n . On the other hand, suppose that $\pi/\theta_1 = P$ an integer. Then it is a straightforward counting argument to verify that if $d < \frac{1}{2}\theta_1$ then it is not possible to have $p + 2$ consecutive values of n such that $n\theta_1 + \phi$ is always in $(-\pi/2 - d/2, \pi/2 + d/2)$ or in $(\pi/2 - d/2, \pi) \cup (-\pi, -\pi/2 + d/2)$ but not in both. Thus once again we find that the block lengths can only be $[\pi/\theta_1]$ or $[\pi/\theta_1] + 1$.

THEOREM 3.1. *The sequence $\{a_n\}$ defined by (1.1) and the initial conditions following has a block decomposition in which all but a finite number of the blocks have length $[\pi/\theta_1]$ or $[\pi/\theta_1] + 1$, where θ_1 is defined in § 2. Further, at least the first $[\pi/\theta_1]$ elements in each such block are nonzero.*

Table 1 gives the block lengths for $l = 2, \dots, 19$.

TABLE 1

L	Block lengths
2	4, 5
3	5, 6
4	7, 8
5	8, 9
6	10, 11
7	11, 12
8	12, 13
9	14, 15
10	15, 16
11	16, 17
12	17, 18
13	19, 20
14	20, 21
15	21, 22
16	23, 24
17	24, 25
18	25, 26
19	27, 28

4. Asymptotic estimate of block length. In order to derive an asymptotic estimate of π/θ_1 we first find a simpler equation which generates the arguments $\theta_i, i = 1, 2, \dots, l$ as roots. We then approximate θ_1 from this equation.

Since $r_j \neq 0$ for any root of (2.1), (2.3) shows that $\sin(l + 1)\theta_j = 0$ if and only if $\sin l\theta_j = 0$. But $\sin(l + 1)\theta_j = \sin l\theta_j \cos \theta_j + \cos l\theta_j \sin \theta_j$, so $\sin l\theta_j = 0$ means that $\cos l\theta_j \sin \theta_j = 0$. In such a case, since $0 < \theta_j < \pi$, $\sin \theta_j > 0$ while $\cos^2 l\theta_j = 1 - \sin^2 l\theta_j = 1$ so that we have a contradiction. Hence both $\sin l\theta_j$ and $\sin(l + 1)\theta_j$ are never zero for any root θ_j . Thus we can replace r_j in (2.2):

$$(4.1) \quad \left(\frac{\sin l\theta_j}{\sin(l + 1)\theta_j}\right)^{l+1} \cos(l + 1)\theta_j = \left(\frac{\sin l\theta_j}{\sin(l + 1)\theta_j}\right)^l \cos l\theta_j - 1,$$

which can be simplified to

$$(4.2) \quad [\sin(l + 1)\theta_j]^{l+1} = \sin \theta_j (\sin l\theta_j)^l.$$

Note that (2.3) implies that $\sin l\theta_j$ and $\sin (l + 1)\theta_j$ have the same sign for any root θ_j , while (4.2) shows that both cannot be negative (since $\sin \theta_j > 0$). Hence $\sin l\theta_j$ and $\sin (l + 1)\theta_j$ are positive for all j .

Now, consider the function

$$(4.3) \quad H(\theta) = [\sin (l + 1)\theta]^{l+1} - \sin \theta(\sin l\theta)^l.$$

We want to find all the roots α_j of $H(\theta)$ which satisfy the additional requirement that

$$(4.4) \quad \min (\sin l\alpha_j, \sin (l + 1)\alpha_j) > 0.$$

Define $J_k = [2\pi(k - 1)/l, \pi(2k - 1)/(l + 1)] = [x_k, y_k]$, $k = 1, 2, \dots, [(l + 1)/2]$. Then $H(x_1) = 0$ and $H(x_k) > 0$ for $k > 1$ while $H(y_k) < 0$ for all k , so it follows that for $k \geq 2$ there is at least one root α_k of $H(\theta)$ in (x_k, y_k) and further, α_k satisfies (4.4). For $k = 1$ a root α_1 satisfying (4.4) can be found in $(0, b_1)$ since near $\theta = 0$, $H(\theta) \sim [(l + 1)^{l+1} - l^l]\theta^{l+1}$ which is positive. Thus there are at least $[(l + 1)/2]$ roots of (4.3) satisfying (4.4), with at least one root in each interval J_k . But any such root satisfies (2.2)–(2.4) and hence is a root of (2.1), and we know that there can only be $[(l + 1)/2]$ such roots. Thus, the pigeonhole principle implies that the α_k are just the roots θ_j identified in § 2.

The root θ_1 is the smallest and hence is identified with α_1 , so $\theta_1 \in J_1 = [0, \pi/(l + 1)]$. It is easy to show that $\pi/2(l + 1) < \theta_1 < \pi/(l + 1)$; in fact, only slightly more work yields $\pi/(1 + \varepsilon)(l + 1) < \theta_1 < \pi/(l + 1)$ for $\varepsilon = \varepsilon(l) < 1$, and $\varepsilon(l) \rightarrow 0$ as $l \rightarrow \infty$. This can be seen by substituting $\theta = \pi/(1 + \varepsilon)(1 + l)$ in (4.3) and showing $H(\theta) > 0$. Thus we can write $\theta_1 = (\pi - a)/(l + 1)$, where $a \equiv a(l) \rightarrow 0$ as $l \rightarrow \infty$. Substituting θ_1 in (4.3) gives

$$(4.5) \quad (\sin a)^{l+1} = \sin \left(\frac{\pi - a}{l + 1} \right) \left[\sin \left(a + \frac{\pi - a}{l + 1} \right) \right]^l.$$

For a small enough, (4.5) can be approximated as

$$a^{l+1} \approx \left(\frac{\pi - a}{l + 1} \right) \left(a + \frac{\pi - a}{l + 1} \right)^l,$$

or

$$(4.6) \quad 1 \approx \frac{b}{l + 1} \left(1 + \frac{b}{l + 1} \right)^l, \quad b = \frac{\pi}{a} - 1.$$

But $b \rightarrow \infty$ as $l \rightarrow \infty$, and (4.6) implies that $b/(l + 1) \rightarrow 0$ as $l \rightarrow \infty$, so we conclude from (4.6) that

$$(4.7) \quad l \approx b e^b.$$

Now $b e^b$ is a monotone function of b , so that it has an inverse. Clearly $\log l - \log \log l < b < \log l$ for large l , so we have that $\pi/(\log l + 1) < a < \pi/(\log l - \log \log l + 1)$. Thus, substituting for a we obtain that, for l big enough,

$$(4.8) \quad \theta_1 = \frac{\pi}{l + 1} \left[1 - \frac{1}{\log l} + o \left(\frac{1}{\log l} \right) \right]$$

and

$$(4.9) \quad \frac{\pi}{\theta_1} = l \left(1 + \frac{1}{\log l} \right) + o \left(\frac{l}{\log l} \right).$$

REFERENCES

- [1] L. BERNSTEIN, *Zeros of the function $f(n) = \sum_i (-1)^i \binom{n-2i}{i}$* , J. Number Theory, 6 (1974), pp. 264–270.
- [2] G. BIRKHOFF AND S. MACLANE, *A Survey of Modern Algebra*, Macmillan, New York, 1965.
- [3] L. CARLITZ, *Some combinatorial identities of Bernstein*, SIAM J. Math. Anal., 9 (1978), pp. 65–75.
- [4] ———, *Recurrences of the third order and related combinatorial identities*, Fibonacci Quart., 16 (1978), pp. 11–18.
- [5] K. CHANDRASEKHARAN, *Introduction to Analytic Number Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1968.
- [6] S. M. TANNY AND M. ZUKER, *Analytic methods applied to a sequence of binomial coefficients*, Discrete Math., 24 (1978), pp. 299–310.

THE BANZHAF INDEX FOR MULTICANDIDATE PRESIDENTIAL ELECTIONS*

EDWARD M. BOLGER†

Abstract. John Banzhaf III [Villanova Law Rev., 13 (1968), pp. 304–332] introduced a measure of voting power for a two-candidate United States presidential election conducted under an electoral college system. This measure indicated that an individual voter in a large state has more voting power than an individual in a small state. In this paper we generalize the Banzhaf index of voting power to voting situations in which there are more than two candidates with one to be elected. We then measure the voting power of an individual voter in a three-candidate and in a four-candidate presidential election under an electoral college system. It is found that there is slightly more bias in favor of an individual in a large state. In the last section, we conjecture what the results will be for an r -candidate United States presidential election.

Key words. voting games, Banzhaf index, electoral college, pivot

1. Introduction. The Banzhaf power index [1] assigns a “value” to each player in a simple game. This value measures the voting power of the players. Simple games serve as models for elections in which a group of voters is to elect one of two candidates to an office. We shall consider elections in which a group of voters is to elect one of several candidates. We shall model such elections by “voting games among r candidates”.

DEFINITION 1. Let N be a finite set. Let S_1, S_2, \dots, S_r be subsets of N . $\{S_1, S_2, \dots, S_r\}$ is called a *partition of N into r subsets* if $S_1 \cup S_2 \cup \dots \cup S_r = N$ and $S_i \cap S_j = \emptyset$ for $i \neq j$.

If precisely one of r candidates is to be elected, then the voters must partition themselves among the r candidates. The voting rules may be specified by stating, for each partition $\{S_1, S_2, \dots, S_r\}$ of the voters, whether each S_j is a winning or losing coalition with respect to this partition. We formalize this in the following definition.

DEFINITION 2. Let N be a finite set. A *voting game on N among r candidates* is a listing of all the partitions of N which consist of r or fewer subsets together with a rule which specifies for each subset in such a partition whether that subset is winning or losing with respect to the given partition. The elements of N are called the voters; subsets of N are called coalitions of voters.

For example, if $N = \{1, 2, \dots, n\}$, then the *simple plurality game on N* among r candidates is determined by the rule which specifies that a subset T of N is winning with respect to a partition \mathcal{P} of which T is a member if and only if T contains more voters than any other member of \mathcal{P} .

We now generalize the concept of pivot set to voting games among r candidates. Recall that for a simple game on N , a subset T of N is said to be a pivot set for voter i if T wins and $T - \{i\}$ loses.

DEFINITION 3. Let $i \in N$. Let $\mathcal{P} = \{S_1, S_2, \dots, S_p\}$ be a partition of the voters into r or fewer subsets, one of which may be empty. Suppose $i \in S_j$. Consider a partition \mathcal{P}' obtained by removing player i from S_j and placing player i in some other member S_k of \mathcal{P} . The mapping $\alpha_{ik}: \mathcal{P} \rightarrow \mathcal{P}'$ defined by

$$\begin{aligned}\alpha_{ik}(S_j) &= S_j - \{i\}, \\ \alpha_{ik}(S_k) &= S_k \cup \{i\}, \\ \alpha_{ik}(S_t) &= S_t \quad \text{for } t \neq j, k\end{aligned}$$

* Received by the editors September 28, 1981, and in final form December 7, 1982.

† Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056.

is called a *move* for player i . Such a move is called a *pivot move* if S_j wins with respect to \mathcal{P} and $S_j - \{i\}$ loses with respect to \mathcal{P}' .

For example, if $N = \{1, 2, 3, 4\}$ and if $r = 3$, then the pivot moves for player 1 in the simple plurality game are:

$$\begin{aligned} &\{\{1, 2, 3\}, \{4\}\} \rightarrow \{\{2, 3\}, \{1, 4\}\} \\ &\{\{1, 2, 4\}, \{3\}\} \rightarrow \{\{2, 4\}, \{1, 3\}\} \\ &\{\{1, 3, 4\}, \{2\}\} \rightarrow \{\{3, 4\}, \{1, 2\}\} \\ &\{\{1, 2\}, \{3\}, \{4\}\} \begin{array}{l} \rightarrow \{\{2\}, \{1, 3\}, \{4\}\} \\ \succ \{\{2\}, \{3\}, \{1, 4\}\} \end{array} \\ &\{\{1, 3\}, \{2\}, \{4\}\} \begin{array}{l} \rightarrow \{\{3\}, \{1, 2\}, \{4\}\} \\ \succ \{\{3\}, \{2\}, \{1, 4\}\} \end{array} \\ &\{\{1, 4\}, \{2\}, \{3\}\} \begin{array}{l} \rightarrow \{\{4\}, \{1, 2\}, \{3\}\} \\ \succ \{\{4\}, \{2\}, \{1, 3\}\} \end{array} \end{aligned}$$

DEFINITION 4. Let v be a voting game on N among r candidates. We let $p(i, v)$ denote the number of pivot moves for player i in the game v .

DEFINITION 5. If $p(i, v) = 0$, we call player i a *dummy* in the game v .

DEFINITION 6. If there are no dummies in the game v , then we define the (normalized) *Banzhaf value* for player i in the game v to be the ratio

$$\frac{p(i, v)}{p(1, v) + p(2, v) + \dots + p(n, v)}$$

where n is the cardinality of N . On the other hand, if there are dummies in v , we assign each of them Banzhaf value equal to 0 and “remove” all of them from the game before assigning Banzhaf values to the other players. (Formally, if $\mathcal{P} = \{S_1, S_2, \dots, S_r\}$ is a partition of N and if D is the set of dummies in v , then $\{S_1 - D, \dots, S_r - D\}$ is a partition of $N - D$. We define an induced game on $N - D$ by saying that $S_j - D$ is winning with respect to the above partition of $N - D$ if and only if S_j is winning with respect to \mathcal{P} . Then, for $i \in N - D$, we define the Banzhaf value of player i for the game v to be the Banzhaf value of player i for the induced game on $N - D$.)

Remark. The above (somewhat cumbersome) definition assures that the addition (or removal) of dummies to a game will not affect the Banzhaf values of the other players. If there are only two candidates, it is not necessary to remove the dummies since the removal of each dummy reduces each $p(i, v)$ by a factor of 0.5, thus leaving the ratio unchanged. However, there are examples of voting games among 4 candidates in which the ratio would be affected by the removal of a dummy.

Example. For the example on the previous page, $p(i, v) = 9$ for all i and thus the Banzhaf value for each voter is 0.25.

In the remaining sections, we shall be primarily concerned with presidential elections conducted under an electoral college system. If we take N to be the set of eligible voters, there will not be any dummies.

2. The number of pivot moves for a voter in a plurality election. In order to determine the number of pivot moves for a voter in a presidential election conducted under an electoral college system, it will be necessary to calculate the number of pivot moves for a voter in a simple plurality (i.e., the candidate, if any, with the most votes wins the election—a tie produces no winner) game. Let $p(1; n, r)$ be the number of pivot moves for voter 1 in the simple plurality game with n voters and r candidates. We consider first the case where $r = 3$ and $n = 6m + 3$. (It turns out that there are

slightly different formulas for $p(1; n, 3)$ for $n = 6m, 6m + 1, 6m + 2, 6m + 3, 6m + 4, 6m + 5$.)

LEMMA 1.

$$p(1; 6m + 3, 3) = 2 \sum_{j=1}^{m+1} \binom{6m+2}{2m+j} \binom{4m+2-j}{2m+j} + \sum_{j=1}^m \binom{6m+2}{2m+1+j} \binom{4m+1-j}{2m+j}.$$

Proof. Voter 1 has exactly one pivot move for each partition in which voter 1's coalition contains exactly two more voters than the next largest coalition. The pivot move is the mapping which moves voter 1 to that next largest coalition, thus creating a "tie". Voter 1 has exactly two pivot moves for each partition in which voter 1's coalition contains exactly one more voter than the next largest coalition. One such pivot move is the mapping which moves player 1 to the next largest coalition; the other is the mapping which moves player 1 to the smallest coalition. To visualize these, it is helpful to represent each such partition by an ordered triple which lists, in descending order, the sizes of the coalitions in the partition. We refer to such an ordered triple as the *type* of the partition. For $n = 6m + 3$, the types which yield precisely one pivot move for voter 1 are $(3m + 2, 3m, 1), (3m + 1, 3m - 1, 3), \dots, (2m + 3, 2m + 1, 2m - 1)$. The types which yield precisely two pivot moves for voter 1 are $(3m + 2, 3m + 1, 0), (3m + 1, 3m, 2), \dots, (2m + 2, 2m + 1, 2m)$.

For $1 \leq j \leq m$, the number of partitions of type $(2m + j + 2, 2m + j, 2m - 2j + 1)$ in which voter 1 belongs to the largest coalition is

$$\binom{6m+2}{2m+j+1} \binom{4m+1-j}{2m+j}.$$

Each of these partitions yields one pivot move for voter 1.

For $1 \leq j \leq m + 1$, the number of partitions of type $(2m + j + 1, 2m + j, 2m - 2j + 2)$ in which voter 1 belongs to the largest coalition is

$$\binom{6m+2}{2m+j} \binom{4m+2-j}{2m+j}.$$

Each of these partitions yields two pivot moves for voter 1.

It follows that the total number of pivot moves for voter 1 is

$$2 \sum_{j=1}^{m+1} \binom{6m+2}{2m+j} \binom{4m+2-j}{2m+j} + \sum_{j=1}^m \binom{6m+2}{2m+j+1} \binom{4m+1-j}{2m+j}.$$

If $r = 4$, we consider the case where $n = 8m$.

LEMMA 2.

$$\begin{aligned} p(1; 8m, 4) &= \frac{1}{2} \sum_{j=1}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m-j} + \binom{8m-1}{4m} \\ &\quad + \sum_{j=1}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \sum_{k=1}^{2j-1} \binom{4m-2j}{2m-j+k} \\ &\quad + \frac{3}{2} \sum_{j=0}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j} \binom{4m-1-2j}{2m+j} \\ &\quad + 3 \sum_{j=1}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j} \sum_{k=0}^{2j-1} \binom{4m-1-2j}{2m-j+k} \end{aligned}$$

where the binomial coefficient $\binom{n}{j}$ is understood to equal 0 if $j > n$.

Proof. Let \mathcal{P} be a partition of the voters into four or fewer subsets. We associate with such partition a 4-tuple which lists in descending order the sizes of the sets in the partition.

For $1 \leq j \leq 2m$ and $0 \leq k \leq \min\{2j-1, 2m-j\}$, a partition of type $(2m+j+1, 2m+j-1, 2m-j+k, 2m-j-k)$ will yield one pivot move for a voter in the largest coalition (move that voter to the second largest coalition) unless $2m+j-1 = 2m-j+k$ in which case such a partition will yield two pivot moves for that voter. (Note that our choice of $n = 8m$ does not allow $2m+j-1 = 2m-j+k = 2m-j-k$.)

We must, however, exercise some care in counting the number of partitions which yield one or two pivot moves. The number of partitions of type $(2m+j+1, 2m+j-1, 2m-j+k, 2m-j-k)$ in which voter 1 belongs to the largest coalition is

$$\binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m-j+k}$$

provided the numbers $2m+j-1, 2m-j+k, 2m-j-k$ are all distinct. These numbers will all be distinct unless $k = 0$ or $k = 2j-1$. If two of these three numbers are equal and positive, then this number of partitions is only

$$\frac{1}{2} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m-j+k}$$

since the order in which the sets of the partition are listed is irrelevant (e.g. the partition $\{\{1, 2, 3\}, \{4, 5\}, \{6, 7\}, \{8\}\}$ is the same as the partition $\{\{1, 2, 3\}, \{6, 7\}, \{4, 5\}, \{8\}\}$). Thus, for $0 < k < 2j-2$, each partition of type $(2m+j+1, 2m+j-1, 2m-j+k, 2m-j-k)$ yields one pivot move for a voter in the largest coalition, and there are

$$\binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m-j+k}$$

such partitions. Furthermore, each partition of type $(2m+j+1, 2m+j-1, 2m-j, 2m-j)$ yields one pivot move, and for $j < 2m$ there are

$$\frac{1}{2} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m-j}$$

such partitions. For $k = 2j-1$, each partition of type $(2m+j+1, 2m+j-1, 2m+j-1, 2m-3j+1)$ yields two pivot moves for a voter in the largest coalition, and there are

$$\frac{1}{2} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m+j-1}$$

such partitions. We have so far counted

$$\begin{aligned} & \frac{1}{2} \sum_{j=1}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m-j} + \binom{8m-1}{4m} \binom{4m-1}{4m-1} \\ & + \sum_{j=1}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \sum_{k=1}^{2j-2} \binom{4m-2j}{2m-j+k} \\ & + 2 \cdot \frac{1}{2} \sum_{j=1}^{2m-1} \binom{8m-1}{2m+j} \binom{6m-1-j}{2m+j-1} \binom{4m-2j}{2m+j-1} \end{aligned}$$

pivot moves for voter 1. We observe that the last two terms can be combined by letting k run from 1 to $2j-1$.

For $0 \leq j \leq 2m - 1$ and $0 \leq k \leq \min \{2j, 2m - 1 - j\}$, a partition of type $(2m + j + 1, 2m + j, 2m - j + k, 2m - j - k - 1)$ yields three pivot moves for a voter in the largest coalition by moving that voter to each of the other three coalitions (one of which may be empty). Here, too, we must be careful when counting the number of partitions. The number of partitions of type $(2m + j + 1, 2m + j, 2m - j + k, 2m - j - k - 1)$ in which voter 1 belongs to the largest coalition is

$$\binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \binom{4m - 1 - 2j}{2m - j + k}$$

provided the numbers $2m + j, 2m - j + k, 2m - j - k - 1$ are all distinct, which they will be unless $k = 2j$. If $k = 2j$, the number of partitions is

$$\frac{1}{2} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \binom{4m - 1 - 2j}{2m + j}.$$

The partitions considered in the preceding paragraph give

$$3 \cdot \frac{1}{2} \sum_{j=0}^{2m-1} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \binom{4m - 1 - 2j}{2m + j} + 3 \sum_{j=1}^{2m-1} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \sum_{k=0}^{2j-1} \binom{4m - 1 - 2j}{2m - j + k}$$

additional pivot moves for voter 1. This completes the proof of Lemma 2.

For computational purposes, it will be better to rewrite $p(1; 8m, 4)$.

COROLLARY.

$$p(1; 8m, 4)$$

$$\begin{aligned} &= \frac{1}{2} \sum_{j=1}^{2m-1} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j - 1} \binom{4m - 2j}{2m - j} + \binom{8m - 1}{4m} \\ &+ \sum_{j=1}^{[(2m+1)/3]} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j - 1} \sum_{k=1}^{2j-1} \binom{4m - 2j}{2m - j + k} \\ &+ \sum_{j=[(2m+1)/3]+1}^{2m-1} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j - 1} \sum_{k=1}^{2m-j} \binom{4m - 2j}{2m - j + k} \\ &+ \frac{3}{2} \sum_{j=0}^{[(2m-1)/3]} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \binom{4m - 1 - 2j}{2m + j} \\ &+ 3 \sum_{j=1}^{[2m/3]} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \sum_{k=0}^{2j-1} \binom{4m - 1 - 2j}{2m - j + k} \\ &+ 3 \sum_{j=[2m/3]+1}^{2m-1} \binom{8m - 1}{2m + j} \binom{6m - 1 - j}{2m + j} \sum_{k=0}^{2m-1-j} \binom{4m - 1 - 2j}{2m - j + k} \end{aligned}$$

where $[x]$ denotes the greatest integer in x .

Proof. The proof follows easily from the assumption that $\binom{n}{j} = 0$ if $j > n$ and from the inequalities:

$$\begin{aligned} 2m - j + k &\leq 4m - 2j \Leftrightarrow k \leq 2m - j, \\ 2m - j &\leq 2j - 1 \Leftrightarrow 2m + 1 \leq 3j, \\ 2m + j &\leq 4m - 1 - 2j \Leftrightarrow 3j \leq 2m - 1, \\ 2m - j + k &\leq 4m - 1 - 2j \Leftrightarrow k \leq 2m - 1 - j, \\ 2m - 1 - j &\leq 2j - 1 \Leftrightarrow 2m \leq 3j. \end{aligned}$$

3. Three- and four-candidate presidential elections conducted under an electoral college system. In the United States, presidents are elected via the electoral college system. Each state is assigned a number of “electoral votes”. After the voters in each state go to the polls on election day and some time after the results are compiled, the simple plurality winner within each state is usually awarded all of that state’s electoral votes. If a candidate receives more than half of the total number of electoral votes, then he or she is declared the winner. If no candidate receives more than half of the electoral votes, then the House of Representatives chooses the president.

In order to obtain a workable model, we shall consider a simplified version of a presidential election. We make the following assumptions:

Assumption 1. Each state is assigned a number of electoral votes.

Assumption 2. There are precisely r candidates.

Assumption 3. The candidate receiving a simple plurality of the votes cast within a state automatically receives all of that state’s electoral votes. In case of a tie, no candidate receives any of the state’s electoral votes.

Assumption 4. In order to win the election, a candidate must receive at least q electoral votes. This number q is called the quota. If no candidate receives at least q votes, then there is no winner.

If we let N be the set of all United States citizens of voting age, then we can model this simplified presidential election as a voting game on N among r candidates.

For simplicity, we first consider the case where there are three candidates and only two states, each with one electoral vote and voting populations $n_1 = 6m_1 + 3$ and $n_2 = 6m_2 + 3$. To win the election, a candidate must win in each state. We shall calculate the number of pivot moves for voter 1 of state 1 in the presidential election. First we need a preliminary result.

LEMMA 3. *Let $W(m)$ be the number of ways of distributing $6m + 3$ voters among three “boxes” B_1, B_2, B_3 in such a way that B_1 contains more voters than either B_2 or B_3 . Then*

$$(1) \quad W(m) = 3^{6m+2} - \frac{1}{3} \binom{6m+3}{2m+1} \binom{4m+2}{2m+1} - \sum_{j=1}^m \binom{6m+3}{2m+1+j} \binom{4m+2-j}{2m+1+j}.$$

Proof. The total number of ways of distributing $6m + 3$ voters among three distinguishable boxes is 3^{6m+3} . The number of ways which produce a tie (i.e. in which the two boxes with the largest number of voters in fact contain the same number of voters) is

$$T(m) = \binom{6m+3}{2m+1} \binom{4m+2}{2m+1} + 3 \sum_{j=1}^m \binom{6m+3}{2m+1+j} \binom{4m+2-j}{2m+1+j}.$$

The rest have the property that one of the boxes contains more voters than either of the other two boxes; in one-third of these, B_1 contains more voters than either of the other boxes. Thus,

$$W(m) = \frac{1}{3}[3^{6m+3} - T(m)].$$

DEFINITION 7. Let $\mathcal{P} = \{S_1, S_2, S_3\}$ be a partition which yields at least one pivot move for voter 1 in a three-candidate presidential election with two states. Assume the subsets S_1, S_2, S_3 are listed in descending order of the number of voters of state 1 in these subsets. (It is not hard to see that if \mathcal{P} is a pivot partition for voter 1 and if $n_1 = 6m_1 + 3$, then there is a unique such ordering of the sets in \mathcal{P} .) We can then represent this partition by an ordered triple of ordered pairs $((v_{11}, v_{12}), (v_{21}, v_{22}), (v_{31}, v_{32}))$ where v_{ij} is the number of voters from state j in S_i . We call this triple the *type* of the partition. (Note that the term “type” is defined differently in the proof of Lemma 1.)

LEMMA 4. *If there are only two states each with one electoral vote, if there are three candidates, and if the number of voters in state i is $6m_i + 3$ where m_i is a positive integer, then the number of pivot moves for voter 1 of state 1 in the presidential election is $P(m_1)W(m_2)$ where $P(m_1)$ is the number of pivot moves for voter 1 in the plurality election within state 1 and $W(m)$ is given by (1) above. (Here we are writing $P(m_1)$ instead of the more cumbersome $p(1; 6m_1 + 3, 3)$.)*

Proof. Let $\mathcal{P} = \{S_1, S_2, S_3\}$ be a partition of the voters which yields a pivot move for voter 1 and assume S_1, S_2, S_3 are listed in order of decreasing size of the number of voters in state 1. Then voter 1 must pivot in the plurality election within state 1 and voter 1’s candidate must win in the plurality election within state 2. Each such partition has type $((v_{11}, v_{12}), (v_{21}, v_{22}), (v_{31}, v_{32}))$ in which

- (a) $1 \in S_1,$
- (b) $v_{21} + 1 \leq v_{11} \leq v_{21} + 2,$
- (c) $v_{21} \geq v_{31},$
- (d) $v_{12} > \max \{v_{22}, v_{32}\}.$

Such a partition yields precisely one pivot move for voter 1 if the triple (v_{11}, v_{21}, v_{31}) is one of $(3m_1 + 2, 3m_1, 1), (3m_1 + 1, 3m_1 - 1, 3), \dots, (2m_1 + 3, 2m_1 + 1, 2m_1 - 1)$, whereas such a partition yields two pivot moves for voter 1 if (v_{11}, v_{21}, v_{31}) is one of $(3m_1 + 2, 3m_1 + 1, 0), (3m_1 + 1, 3m_1, 2), \dots, (2m_1 + 2, 2m_1 + 1, 2m_1)$.

To form all pivot partitions for voter 1 we first partition the voters of state 1 into partitions of type (v_{11}, v_{21}, v_{31}) satisfying (a), (b), (c) above, and then we distribute the voters of state 2 among the subsets of these partitions to form partitions of all the voters in such a way that these partitions will be of types $((v_{11}, v_{12}), (v_{21}, v_{22}), (v_{31}, v_{32}))$ which also satisfy (d).

Recall that the number of partitions of the voters of state 1 into partitions which yield two pivot moves for voter 1 in the plurality election within state 1 is

$$\sum_{j=1}^{m_1+1} \binom{6m_1+2}{2m_1+j} \binom{4m_1+2-j}{2m_1+j}.$$

After forming such a partition, we then wish to distribute the voters of state 2 among the three subsets in such a way that voter 1’s subset will receive more voters from state 2 than either of the other subsets. This can be done in $W(m_2)$ ways. Thus, the

number of partitions which yield two pivot moves for voter 1 is

$$\sum_{j=1}^{m_1+1} \binom{6m_1+2}{2m_1+j} \binom{4m_1+2-j}{2m_1+j} \cdot W(m_2).$$

Similarly, the number of partitions of the voters which yield one pivot move for voter 1 is

$$\sum_{j=1}^{m_1} \binom{6m_1+2}{2m_1+1+j} \binom{4m_1+1-j}{2m_1+j} \cdot W(m_2).$$

The result follows immediately.

Using Lemma 4, we obtain the following theorem.

THEOREM 1. *If there are only two states each with one electoral vote, if state i has $6m_i + 3$ voters, and if there are three candidates, then the ratio of the number of pivot moves for a voter of state 1 to the number of pivot moves for a voter of state 2 in the presidential election equals*

$$(2) \quad \frac{P(m_1)W(m_2)}{P(m_2)W(m_1)}.$$

Remark. Each of the factors in (2) becomes quite large even for moderate values of m_1 or m_2 . For example, $P(1) = 1540$, $W(1) = 5371$. Thus it seems desirable to find an alternate method for calculating the ratio $P(m)/W(m)$. We write

$$\sum_{j=1}^{m+1} \binom{6m+2}{2m+j} \binom{4m+2-j}{2m+j} = \frac{(6m+2)!}{[(2m+1)!]^2(2m)!} \sum_{j=1}^{m+1} \frac{[(2m+1)!]^2(2m)!}{[(2m+j)!]^2(2m+2-2j)!}$$

and then let

$$S_1(m) = \sum_{j=1}^{m+1} \frac{[(2m+1)!]^2(2m)!}{[(2m+j)!]^2(2m+2-2j)!}.$$

Similarly,

$$\sum_{j=1}^m \binom{6m+2}{2m+1+j} \binom{4m+1-j}{2m+j} = \frac{(6m+2)!}{(2m+2)!(2m+1)!(2m-1)!} S_2(m)$$

where

$$S_2(m) = \sum_{j=1}^m \frac{(2m+2)!(2m+1)!(2m-1)!}{(2m+1+j)!(2m+j)!(2m+1-2j)!},$$

and lastly,

$$\sum_{j=1}^m \binom{6m+3}{2m+1+j} \binom{4m+2-j}{2m+1+j} = \frac{(6m+3)!}{[(2m+2)!]^2(2m-1)!} S_3(m)$$

where

$$S_3(m) = \sum_{j=1}^m \frac{[(2m+2)!]^2(2m-1)!}{[(2m+1+j)!]^2(2m+1-2j)!}.$$

Then

$$\frac{P(m)}{W(m)} = \frac{2S_1(m) + \frac{m}{m+1}S_2(m)}{\frac{[(2m+1)!]^2(2m)!3^{6m+2}}{(6m+2)!} - 1 - \frac{m(6m+3)}{(m+1)(2m+2)}S_3(m)}$$

For $m = 10$, the above ratio is approximately 0.0985 whereas for $m = 16$, this ratio is approximately 0.0776. Thus, for a nation of two states with voting populations 63 and 99, the ratio of the number of pivot moves of a voter in the smaller state to that of a voter in the larger state is approximately $(0.0985) \div (0.0776) \approx 1.27$. This ratio is fairly close to $\sqrt{99} \div \sqrt{63} \approx 1.25$, the approximate ratio for a two-candidate presidential election in a two-state nation (see Banzhaf [1] or Lucas [4]). (The reader is reminded that the ratio of the numbers of pivot moves is the same as the ratio of the Banzhaf values.) This leads us to wonder if this “square-root” law holds in general for a three-candidate election in a two-state nation. To see that this is indeed the case, we use Stirling’s formula to approximate the factorials. We get, for large m ,

$$\frac{[(2m+1)!]^2(2m)!3^{6m+2}}{(6m+2)!} \approx \frac{4\pi m}{\sqrt{3}}$$

and

$$\begin{aligned} \frac{P(m)}{W(m)} &\approx \frac{2S_1(m) + S_2(m)}{4\pi m/\sqrt{3} - 1 - 3S_3(m)} \\ &= \frac{\frac{2S_1(m)}{\sqrt{m}} + \frac{S_2(m)}{\sqrt{m}}}{\frac{4\pi\sqrt{m}}{\sqrt{3}} - \frac{1}{\sqrt{m}} - \frac{3S_3(m)}{\sqrt{m}}} \end{aligned}$$

For $50,000 \leq m \leq 10,000,000$,

$$\frac{2S_1(m)}{\sqrt{m}} + \frac{S_2(m)}{\sqrt{m}} \approx 2.17$$

and

$$2.16 < \frac{3S_3(m)}{\sqrt{m}} < 2.17.$$

For such values of m ,

$$\frac{P(m)}{W(m)} \approx \frac{2.17}{4\pi\sqrt{m}/\sqrt{3}}$$

If m_1 and m_2 are each in the above range,

$$\frac{P(m_1)W(m_2)}{P(m_2)W(m_1)} \approx \frac{\sqrt{m_2}}{\sqrt{m_1}}$$

Remark. For the U.S. presidential election, the values of m are in the interval from 50,000 to 10,000,000.

It is not hard to see that if there are k states, if there are three candidates, if the number of voters in state i is $6m_i + 3$, and if the winner must carry *each* state, then the number of pivot moves for a voter of state 1 is $P(m_1)[W(m_2)W(m_3) \cdots W(m_k)]$.

We proceed now to presidential elections in which a candidate does not have to carry each state in order to win the election. It will be convenient to let $L(m)$ denote the number of ways of distributing $6m + 3$ voters among three boxes in such a way that the number of voters in box 1 is less than or equal to the number in at least one of the other boxes. Clearly, $L(m) = 3^{6m+3} - W(m)$.

Now suppose there are three states and three candidates and that a candidate must carry two states to win the election. If there are $6m_i + 3$ voters in state i , then the number of pivot moves for voter 1 in state 1 is $P(m_1) \cdot [W(m_2)L(m_3) + W(m_3)L(m_2)]$ (since that voter must pivot in the plurality election within state 1 and that voter's candidate must win precisely one other state in order for the voter in state 1 to have a pivot move in the presidential election).

In general, let there be k states with electoral votes w_1, w_2, \dots, w_k and voting populations $6m_1 + 3, 6m_2 + 3, \dots, 6m_k + 3$. Let $[q; w_1, w_2, \dots, w_k]$ denote the k -player weighted voting game with quota q , i.e., the simple game in which a coalition \mathcal{S} is winning if $\sum_{i \in \mathcal{S}} w_i \geq q$. Let $c(j)$ denote the number of pivot sets for player j in this simple game and let $\mathcal{S}_{j1}, \mathcal{S}_{j2}, \dots, \mathcal{S}_{jc(j)}$ be a listing of the pivot sets for player j . A voter of state 1 will have at least one pivot move in the presidential election if this voter has a pivot move for the plurality election within state 1 and if his candidate carries just enough other states to make the quota, q . It follows that the number of pivot moves for a voter of state 1 in the presidential election is

$$P(m_1) \sum_{i=1}^{c(1)} \prod_{j \in \mathcal{S}_{1i}, j \neq 1} W(m_j) \prod_{j \notin \mathcal{S}_{1i}} L(m_j).$$

This yields the following theorem.

THEOREM 2. *If there are k states with electoral votes w_1, w_2, \dots, w_k , if state i has $6m_i + 3$ voters, if there are three candidates, and if a candidate must receive q electoral votes to win the election, then the ratio of the number of pivot moves for a voter of state 1 to the number of pivot moves for a voter of state 2 is*

$$(3) \quad \frac{P(m_1) \sum_{i=1}^{c(1)} \prod_{j \in \mathcal{S}_{1i}, j \neq 1} W(m_j) \prod_{j \notin \mathcal{S}_{1i}} L(m_j)}{P(m_2) \sum_{i=1}^{c(2)} \prod_{j \in \mathcal{S}_{2i}, j \neq 2} W(m_j) \prod_{j \notin \mathcal{S}_{2i}} L(m_j)}.$$

COROLLARY. *If there are k states with electoral votes w_1, w_2, \dots, w_k , if state i has $6m_i + 3$ voters, if there are three candidates, and if a candidate must receive q electoral votes to win, then for $50,000 \leq m \leq 10,000,000$, the ratio of the number of pivot moves for a voter of state 1 to the number of pivot moves for a voter of state 2 is approximately*

$$(4) \quad \frac{\sum_{i=1}^{c(1)} \left(\frac{1}{2}\right)^{|\mathcal{S}_{1i}|}}{\sum_{i=1}^{c(2)} \left(\frac{1}{2}\right)^{|\mathcal{S}_{2i}|}} \cdot \frac{\sqrt{m_2}}{\sqrt{m_1}}$$

where $|\mathcal{S}_{ji}|$ denotes the number of elements in the set \mathcal{S}_{ji} .

Proof of the corollary. For large m ,

$$\begin{aligned} \frac{[(2m+1)!]^2(2m)!}{(6m+2)!} P(m) &\approx 2S_1(m) + S_2(m), \\ \frac{[(2m+1)!]^2(2m)!}{(6m+2)!} W(m) &\approx \frac{4\pi m}{\sqrt{3}} - 1 - 3S_3(m), \\ \frac{[(2m+1)!]^2(2m)!}{(6m+2)!} L(m) &\approx \frac{8\pi m}{\sqrt{3}} + 1 + 3S_3(m). \end{aligned}$$

Now multiply the numerator and denominator of (3) by

$$\frac{[(2m_1+1)!]^2(2m_1)!}{(6m_1+2)!} \frac{[(2m_2+1)!]^2(2m_2)!}{(6m_2+2)!} \dots \frac{[(2m_k+1)!]^2(2m_k)!}{(6m_k+2)!}$$

and by $\sqrt{m_1 m_2 \dots m_k}$.

The ratio in (3) is then approximately equal to

$$\begin{aligned} (2.17) \quad & \sum_{i=1}^{c(1)} \prod_{j \in \mathcal{S}_{1i}, j \neq 1} \frac{4\pi\sqrt{m_j}}{\sqrt{3}} \prod_{j \notin \mathcal{S}_{1i}} \frac{8\pi\sqrt{m_j}}{\sqrt{3}} \\ (2.17) \quad & \sum_{i=1}^{c(2)} \prod_{j \in \mathcal{S}_{2i}, j \neq 2} \frac{4\pi\sqrt{m_j}}{\sqrt{3}} \prod_{j \notin \mathcal{S}_{2i}} \frac{8\pi\sqrt{m_j}}{\sqrt{3}} \\ &= \frac{\sum_{i=1}^{c(1)} \left(\frac{4\pi}{\sqrt{3}}\right)^{|\mathcal{S}_{1i}|-1} \left(\frac{8\pi}{\sqrt{3}}\right)^{k-|\mathcal{S}_{1i}|}}{\sum_{i=1}^{c(2)} \left(\frac{4\pi}{\sqrt{3}}\right)^{|\mathcal{S}_{2i}|-1} \left(\frac{8\pi}{\sqrt{3}}\right)^{k-|\mathcal{S}_{2i}|}} \cdot \frac{\sqrt{m_2 m_3 \dots m_k}}{\sqrt{m_1 m_3 m_4 \dots m_k}} \\ &= \frac{\left(\frac{4\pi}{\sqrt{3}}\right)^{k-1} \sum_{i=1}^{c(1)} 2^{k-|\mathcal{S}_{1i}|}}{\left(\frac{4\pi}{\sqrt{3}}\right)^{k-1} \sum_{i=1}^{c(2)} 2^{k-|\mathcal{S}_{2i}|}} \cdot \frac{\sqrt{m_2}}{\sqrt{m_1}} \\ &= \frac{\sum_{i=1}^{c(1)} \left(\frac{1}{2}\right)^{|\mathcal{S}_{1i}|}}{\sum_{i=1}^{c(2)} \left(\frac{1}{2}\right)^{|\mathcal{S}_{2i}|}} \cdot \frac{\sqrt{m_2}}{\sqrt{m_1}}. \end{aligned}$$

Remark. Banzhaf [1] showed that in the case of two candidates, the approximate ratio of voting powers of individuals is

$$\frac{c(1)}{c(2)} \cdot \frac{\sqrt{m_2}}{\sqrt{m_1}}$$

In order to calculate the ratio in (4) for the states in a United States presidential election, we need the numbers $|\mathcal{S}_{ji}|$ for each pair (i, j) . Several computer programs are available for this computation. One such program was written by Ken Fox while he was a student at Miami University. Using his program and the numbers of electoral votes assigned to the various states for the 1972 election, we find that, for the case of three candidates, the ratio of the number of pivot moves of a voter in California

(45 electoral votes) to the number of pivot moves of a voter in Utah (4 electoral votes) is approximately $(12.3)(0.23) = 2.8$. (The second factor on the left-hand side of the previous equality is the ratio of the square root of the population of Utah to the square root of the population of California.) If there were only two candidates, then the ratio of voting powers would be approximately 12.2 times the square root of the ratio of the numbers of voters in these two states. This indicates that with three candidates there is slightly more bias in favor of an individual in a large state. This raises the question as to whether the bias will increase as the number of candidates increases. We now consider the case of four candidates.

The reasoning which led to formula (3) for the case of three candidates is still valid for a four-candidate presidential election. However, the expressions $P(m)$, $W(m)$, and $L(m)$ will have to be modified if there are four or more candidates. Here we are writing $P(m)$ instead of $p(1; 8m, 4)$. $W(m)$ will be the number of ways of distributing $8m$ voters among four distinguishable boxes in such a way that box 1 will contain more voters than any of the boxes. $L(m)$ is the number of ways of distributing $8m$ voters among four distinguishable boxes so that the number in box 1 is less than or equal to the number of voters in some other box. Hence, $L(m) = 4^{8m} - W(m)$. The major problem will be the computation of $P(m)$, $W(m)$, and $L(m)$. If one wishes to skip the incredibly messy details, one may proceed directly to the main result which is presented in Theorem 3.

To compute $P(m)$, we let

$$\begin{aligned}
 T_1(m) &= \sum_{j=1}^{2m-1} \frac{(2m+1)!(2m)!(2m-1)!(2m-1)!}{(2m+j)!(2m+j-1)!(2m-j)!(2m-j)!} \\
 T_2(m) &= \sum_{j=1}^{\lceil(2m+1)/3\rceil} \sum_{k=1}^{2j-1} \frac{(2m+1)!(2m)!(2m)!(2m-2)!}{(2m+j)!(2m+j-1)!(2m-j+k)!(2m-j-k)!} \\
 T_3(m) &= \sum_{j=\lfloor(2m+1)/3\rfloor+1}^{2m-1} \sum_{k=1}^{2m-j} \frac{(2m+1)!(2m)!(2m)!(2m-2)!}{(2m+j)!(2m+j-1)!(2m-j+k)!(2m-j-k)!} \\
 T_4(m) &= \sum_{j=0}^{\lceil(2m-1)/3\rceil} \frac{(2m)!(2m)!(2m)!(2m-1)!}{(2m+j)!(2m+j)!(2m+j)!(2m-1-3j)!} \\
 T_5(m) &= \sum_{j=1}^{\lfloor 2m/3 \rfloor} \sum_{k=1}^{2j} \frac{(2m+1)!(2m+1)!(2m-1)!(2m-2)!}{(2m+j)!(2m+j)!(2m-1-j+k)!(2m-j-k)!} \\
 T_6(m) &= \sum_{j=\lfloor 2m/3 \rfloor+1}^{2m-1} \sum_{k=1}^{2m-j} \frac{(2m+1)!(2m+1)!(2m-1)!(2m-2)!}{(2m+j)!(2m+j)!(2m-1-j+k)!(2m-j-k)!}
 \end{aligned}$$

Then

$$\begin{aligned}
 P(m) &= \frac{1}{2} \frac{(8m-1)!}{(2m+1)!(2m)!(2m-1)!(2m-1)!} T_1(m) \\
 &\quad + \frac{(8m-1)!}{(2m+1)!(2m)!(2m)!(2m-2)!} [T_2(m) + T_3(m)] \\
 &\quad + \frac{(8m-1)!}{(4m)!(4m-1)!} + \frac{3}{2} \frac{(8m-1)!}{(2m)!(2m)!(2m)!(2m-1)!} T_4(m) \\
 &\quad + 3 \frac{(8m-1)!}{(2m+1)!(2m+1)!(2m-1)!(2m-2)!} [T_5(m) + T_6(m)].
 \end{aligned}$$

Thus,

$$\begin{aligned} & \frac{(2m+1)!(2m)!(2m-1)!(2m-1)!}{(8m-1)!} P(m) \\ &= \frac{1}{2} T_1(m) + \frac{2m-1}{2m} (T_2(m) + T_3(m)) \\ & \quad + \frac{(2m+1)!(2m)!(2m-1)!(2m-1)!}{(4m)!(4m-1)!} + \frac{3}{2} \frac{2m+1}{2m} T_4(m) \\ & \quad + 3 \frac{2m-1}{2m+1} (T_5(m) + T_6(m)) \\ & \approx \frac{1}{2} T_1(m) + T_2(m) + T_3(m) + \frac{3}{2} T_4(m) + 3(T_5(m) + T_6(m)). \end{aligned}$$

Tedious calculations show that for $40,000 \leq m \leq 3,000,000$,

$$\frac{1}{2} \frac{T_1(m)}{m} + \frac{T_2(m) + T_3(m)}{m} + \frac{3}{2} \frac{T_4(m)}{m} + \frac{3(T_5(m) + T_6(m))}{m} \approx 2.70,$$

and thus

$$(5) \quad \frac{1}{m} \frac{(2m+1)!(2m)![(2m-1)!]^2}{(8m-1)!} P(m) \approx 2.70.$$

As in the case of 3 candidates, it is more convenient to obtain $W(m)$ indirectly. The total number of ways of distributing $8m$ voters among 4 distinguishable boxes is 4^{8m} . Let $T(m)$ be the number of these in which there is a tie (i.e., the two boxes with the largest number of voters contain the same numbers of voters). Then $W(m) = \frac{1}{4}(4^{8m} - T(m))$.

LEMMA 5.

$$\begin{aligned} T(m) &= \binom{8m}{2m} \binom{6m}{2m} \binom{4m}{2m} \\ & \quad + 4 \sum_{j=1}^{[2m/3]} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m+j} \\ & \quad + 6 \sum_{j=1}^{2m} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} \\ & \quad + 12 \sum_{j=1}^{2m-1} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \sum_{k=1}^{2j-1} \binom{4m-2j}{2m-j+k}. \end{aligned}$$

Proof. A tie can occur in the following ways:

(i) All four boxes contain $2m$ voters; this accounts for the first term on the right-hand side.

(ii) For $1 \leq j \leq [2m/3]$, three of the boxes each contain $2m+j$ voters and the other box contains $2m-3j$. There are 4 ways to choose which box gets the $2m-3j$ voters. The voters can then be chosen in

$$\binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m+j}$$

ways.

(iii) For $1 \leq j \leq 2m$, two of the boxes each contain $2m + j$ voters and the other two boxes each contain $2m - j$ voters. There are 6 ways to choose the two boxes which will each get $2m - j$ voters. The voters can then be chosen in

$$\binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j}$$

ways.

(iv) For $1 \leq j \leq 2m - 1$ and $1 \leq k \leq 2j - 1$, the numbers of voters in the boxes are $2m + j$, $2m + j$, $2m - j + k$, $2m - j - k$. There are 4 ways to choose the box which will get $2m - j + k$ voters, after which there are 3 ways to choose the box which will get $2m - j - k$ voters. The voters can then be chosen in

$$\binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j+k}$$

ways.

For computational purposes it will be convenient to remove from the double sum those binomial coefficients which equal 0.

COROLLARY.

$$\begin{aligned} T(m) &= \binom{8m}{2m} \binom{6m}{2m} \binom{4m}{2m} \\ &+ 4 \sum_{j=1}^{[2m/3]} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} \\ &+ 6 \sum_{j=1}^{2m} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} \\ &+ 12 \sum_{j=1}^{[2m/3]} \sum_{k=1}^{2j-1} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j+k} \\ &+ 12 \sum_{j=[2m/3]+1}^{2m-1} \sum_{k=1}^{2m-j} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j+k}. \end{aligned}$$

Proof. We note that $2m - j + k \leq 4m - 2j \Leftrightarrow k \leq 2m - j$. Further, $j \leq [2m/3] \Rightarrow 3j \leq 2m + 1 \Rightarrow 2j - 1 \leq 2m - j$, whereas $j \geq [2m/3] + 1 \Rightarrow 3j \geq 2m + 1 \Rightarrow 2j - 1 \geq 2m - j$.

COROLLARY.

$$\begin{aligned} T(m) &= \binom{8m}{2m} \binom{6m}{2m} \binom{4m}{2m} \\ &+ 4 \sum_{j=1}^{[2m/3]} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m+j} \\ &+ 12 \sum_{j=1}^{[2m/3]} \sum_{k=1}^{2j-1} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j+k} \\ &+ 6 \sum_{j=1}^{[2m/3]} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} \\ &+ 6 \sum_{j=[2m/3]+1}^{2m} \binom{8m}{2m+j} \binom{6m-j}{2m+j} 2^{4m-2j}. \end{aligned}$$

Proof. Observe that

$$\sum_{k=1}^{2m-j} \binom{4m-2j}{2m-j+k} = \sum_{i=0}^{2m-j-1} \binom{4m-2j}{i} = \frac{1}{2} \left[(2^{4m-2j}) - \binom{4m-2j}{2m-j} \right].$$

Then

$$\begin{aligned} 12 \sum_{j=[2m/3]+1}^{2m-1} \sum_{k=1}^{2m-j} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j+k} \\ = 6 \sum_{j=[2m/3]+1}^{2m-1} \binom{8m}{2m+j} \binom{6m-j}{2m+j} 2^{4m-2j} \\ - 6 \sum_{j=[2m/3]+1}^{2m-1} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j}. \end{aligned}$$

The proof is completed by writing

$$\begin{aligned} \sum_{j=1}^{2m} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} \\ = \sum_{j=1}^{[2m/3]} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} \\ + \sum_{j=[2m/3]+1}^{2m-1} \binom{8m}{2m+j} \binom{6m-j}{2m+j} \binom{4m-2j}{2m-j} + \binom{8m}{4m} \binom{4m}{4m} \end{aligned}$$

and then combining each of these terms with the appropriate summation.

Now let $N = [2m/3]$ and

$$\begin{aligned} U_1(m) &= \sum_{j=1}^N \frac{(2m+1)!(2m+1)!(2m+1)!(2m-3)!}{(2m+j)!(2m+j)!(2m+j)!(2m-3j)!}, \\ U_2(m) &= \sum_{j=1}^N \sum_{k=1}^{2j-1} \frac{(2m+1)!(2m+1)!(2m)!(2m-2)!}{(2m+j)!(2m+j)!(2m-j+k)!(2m-j-k)!}, \\ U_3(m) &= \sum_{j=1}^N \frac{(2m+1)!(2m+1)!(2m-1)!(2m-1)!}{(2m+j)!(2m+j)!(2m-j)!(2m-j)!}, \\ U_4(m) &= \sum_{j=N+1}^{2m} \frac{(2m+N+1)!(2m+N+1)!(4m-2N-2)!4^{N+1}}{(2m+j)!(2m+j)!(4m-2j)!4^j}. \end{aligned}$$

Then

$$\begin{aligned} \frac{(2m+1)!(2m)!(2m-1)!(2m-1)!}{(8m-1)!} T(m) \\ = \frac{4m+2}{m} + \frac{16(2m-1)(2m-2)}{(2m+1)^2} U_1(m) + \frac{48(2m-1)}{2m+1} U_2(m) \\ + \frac{48m}{2m+1} U_3(m) + \frac{(48m)(2m+1)!(2m)!(2m-1)!(2m-1)!2^{4m-2N-2}}{(2m+N+1)!(2m+N+1)!(4m-2N-2)!} U_4(m) \\ \approx 4 + 16U_1(m) + 48U_2(m) + 24U_3(m) + 0 \cdot U_4(m). \end{aligned}$$

Tedious calculations show that for $40,000 \leq m \leq 3,000,000$

$$32.3 < \frac{4}{m} + \frac{16U_1(m)}{m} + \frac{48U_2(m)}{m} + \frac{24U_3(m)}{m} < 32.5,$$

so that for $40,000 \leq m \leq 3,000,000$

$$(6) \quad \frac{(2m+1)!(2m)![(2m-1)!]^2}{(8m-1)!} \frac{T(m)}{m} \approx 32.$$

Since $W(m) = \frac{1}{4}(4^{8m} - T(m))$,

$$(7) \quad \begin{aligned} & \frac{(2m+1)!(2m)![(2m-1)!]^2}{(8m-1)!} \frac{W(m)}{m} \\ & \approx \frac{(2m+1)!(2m)![(2m-1)!]^2 4^{8m-1}}{m(8m-1)!} - 8 \\ & \approx 4\pi^{3/2} m^{1/2} - 8 \approx 4\pi^{3/2} m^{1/2}. \end{aligned}$$

Similarly, since $L(m) = 4^{8m} - W(m)$,

$$(8) \quad \frac{(2m+1)!(2m)![(2m-1)!]^2}{(8m-1)!} \frac{L(m)}{m} \approx 3(4\pi^{3/2} m^{1/2}).$$

We are now ready to state the main result for the case of four candidates.

THEOREM 3. *If there are k states with electoral votes w_1, w_2, \dots, w_k , if state i has $8m_i$ voters, if there are four candidates, and if a candidate must receive q electoral votes to win the election, then the ratio of the number of pivot moves for a voter of state 1 to the number of pivot moves for a voter of state 2 is approximately*

$$(9) \quad \frac{\sum_{i=1}^{c(1)} \left(\frac{1}{3}\right)^{|\mathcal{S}_{1i}|}}{\sum_{i=1}^{c(2)} \left(\frac{1}{3}\right)^{|\mathcal{S}_{2i}|}} \cdot \frac{\sqrt{m_2}}{\sqrt{m_1}}$$

where $\mathcal{S}_{j1}, \mathcal{S}_{j2}, \dots, \mathcal{S}_{j,c(j)}$ are the pivot sets for state j in the simple game $[q; w_1, w_2, \dots, w_k]$.

Proof. For the case of four candidates, we multiply and divide the ratio in (3) by

$$\prod_{i=1}^k \frac{(2m_i+1)!(2m_i)!(2m_i-1)!(2m_i-1)!}{m_i(8m_i-1)!}.$$

Formula (3) becomes (approximately)

$$(2.70) \quad \frac{\sum_{i=1}^{c(1)} \prod_{j \in \mathcal{S}_{1i}, j \neq 1} (4\pi^{3/2} m_j^{1/2}) \prod_{j \notin \mathcal{S}_{1i}} (12\pi^{3/2} m_j^{1/2})}{\sum_{i=1}^{c(2)} \prod_{j \in \mathcal{S}_{2i}, j \neq 2} (4\pi^{3/2} m_j^{1/2}) \prod_{j \notin \mathcal{S}_{2i}} (12\pi^{3/2} m_j^{1/2})} = \frac{\sum_{i=1}^{c(1)} 3^{k-|\mathcal{S}_{1i}|}}{\sum_{i=1}^{c(2)} 3^{k-|\mathcal{S}_{2i}|}} \cdot \frac{\sqrt{m_2}}{\sqrt{m_1}}.$$

Using the numbers of electoral votes assigned to the various states for the 1972 presidential election, we find that the ratio of the number of pivot moves of a voter in California to the number of pivot moves of a voter in Utah is approximately 12.6 times the square root of the ratio of the number of voters in Utah to the number of voters in California. Thus, the bias in favor of a voter in California relative to a voter

in Utah does indeed increase as the number of candidates increases from 3 to 4. Indeed, the bias in favor of a voter in California relative to a voter in a state with 3, 4, 5, 6, or 7 electoral votes increases as the number of candidates increases from 3 to 4.

4. Conjectures concerning multicandidate presidential elections. Formulas (4) and (9) immediately suggest the following:

CONJECTURE. *If there are k states with electoral votes w_1, w_2, \dots, w_k , if state i has v_i voters, if there are r candidates, and if a candidate must receive q electoral votes to win the election, then the ratio of the numbers of pivot moves for a voter of state 1 to the number of pivot moves for a voter of state 2 is approximately*

$$\frac{\sum_{i=1}^{c(1)} \left(\frac{1}{r-1} \right)^{|\mathcal{P}_{1i}|} \sqrt{v_2}}{\sum_{i=1}^{c(2)} \left(\frac{1}{r-1} \right)^{|\mathcal{P}_{2i}|} \sqrt{v_1}}$$

Unfortunately, I have not been able to obtain formulas for $P(m)$, $W(m)$, $L(m)$ in the case of r candidates. Consequently, I have been unable to prove the above conjecture.

However, *if the conjecture is true*, then when there are ten candidates, an individual in California has approximately three and one-half times as much voting power as an individual in Utah. For two candidates, the ratio is approximately 2.8. (In calculating these ratios, we used the electoral votes for 1972 and populations of the states rather than the numbers of citizens of voting age.) It does indeed appear that as the number of candidates increases, the bias in favor of a voter in California relative to a voter in Utah also increases.

REFERENCES

- [1] J. R. BANZHAF III, *One man, 3.312 votes*, Villanova Law Rev., 13 (1968), pp. 304–332.
- [2] S. J. BRAMS, *Game Theory and Politics*, Free Press, New York, 1975.
- [3] P. DUBEY AND L. S. SHAPLEY, *Mathematical properties of the Banzhaf power index*, Math. Oper. Res., 4 (1979), pp. 99–131.
- [4] W. F. LUCAS, *Measuring power in weighted voting systems*, in Political and Related Models, S. J. Brams, W. F. Lucas, and P. D. Straffin, Jr., eds., Modules in Applied Mathematics, 2, Springer-Verlag, New York, 1982.
- [5] I. MANN AND L. S. SHAPLEY, *The a priori voting strength of the electoral college*, in Game Theory and Related Approaches to Social Behavior, M. Shubik, ed. Wiley, New York, 1964.
- [6] G. OWEN, *Evaluation of a presidential election*, Amer. Political Sci. Rev., 69 (1975), pp. 947–953.
- [7] ———, *Multilinear extensions and the Banzhaf value*, Naval Res. Logist. Quart., 22 (1975), pp. 741–750.
- [8] P. D. STRAFFIN, JR., *Power indices in politics*, in Political and Related Models, S. J. Brams, W. F. Lucas, and P. D. Straffin, Jr., eds., Modules in Applied Mathematics, 2, Springer-Verlag, New York, 1982.

RUNS, SLIDES AND MOMENTS*

LOUIS W. SHAPIRO,† WEN-JIN WOAN† AND SEYOUM GETU†

Abstract. The moments of a Catalan triangle are computed. The method is similar to that used for probability generating functions. This however does not explain the elegance of the results until further combinatorial connections are developed. These include Eulerian numbers, runs, zig-zag permutations, tangent numbers and a kind of nontrivial run called a slide.

Key words. Catalan triangle, runs, Eulerian numbers, moments, slides, tangent numbers, zig-zag permutations

1. Background and motivation. This study started with a problem in computing the moments of the Catalan triangle. This project became more interesting when the Eulerian numbers appeared and far more so when the tangent numbers showed up.

The Catalan triangle is defined by $B_{nk} = (k/n) \binom{2n}{n+k}$, and its first few entries are as follows:

k	1	2	3	4	5	6
n	1	0	0	0	0	0
1	1	0	0	0	0	0
2	2	1	0	0	0	0
3	5	4	1	0	0	0
4	14	14	6	1	0	0
5	42	48	27	8	1	0
6	132	165	110	44	10	1

Many properties of this triangle are discussed in a paper entitled *A Catalan Triangle* [11]. This paper will be denoted ACT. Other related papers are Epplett [4], Moon [9], Rogers [10] and Strehl [13]. The interpretation discussed in ACT is by pairs of nonintersecting paths, but there are several other interesting interpretations. For example, if a coin is flipped $2n$ times and the running total of heads always exceeds tails, then there are B_{nk} possible sequences where $2k$ is the excess of heads over tails. Yet another interpretation is by linear forests of rooted planar trees with k nontrivial trees arranged in a line having a total of n edges.

An interesting observation concerning the triangle is

$$\begin{aligned}
 1 \cdot 1 &= 1, \\
 2 \cdot 1 + 1 \cdot 2 &= 4, \\
 5 \cdot 1 + 4 \cdot 2 + 1 \cdot 3 &= 4^2, \\
 14 \cdot 1 + 14 \cdot 2 + 6 \cdot 3 + 1 \cdot 4 &= 4^3,
 \end{aligned}$$

which leads to the immediate guess about first moments

$$\sum_{k=1}^n kB_{nk} = 4^{n-1} \quad \text{for } n \geq 1,$$

* Received by the editors October 30, 1981, and in revised form February 23, 1982.

† Department of Mathematics, Howard University, Washington, DC 20001.

or

$$\sum_{n=1}^{\infty} \left(\sum_{k=1}^n kB_{nk} \right) x^n = \frac{x}{1-4x} = F_1(x).$$

In ACT a related result is proved:

$$\sum_{k=1}^n B_{nk} = \frac{1}{2} \binom{2n}{n},$$

which can be rephrased as

$$\sum_{n=1}^{\infty} \left(\sum_{k=1}^n B_{nk} \right) x^n = \frac{x}{2\sqrt{1-4x}} = F_0(x).$$

We consider next the second moments:

$$\begin{aligned} 1 \cdot 1^2 &= 1, \\ 2 \cdot 1^2 + 1 \cdot 2^2 &= 6, \\ 5 \cdot 1^2 + 4 \cdot 2^2 + 1 \cdot 3^2 &= 30, \\ 14 \cdot 1^2 + 14 \cdot 2^2 + 6 \cdot 3^2 + 1 \cdot 4^2 &= 140. \end{aligned}$$

This sequence is not in Sloane’s *Handbook* [12], but we observe

$$\begin{aligned} 1 + 6x + 30x^2 + 140x^3 + \dots &= (1 + 4x + 16x^2 + 64x^3 + \dots) \\ &\cdot (1 + 2x + 6x^2 + 20x^3 + \dots), \end{aligned}$$

which suggests the relationship

$$F_2(x) = \sum_{n=1}^{\infty} \left(\sum_{k=1}^n k^2 B_{nk} \right) x^n = \frac{x}{1-4x} \frac{1}{\sqrt{1-4x}} = \frac{x}{(1-4x)^{3/2}}.$$

By computing the first few terms of each moment sequence, the following seem reasonable:

$$\begin{aligned} F_3(x) &= \frac{x + 2x^2}{(1-4x)^2}, \\ F_4(x) &= \frac{x + 8x^2}{(1-4x)^{5/2}} \\ (2) \quad F_5(x) &= \frac{x + 22x^2 + 16x^2}{(1-4x)^3} \end{aligned}$$

and

$$F_n(x) = \frac{\sum m(n, s)x^s}{(1-4x)^{(n+1)/2}}.$$

The numerators are the interesting feature, and if we put the coefficients in tabular form we obtain

		$m(n, s)$			
		1	2	3	4
$n \backslash s$					
1		1	0	0	0
2		1	0	0	0
3		1	2	0	0
4		1	8	0	0
5		1	22	16	0
6		1	52	136	0
7		1	114	720	272

At this point several observations can be made. The $m(n, s)$ seem to be nonnegative integers and thus possibly of combinatorial origin. A less obvious but more intriguing observation is that the boldfaced entries are the first few tangent numbers, T_n , where $\tan x = \sum_{n=0}^{\infty} T_n x^{2n+1}/(2n+1)!$. If we divide each entry in the second column by 2, we get some of the Eulerian numbers. In § 2 we will compute some moments in combinatorial settings. In § 3 we will give a short derivation of (2) above, which however will depend on a combinatorial setting for the $m(n, s)$ and an identity which will be proven in § 4.

2. Moments. Let (a_{ij}) be an infinite matrix with all but a finite number of zeros in each row. Let $D_j(x) = \sum_{i=1}^{\infty} a_{ij}x^i$ be the generating function for the j th column. The exponential generating function for the moments is given by

$$(4) \quad M = D_1(x) e^z + D_2(x) e^{2z} + \dots = \sum_{k=1}^{\infty} D_k(x) e^{kz}$$

in the following sense: the coefficient of $z^k/k!$ is

$$M_k = 1^k D_1(x) + 2^k D_2(x) + \dots = \sum_{m=1}^{\infty} m^k D_m(x).$$

This generalizes slightly the moment generating function associated with a probability density function (see Freund [7] and Feller [6]), since we have a column, $D_j(x)$, instead of a probability, p_j .

If $f(x) = \sum_{k=0}^{\infty} a_k x^k$, then define $\mathcal{C}_n(f(x)) = a_n$.

The examples given below depend on the following situation. Suppose $D_i(x) = [D(x)]^i$ for a given generating function, $D(x)$. Then

$$(5) \quad M = \frac{D(x) e^z}{1 - D(x) e^z}.$$

LEMMA. Letting $D(x) = D$ we have

$$(6) \quad \frac{\partial^n}{\partial z^n} \left(\frac{D e^z}{1 - D e^z} \right) = \frac{\sum_{k=1}^n A(n, k) (D e^z)^k}{(1 - D e^z)^{n+1}},$$

where the $A(n, k)$ are the Eulerian numbers.

Proof. The key recurrence is

$$(7) \quad A(n+1, k) = kA(n, k) + (n+1-k)A(n, k-1),$$

which can be used to define the Eulerian numbers. \square

Another definition of $A(n, k)$ is as the number of permutations of $[n]$ with k runs. This yields the same recurrence. See Knuth [8] for a good exposition of this material.

The polynomial $A_n(x)$ is defined as $\sum_{k=1}^n A(n, k)x^k$.

Example 1. Let $D(x) = x$. This simplest case goes back to Euler [5] and already is of some interest.

$$(8) \quad \left. \frac{\partial^n}{\partial z^n} \left(\frac{x e^z}{1 - x e^z} \right) \right|_{z=0} = \frac{\sum_{k=1}^n A(n, k)x^k}{(1-x)^{n+1}} = \frac{A_n(x)}{(1-x)^{n+1}} = \sum_{m=1}^{\infty} m^n x^m.$$

Example 2. We want to compute the moments of successive Bernoulli trials. Consider the matrix

$$(9) \quad B = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ q & p & 0 & 0 & \cdots \\ q^2 & 2pq & p^2 & 0 & \cdots \\ q^3 & 3pq^2 & 3p^2q & p^3 & \cdots \end{pmatrix},$$

where p and q are probabilities of success and failure, respectively. Now consider the matrix pB . For pB one obtains $D_1 = D = px/(1 - qx)$ and $D_i = D^i$.

Let $M^* = D + D_2 e^z + D_3 e^{2z} + \cdots = D/(1 - D e^z)$. Then

$$(10) \quad \frac{\partial M^*}{\partial z} = \frac{D^2 e^z}{(1 - D e^z)^2}.$$

Comparing this with (5) yields

$$\frac{\partial^n M^*}{\partial z^n} = D \frac{\partial^n M}{\partial z^n} \quad \text{for } n \geq 1.$$

Thus

$$(11) \quad \frac{\partial^m M^*}{\partial z^m} = D \frac{\sum_{k=1}^m A(m, k)(D e^z)^k}{(1 - D e^z)^{m+1}}.$$

Letting $z = 0$ and recalling that $D = px/(1 - qx)$, we get

$$(12) \quad \frac{\sum_{k=1}^m A(m, k)(px/(1 - k))^k}{(1 - x)^{m+1}} (1 - qx)^m = M_m.$$

Then $(1/p)\mathcal{C}_{n+1}(M_m)$ yields the m th moment of $(q + px)^n$.

Indeed

$$(13) \quad \left. \frac{\partial M^*}{\partial z} \right|_{z=0} = \frac{D^2}{(1 - D)^2} = p^2 x^2 + 2p^2 x^3 + 3p^2 x^4 + \cdots,$$

and

$$(14) \quad \frac{1}{p} \mathcal{C}_{n+1} \left(\frac{D^2}{(1 - D)^2} \right) = np.$$

Similarly

$$(15) \quad \left. \frac{\partial^2 M^*}{\partial z^2} \right|_{z=0} - \left(\left. \frac{\partial M^*}{\partial z} \right|_{z=0} \right)^2$$

$$(16) \quad = \frac{D^2 + D^3}{(1 - D)^3} - \left(\frac{D^2}{(1 - D)^2} \right)^2$$

$$(17) \quad = (p^2x^2 - p^2qx^3 + p^3x^3) \sum_{k=0}^{\infty} \binom{2+k}{n} x^k = V.$$

Hence

$$(18) \quad \frac{1}{p} \mathcal{C}_{n+1}(V) = npq,$$

which is the variance.

3. The Catalan triangle. We now prove (2), which gives the generating function for the n th moments of the Catalan triangle. We will need the following identity, which will be proved in § 4:

$$(19) \quad A_n(x) = \sum_{k=1}^n A(n, k)x^k = \sum_{s=1}^{\lfloor n/2 \rfloor} m(n, s)x^s(1+x)^{n+1-2s}.$$

PROPOSITION 1. Let $F_n(x)$ be defined as in (2) in § 1. Then

$$F_n(x) = \frac{\sum_{s=1}^{\lfloor n/2 \rfloor} m(n, s)x^s}{(1-4x)^{(n+1)/2}}.$$

Proof. Start by noting the following facts about the generating function $c = c(x)$ of Catalan numbers:

$$(20) \quad c - 1 = xc^2, \quad 2 - c = c\sqrt{1 - 4x},$$

and for the Catalan triangle $D = c - 1$, by (6),

$$(21) \quad \begin{aligned} F_n(x) &= \left. \frac{\partial^n M}{\partial z^n} \right|_{z=0} = \frac{\sum_{k=1}^{\lfloor n/2 \rfloor} A(n, k)(c-1)^k}{[1 - (c-1)]^{n+1}} \\ &= \frac{\sum_{s=1}^{\lfloor n/2 \rfloor} m(n, s)(c-1)^s(1+c-1)^{n+1-2s}}{(2-c)^{n+1}} \\ &= \left(\frac{c}{2-c} \right)^{n+1} \sum_{s=1}^{\lfloor n/2 \rfloor} m(n, s) \left(\frac{c-1}{c^2} \right)^s \\ &= \frac{\sum_{s=1}^{\lfloor n/2 \rfloor} m(n, s)x^s}{(1-4x)^{(n+1)/2}}. \quad \square \end{aligned}$$

4. The missing identity and other results. Let $a_1 a_2 \cdots a_n$ be a permutation of $[n]$. Following Knuth we put bars at each end of the permutation and also between a_j and a_{j+1} whenever $a_j > a_{j+1}$. The runs are the segments between bars. For instance, $|357|1689|4|2|$ has four runs. Note that this allows runs of length one.

Start again, except now adjoin $a_0 = \infty$. Put asterisks at each end and also between a_j and a_{j+1} whenever $a_j < a_{j+1}$. For instance, the same permutation gives

$$* \infty 3 * 5 * 71 * 6 * 8 * 942 *.$$

A slide is any segment between asterisks of length at least two. Here we have three slides $\infty 3, 71$ and 942 .

Let $W(n, k, s)$ be the set of all permutations of $[n]$ with k runs and s slides. Then let $w(n, k, s) = |W(n, k, s)|$. The next proposition singles out the $w(n, s, s)$, and we

define $m(n, s) = w(n, s, s)$. Since the end of any slide is the beginning of a run, we will always have $k \geq s$.

PROPOSITION 2.

$$(22) \quad w(n, k, s) = \binom{n+1-2s}{k-s} w(n, s, s).$$

Proof. Any permutation in $W(n, s, s)$ has each slide of length just two. [Let $\infty 2 * 51 * 3 * 6 * 84 * 7 * 9 *$ be an example. The slides are $\infty 2, 51$ and 84 , while the runs are $25, 1368, 479$. Thus $s = k = 3$]. Counting $a_0 = \infty$, there are $n + 1$ symbols and $n + 1 - 2s$ that are not included in the slides. Choose $k - s$ of these $n + 1 - 2s$ elements, move each chosen element a_k to the left into the nearest slide $* a_j a_{j+1} *$ with $a_j > a_k > a_{j+1}$. [Say $k - s = 2$ and we choose 3 and 9 ; then $\infty 2 * 51 * 3 * 6 * 84 * 7 * 9 *$ becomes $\infty 9 2 * 531 * 6 * 84 * 7 *$.]

Figure 1 illustrates the one-to-one correspondence.

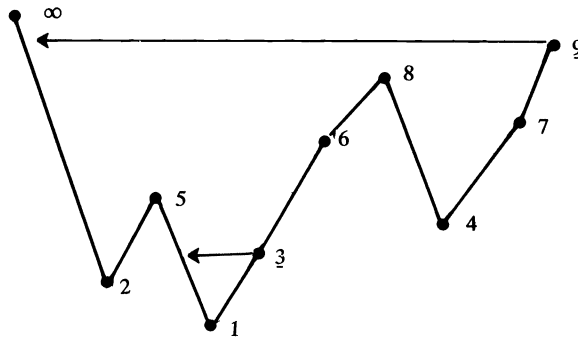


FIG. 1

Now let $A(n, k)$ be the number of permutations on $[n]$ with k runs. Then we have PROPOSITION 3.

$$(23) \quad A(n, k) = \sum_{1 \leq s \leq k} w(n, k, s) = \sum_{1 \leq s \leq k} \binom{n+1-2s}{k-s} m(n, s).$$

Recall that $A_n(x) = \sum_{k=1}^n A(n, k)x^k$. Then we obtain PROPOSITION 4.

$$(24) \quad A_n(x) = \sum_{s \geq 1} m(n, s)x^s(1+x)^{n+1-2s}.$$

Proof. Examining coefficients of x^k yields the equation given in Proposition 3.

This is identity (19) used in § 3. We can now easily confirm all the observations mentioned in § 1. Since $m(n, s) = w(n, s, s)$, all the $m(n, s)$ must indeed be positive integers, for $1 \leq s \leq [n/2]$.

The set $W(2k + 1, k, k)$ gives just zig-zag (or alternating) permutations of length $2k + 1$. By André's famous result on zig-zag permutations [1], [2],

$$(25) \quad \sum_{n=0}^{\infty} Z_n \frac{x^n}{n!} = \tan x + \sec x.$$

The odd part of this is indeed $\tan x$, which confirms that observation. One might wonder where the secant numbers are. These arise by discarding $a_0 = \infty$.

This leads to the following:

(26)

$n \backslash s$	0	1	2	3
1	1			
2	1	1		
3	1	5		
4	1	18	5	
5	1	58	61	
6	1	543	3111	1385

For n up to 15 see David, Kendall, and Barton's *Symmetric Functions and Allied Tables* [3, Table 7.2.2].

There are some other results of interest which however have long proofs. The proofs will be left out but are available from the authors.

Start by rearranging the $m(n, s)$ and putting them in a double generating function as follows (so as to put the tangent numbers in the first column):

(27)

$n \backslash s$	1	2	3	4	5
1	$1x$				
2		$1x^2y/2!$			
3	$2x^3/3:$		$1x^3y^2/3!$		
4		$8x^4y/4!$		$1x^4y^3/4!$	
5	$16x^5/5!$		$22x^5y^2/5!$		$1x^5y^4/5!$

Let $M = M(x, y)$ be the sum of all these terms.

PROPOSITION 5. M satisfies the quasi-linear partial differential equation

$$4M_y = 2M_x - 2 - xyM_x + y^2M_y - yM.$$

When $y = 0$, $M = \tan x$. Solving this P.D.E. yields

$$M = -\frac{y}{2} + \sqrt{1 - \left(\frac{y}{2}\right)^2} \tan \left(x \sqrt{1 - \left(\frac{y}{2}\right)^2} + \arcsin \frac{y}{2} \right).$$

This proposition has an interesting consequence. Let $r_n = \sum_{s \geq 1} m(n, s)$ be the number of what we will call *reduced* permutations of $[n]$. Then let $R(x) = \sum_{n=1}^{\infty} r_n (x^n/n!)$ be the exponential generating function for the r_n .

COROLLARY.

$$R(x) = \frac{\sqrt{3}}{2} \frac{\tan(\sqrt{3}x/2) + 1/\sqrt{3}}{1 - (1/\sqrt{3}) \tan(\sqrt{3}x/2)} - \frac{1}{2} = \frac{(1/\sqrt{3}) \tan(x\sqrt{3}/2)}{1 - (1/\sqrt{3}) \tan(x\sqrt{3}/2)}.$$

PROPOSITION 6. $m(n, s) = sm(n - 1, s) + 2(n - 2s + 2)m(n - 1, s - 1)$.

If we consider the number of permutations of $[n]$ with s slides, we immediately obtain:

PROPOSITION 7. There are $2^{n+1-2s}m(n, s)$ permutations of $[n]$ with s slides.

See David, Kendall, and Barton [3, Table 7.3]. Then summing over s yields

PROPOSITION 8. $\sum_{s \geq 1} m(n, s)2^{n+1-2s} = n!$

REFERENCES

- [1] D. ANDRÉ, *Développement de $\sec x$ et $\operatorname{tg} x$* , C.R. Acad. Sci. Paris, 97 (1879a), pp. 965–967.
- [2] L. COMTET, *Advanced Combinatorics*, D. Reidel, Boston, 1974.
- [3] F. DAVID, M. KENDALL AND D. BARTON, *Symmetric Functions and Allied Tables*, Cambridge Univ. Press, New York, 1966.
- [4] W. J. R. EPPLETT, *A note about the Catalan triangle*, Discr. Math., 25 (1979), pp. 289–291.
- [5] L. EULER, *Institutio calculi differentialis*, St. Petersburg 1755, pp. 483–485 or Opera Omnia (1), 10 (1913), pp. 373–375.
- [6] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. 1, third ed., John Wiley, New York, 1968.
- [7] J. FREUND, *Mathematical Statistics*, second ed., Prentice Hall, Englewood Cliffs, NJ, 1971.
- [8] D. KNUTH, *The Art of Computer Programming, Vol. 3, Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.
- [9] J. W. MOON, *Some enumeration problems for similarity relations*, Discr. Math., 26 (1979), pp. 251–260.
- [10] D. G. ROGERS, *Pascal triangles, Catalan numbers, and renewal arrays*, Discr. Math., 22 (1978), pp. 301–310.
- [11] L. W. SHAPIRO, *A Catalan triangle*, Discr. Math., 14 (1976), pp. 83–90.
- [12] N. J. A. SLOANE, *A Handbook of Integer Sequences*, Academic Press, New York, 1973.
- [13] V. STREHL, *A note on similarity relations*, Discr. Math., 19 (1977), pp. 99–101.

A MATROID ABSTRACTION OF THE BOTT-DUFFIN CONSTRAINED INVERSE*

SETH CHAIKEN†

Abstract. Let (E_1, E_2, \mathcal{G}) be a linking system with linking function γ as defined by Schrijver or bimatroid as defined by Kung. That is, there is a matroid on the disjoint union of E_1 and E_2 whose bases are E_1 and $(E_1 \setminus X) \cup Y$ for $(X|Y) \in \mathcal{G} \subset 2^{E_1} \times 2^{E_2}$. \mathcal{G} abstracts to matroid theory some properties of the nonsingular minors of a matrix and γ abstracts the submatrix rank function. For $i = 1, 2$ let \mathcal{M}_i be a matroid on E_i with rank function r_i and bases \mathcal{B}_i . Suppose $r_1(E_1) = r_2(E_2) = R$ and there are bases B_i in \mathcal{M}_i such that $(B_1|B_2) \in \mathcal{G}$. We show (E_2, E_1, \mathcal{T}) is a linking system where $(Y|X) \in \mathcal{T}$ iff there exist $F_i \subset E_i$ s.t. $F_1 \cap X = F_2 \cap Y = \phi$, $F_1 \cup X \in \mathcal{B}_1$, $F_2 \cup Y \in \mathcal{B}_2$ and $(F_1|F_2) \in \mathcal{G}$. The linking function

$$\tau(Y, X) = \min_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} [r_1(F_1 \cup X) + \gamma(F_1^c, F_2^c) + r_2(F_2 \cup Y)] - R \text{ for } \mathcal{T}$$

and Schrijver's extension of Edmond's intersection theorem are used in the proof.

As a special case, suppose \mathcal{M} is a matroid on E . Let E_1 and E_2 be disjoint copies of E , and X_1, Y_2 be the images of $X, Y \subset E$ in E_1, E_2 respectively. Then $\{(E_1 \setminus X_1) \cup Y_2\}$ there exists a base B in \mathcal{M} with $X \subset B$ and $(B \setminus X) \cup Y$ is a base in \mathcal{M} is the collection of bases of a matroid.

When \mathcal{M}_i is coordinatized by cycle space \mathcal{C}_i , \mathcal{G} is coordinatized generically by matrix G and $\mathcal{M}_i, \mathcal{G}$ satisfy the above conditions, then the Bott-Duffin inverse problem, which is to find v for i_0 such that $v \in \mathcal{C}_2^\perp$ and $Gv - i_0 \in \mathcal{C}_1$, has a unique solution $v = Ti_0$, and conversely. We show then that matrix T coordinatizes the linking system (E_2, E_1, \mathcal{T}) .

1. Introduction. Matroid theory results from abstracting the following combinatorial properties which the collection \mathcal{I} of linearly independent subsets of columns in a matrix obeys. (See Welsh [6] for a general matroid theory reference.)

(I1) If $I \in \mathcal{I}$ and $J \subset I$ then $J \in \mathcal{I}$.

(I2) If $I_1, I_2 \in \mathcal{I}$ and $|I_2| = |I_1| + 1$ then for some $x \in I_2 \setminus I_1$, $I_1 \cup \{x\} \in \mathcal{I}$.

Thus, a matroid $\mathcal{M} = (E, \mathcal{I})$ consists of a finite set E and a nonempty collection \mathcal{I} of subsets of E that satisfies (I1) and (I2). When \mathcal{I} is the collection of independent sets of columns in a matrix M , we say M represents (or coordinatizes) \mathcal{M} . The axioms in an equivalent definition of a matroid are the following properties of the nonempty collection \mathcal{B} of maximal independent sets. The members of \mathcal{B} are called bases. (We will abbreviate $\{x\}$ by x .)

(B1) If $B_1, B_2 \in \mathcal{B}$ then $|B_1| = |B_2|$.

(B2) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1$, there is some $y \in B_2$ such that $B_2 \setminus y \cup x \in \mathcal{B}$.

Matroid theory is also used to abstract the combinatorial properties of the collection of nonsingular minors (i.e. square submatrices) of a matrix. If M is a matrix with row set E_1 and column set E_2 (briefly, M is $E_1 \times E_2$; assume $E_1 \cap E_2 = \emptyset$), consider the matrix M' with columns $E_1 \cup E_2$ and rows E_1 obtained by appending an $E_1 \times E_1$ identity matrix to M :

$$M' = \left[\begin{array}{c|c} \overbrace{\begin{matrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{matrix}}^{E_1} & \overbrace{\begin{matrix} & & & \\ & & & \\ & & & \\ & & & M \end{matrix}}^{E_2} \end{array} \right].$$

* Received by the editors July 4, 1982, and in revised form January 6, 1983. This research was supported in part by the National Science Foundation under grant MCS-8204796. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26-29, 1982.

† Department of Computer Science, State University of New York, Albany, New York 12222.

It is easy to verify that $(E_1 \setminus X) \cup Y$ is a base in the matroid represented by M' if and only if $(X \mid Y)$ indexes a nonsingular minor of M . Now let \mathcal{M} be any matroid on E where $E = E_1 \cup E_2$, $E_1 \cap E_2 = \emptyset$ and E_1 is a base in \mathcal{M} . The above motivates us to consider the collection Λ of pairs of sets $(X \mid Y)$ such that $X \subset E_1$, $Y \subset E_2$ and $(E_1 \setminus X) \cup Y$ is a base in \mathcal{M} . Triples (E_1, E_2, Λ) associated with a matroid as above have been studied by Schrijver [4], [5] who called them linking systems and the elements of Λ linked pairs, and by Kung [2] who called them bimatroids and the elements of Λ nonsingular minors. Actually, Schrijver and Kung defined linking systems with axioms. They then proved every linking system so defined is associated with a matroid as above and conversely.

Kung's axioms consist of $(\emptyset \mid \emptyset) \in \Lambda$ and an axiom that is derived from the strong base exchange property applied to the associated matroid:

- (B3) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1$ then there is some $y \in B_2$ such that $(B_1 \setminus x) \cup y \in \mathcal{B}$ and $(B_2 \setminus y) \cup x \in \mathcal{B}$.

Schrijver's axioms are

- (L1) If $(X \mid Y) \in \Lambda$ and $x \in X$ then $(X \setminus x \mid Y \setminus y) \in \Lambda$ for some $y \in Y$.
- (L2) If $(X \mid Y) \in \Lambda$ and $y \in Y$ then $(X \setminus x \mid Y \setminus y) \in \Lambda$ for some $x \in X$.
- (L3) If $(X_1 \mid Y_1) \in \Lambda$ and $(X_2 \mid Y_2) \in \Lambda$ then for some X, Y such that

$$X_1 \subset X \subset X_1 \cup X_2 \text{ and } Y_2 \subset Y \subset Y_1 \cup Y_2, (X \mid Y) \in \Lambda.$$

Linking systems are also characterized by their linking function $\lambda : 2^{E_1} \times 2^{E_2} \rightarrow \mathbb{Z}^+$. $\lambda(X, Y)$ is the cardinality of the largest $X_1 \subset X$ and $Y_1 \subset Y$ such that $(X_1 \mid Y_1) \in \Lambda$. Note that these X_1, Y_1 have equal cardinality. λ is the abstraction of the rank function of submatrices. Linking functions are functions into \mathbb{Z} that satisfy

- (F1) $0 \leq \lambda(X, Y) \leq \min\{|X|, |Y|\}$.
- (F2) If $X' \subset X$ and $Y' \subset Y$ then $\lambda(X', Y') \leq \lambda(X, Y)$.
- (F3) $\lambda(X_1 \cap X_2, Y_1 \cup Y_2) + \lambda(X_1 \cup X_2, Y_1 \cap Y_2) \leq \lambda(X_1, Y_1) + \lambda(X_2, Y_2)$.

These axioms were given by Schrijver. Kung's axioms are equivalent but slightly different.

Since linking systems are matroids with a distinguished base, in a different guise, the reader may ask for the point of defining and using linking systems. Linking systems are interesting because they reveal and motivate aspects of matroid theory that would otherwise look strange and be obscure. For example, (L3) is the generalization for linking systems of the Dulmage–Mendelsohn theorem for matchings in bipartite graphs.

Important matroid analogies of some operations in linear algebra are clearly described with linking systems. For example if (E_1, E_2, Λ_{12}) and (E_2, E_3, Λ_{23}) are linking systems then (E_1, E_3, Λ_{13}) is a linking system where

$$\Lambda_{13} = \{(X \mid Y) \mid \text{there exists } Z \subset E_2 \text{ such that } (X \mid Z) \in \Lambda_{12} \text{ and } (Z \mid Y) \in \Lambda_{23}\}$$

(Schrijver [4], [5] and Kung [2]). $\Lambda_{13} = \Lambda_{12} \circ \Lambda_{23}$ is called the composition or bimatroid product. If matrix M_{12} represents Λ_{12} , M_{23} represents Λ_{23} , and G is a diagonal matrix with rows and columns E_2 and entries that are algebraically independent indeterminates, then $M_{13} = M_{12}GM_{23}$ represents Λ_{13} . This is proven using the Cauchy–Binet theorem: $(M(X \mid Y))$ denotes the minor of matrix M with rows X and columns Y .)

$$\det M_{13}(X \mid Y) = \sum_{\substack{Z \subset E_2 \\ |Z|=|X|}} \det M_{12}(X \mid Z) \det G(Z \mid Z) \det M_{23}(Z \mid Y).$$

Schrijver shows that inversion of a nonsingular matrix has a linking system analogue.

If (E_1, E_2, Λ) is a linking system for which $(E_1|E_2) \in \Lambda$ then (E_2, E_1, Λ^{-1}) below is a linking system:

$$(1.1) \quad \Lambda^{-1} = \{(Y|X) | (E_1 \setminus X | E_2 \setminus Y) \in \Lambda\}.$$

Proof. Λ^{-1} and Λ have the same associated matroid.

This is an analogue of matrix inversion because Jacobi’s theorem shows that for a nonsingular matrix M , $M^{-1}(Y|X)$ is nonsingular if and only if $M(E_1 \setminus X | E_2 \setminus Y)$ is nonsingular.

2. Summary of results. The main result of this paper is that another construction in linear algebra, the Bott–Duffin [1] constrained (or generalized) inverse, can be abstracted to matroid theory in the same sense matrix multiplication and inversion were abstracted. See also Rao and Mitra [3].

Let $\mathcal{M}_1 = (E_1, \mathcal{I}_1)$ and $\mathcal{M}_2 = (E_2, \mathcal{I}_2)$ be matroids and (E_1, E_2, \mathcal{G}) be a linking system. We will give a combinatorial definition (2.2) of a collection \mathcal{T} of pairs $(Y|X)$, for which $Y \subset E_2$ and $X \subset E_1$. We then show that when condition (2.1) holds, (2.2) is a linking system.

We will next consider when $\mathcal{M}_1, \mathcal{M}_2$, and \mathcal{G} respectively are “generically” coordinatized by cycle spaces $\mathcal{C}_1, \mathcal{C}_2$ and a matrix G . The condition (2.1) then turns out to be equivalent to the condition that the Bott–Duffin constrained inverse problem $Gv - i \in \mathcal{C}_1, x \in \mathcal{C}_2^\perp$ has a unique solution v for all i . When this is true, we will show that \mathcal{T} is coordinatized by the matrix T for which $v = Ti$ for all i .

Our condition for the existence of the Bott–Duffin inverse linking system of \mathcal{G} with respect to \mathcal{M}_1 and \mathcal{M}_2 , is

$$(2.1) \quad \mathcal{M}_1 \text{ has a base } F_1 \text{ and } \mathcal{M}_2 \text{ has a base } F_2 \text{ such that } (F_1|F_2) \in \mathcal{G}.$$

This implies that \mathcal{M}_1 and \mathcal{M}_2 have equal rank R .

THEOREM A. Suppose $\mathcal{M}_i = (E_i, \mathcal{I}_i)$, with rank function $r_i, i = 1, 2$ and (E_1, E_2, \mathcal{G}) , with linking function γ , satisfy the above condition. Then (E_2, E_1, \mathcal{T}) is a linking system with linking function τ , where

$$(2.2) \quad \begin{aligned} \mathcal{T} = \{ & (Y|X) | Y \in \mathcal{I}_2, X \in \mathcal{I}_1, \text{ and there exist } F_1 \subset E_1, F_2 \subset E_2 \text{ such that} \\ & Y \cap F_2 = X \cap F_1 = \emptyset, Y \cup F_2 \in \mathcal{B}_2, X \cup F_1 \in \mathcal{B}_1, \text{ and } (F_1|F_2) \in \mathcal{G}\}, \\ \tau(Y, X) = & \min_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} \{r_1(X \cup F_1) + \gamma(F_1^c, F_2^c) + r_2(Y \cup F_2)\} - R. \end{aligned}$$

Here $F_i^c = E_i \setminus F_i$.

The following two corollaries are special cases.

COROLLARY 1. If (E_1, E_2, Λ) is a linking system with $(E_1|E_2) \in \Lambda$, then (E_2, E_1, Λ^{-1}) where Λ^{-1} given by (1.1) is a linking system.

Proof. Take $\mathcal{G} = \Lambda$ and \mathcal{M}_i the (free) matroid on E_i for which all subsets of E_i are independent. Since E_i is a base (the only base) in \mathcal{M}_i and $(E_1|E_2) \in \Lambda$, the corollary follows from Theorem A.

COROLLARY 2. If $\mathcal{M} = (E, \mathcal{I})$ is a matroid, then $T(\mathcal{M}) = (E, E, \mathcal{T})$ is a linking system where

$$\begin{aligned} \mathcal{T} = \{ & (X|Y) | X, Y \in \mathcal{I} \text{ and there exists } F \in \mathcal{I} \\ & \text{such that } F \cap (X \cup Y) = \emptyset \text{ and } F \cup X, F \cup Y \text{ are bases in } \mathcal{M}\}. \end{aligned}$$

Proof. Take $\mathcal{G} = \{(X|X) | X \subset E\}$ and $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}$. Clearly the hypotheses of Theorem A hold and \mathcal{T} is as claimed because $(F_1|F_2) \in \mathcal{G}$ if and only if $F_1 = F_2$.

Corollary 2 is an interesting fact about the basis exchange or pivoting operation in matroids. Let $\mathcal{M} = (E, \mathcal{I})$ be a matroid on $E = E_0 \cup E'$ where E_0 is a base. Then, in the linking system (E_0, E', Λ) associated with \mathcal{M} with distinguished base E_0 , $(X|Y) \in \Lambda$ if and only if X can be exchanged for Y in (fixed) base E_0 , that is, $(E_0 \setminus X) \cup Y$ is a base in \mathcal{M} . However, in (E, E, \mathcal{T}) , $(X|Y) \in \mathcal{T}$ if and only if X can be exchanged for Y in *some* base in \mathcal{M} . The sets $(E_1 \setminus X_1) \cup Y_2$, where $E_i, i = 1, 2$ are disjoint copies of E and X_1, Y_2 are the corresponding images of X, Y when $(X|Y) \in \mathcal{T}$, comprise the bases of a new matroid on $E_1 \cup E_2$! Note that the rank of the new matroid equals the cardinality of the set of points E of \mathcal{M} . The independent sets \mathcal{I} are recovered from \mathcal{T} by $\mathcal{I} = \{X | (X|X) \in \mathcal{T}\}$.

Let us apply Corollary 2 to the polygon matroid of a (for simplicity) connected graph \mathcal{N} with edges E . The bases are the spanning trees. We obtain a new matroid associated with a graph.

COROLLARY 3. *Let E_1 and E_2 be two disjoint copies of E . Let \mathcal{B} be the collection of sets of the form $(E_1 \setminus X) \cup Y, X \subset E_1, Y \subset E_2$ for which*

- (1) *the sets of edges in \mathcal{N} corresponding to X and Y respectively are equicardinal forests, and*
- (2) *there exists a forest F in \mathcal{N} such that $F \cap (X \cup Y) = \emptyset, X \cup F$ is a spanning tree and $Y \cup F$ is a spanning tree.*

Then \mathcal{B} is the collection of bases of a matroid on $E_1 \cup E_2$.

Property (L3) implies the following result in graph theory.

COROLLARY 4. *Suppose $X_1 \cup F_1, Y_1 \cup F_1, X_2 \cup F_2, Y_2 \cup F_2$ are spanning trees and $X_i \cap F_i = Y_i \cap F_i = \emptyset$ for $i = 1, 2$. Then there exist forests $X, Y,$ and F such that $X_1 \subset X \subset X_1 \cup X_2, Y_2 \subset Y \subset Y_1 \cup Y_2, X \cap F = Y \cap F = \emptyset, F \subset X_1 \cup X_2 \cup Y_1 \cup Y_2 \cup F_1 \cup F_2,$ and $X \cup F, Y \cup F$ are both spanning trees.*

The linking system in Corollary 3 turns out to be coordinatized by the open circuit resistance matrix of the resistor network with underlying graph \mathcal{N} and resistors with algebraically independent values. To explain this and introduce the Bott–Duffin constrained inverse, we formulate the equations for the currents and voltages in a resistive network.

Let E be the set of edges of an electrical network \mathcal{N} of resistors. Let \mathcal{C} be the cycle space and \mathcal{C}^\perp be the cocycle space of the graph of \mathcal{N} . All vectors w here will be tuples indexed by the elements of E ; $w(e)$ is the component corresponding to $e \in E$. Suppose current $i_0(e)$ is applied across edge e in \mathcal{N} and we wish to compute the voltages that appear across the edges as a result. Let $i(e)$ be the current in edge e and $v(e)$ be the voltage across edge e .

Kirchhoff’s current law requires

$$i - i_0 \in \mathcal{C}.$$

Kirchhoff’s voltage law requires

$$v \in \mathcal{C}^\perp.$$

Ohm’s law relates v to i ,

$$i = Gv$$

where G is a diagonal matrix of edge conductances. (Conductance is the reciprocal of resistance.) Thus, we have an example of a constrained inverse problem. When for

all i_0 the solution v exists and is unique, the constrained inverse T exists and

$$v = Ti_0.$$

Here, T is called the open-circuit resistance or “transpedence” matrix of \mathcal{N} .

We consider the following slight generalization of Bott and Duffin’s constrained inverse problem.

Let E_1, E_2 be two finite sets which are fixed coordinate sets for vector spaces $\mathcal{E}_1, \mathcal{E}_2$. Let \mathcal{C}_1 and \mathcal{C}_2 be given subspaces of \mathcal{E}_1 and \mathcal{E}_2 respectively. Let $G : \mathcal{E}_2 \rightarrow \mathcal{E}_1$ be a linear map from \mathcal{E}_2 to \mathcal{E}_1 . G can be considered here to be an $E_1 \times E_2$ matrix.

Given $i_0 \in \mathcal{E}_1$ we wish to find $v \in \mathcal{E}_2$ such that

$$(2.3) \quad Gv - i_0 \in \mathcal{C}_1,$$

$$(2.4) \quad v \in \mathcal{C}_2^\perp$$

where \mathcal{C}^\perp is the orthogonal complement of \mathcal{C} . When these relations have a unique solution for all $i_0 \in \mathcal{E}_1$ then it can be shown there is a matrix T such that $v = Ti_0$.

The orthogonal complement is taken with respect to the usual scalar product. Let $\mathcal{M}_1, \mathcal{M}_2$ be the matroids on E_1, E_2 respectively that are coordinatized by cycle spaces $\mathcal{C}_1, \mathcal{C}_2$. (In other words, \mathcal{M}_i is represented by chain group \mathcal{C}_i .) Let B_1, B_2 be bases in $\mathcal{M}_1, \mathcal{M}_2$ respectively. Let M_{B_1}, M_{B_2} respectively be the fundamental cocycle matrices of $\mathcal{M}_1, \mathcal{M}_2$ with respect to B_1, B_2 . That is, M_{B_i} has rows B_i and columns E_i . Row f of M_{B_i} is the unique vector v in \mathcal{C}_i^\perp such that

$$v(f) = 1, \quad v(e) = 0 \quad \text{for all } e \in B_i \setminus f.$$

The following lemma is an easy generalization of the results of Bott and Duffin.

LEMMA. Equations (2.3) and (2.4) have a unique solution v for all $i_0 \in \mathcal{E}_2$ if and only if

$$\det(M_{B_1}GM_{B_2}^t) \neq 0.$$

If so, $v = Ti_0$ where

$$T = M_{B_2}^t(M_{B_1}GM_{B_2}^t)^{-1}M_{B_1}.$$

G represents or coordinatizes \mathcal{G} when $(X|Y) \in \mathcal{G}$ if and only if submatrix $G(X|Y)$ is nonsingular.

Our connection between the Bott–Duffin inverse and the linking system in Theorem A holds when G coordinatizes \mathcal{G} generically. That is, assume G has the form

$$(2.5) \quad G = \begin{pmatrix} h_1 & & & 0 \\ & h_2 & & \\ 0 & & \dots & \\ & & & h_{|E_1|} \end{pmatrix} \cdot G' \cdot \begin{pmatrix} k_1 & & & 0 \\ & k_2 & & \\ 0 & & \dots & \\ & & & k_{|E_2|} \end{pmatrix}$$

where h_i and k_j are algebraically independent indeterminates.

THEOREM B. When G is generic, (2.3) and (2.4) have a unique solution $v = Ti_0$ for all $i_0 \in \mathcal{E}_1$ if and only if there are bases F_1 in \mathcal{M}_1 and F_2 in \mathcal{M}_2 so that $(F_1|F_2) \in \mathcal{G}$. (Hence $\text{rank}(\mathcal{M}_1) = \text{rank}(\mathcal{M}_2)$. When G is arbitrary, the condition is necessary.) When G is “generic” and the condition holds, T represents the linking system (E_2, E_1, \mathcal{T}) in Theorem A.

Note that when G is a nonsingular, generic diagonal matrix and $\mathcal{C}_1 = \mathcal{C}_2$ the condition holds automatically. In this case, T coordinatizes the linking system in Corollary 2.

3. Proof of the general matroid result. Let (E_1, E_2, \mathcal{G}) be a linking system with linking function γ . Let $\mathcal{M}_i = (E_i, \mathcal{F}_i)$, $i = 1, 2$ be matroids with rank functions r_i respectively.

If \mathcal{M} is a matroid on E , $r_{\mathcal{M}}$ denotes \mathcal{M} 's rank function. If $S \subset E$, \mathcal{M}/S denotes \mathcal{M} with S contracted. We take \mathcal{M}/S to be a matroid on E ; the elements of S are loops in \mathcal{M}/S . Thus, for $X \subset E$,

$$(3.1) \quad r_{\mathcal{M}/S}(X) = r_{\mathcal{M}}(X \cup S) - r_{\mathcal{M}}(S).$$

We assume the reader is familiar with the elementary properties of matroid rank functions, see Welsh [6]. In particular, $X \in \mathcal{F}$ if and only if $r(X) = |X|$.

Throughout this section, assume

HYPOTHESIS. \mathcal{M}_1 has a base F_1 and \mathcal{M}_2 has a base F_2 such that $(F_1|F_2) \in \mathcal{G}$. This implies that \mathcal{M}_1 and \mathcal{M}_2 have equal rank R .

DEFINITION. Let $T \subset E_2$ and $S \subset E_1$. Denote $E_i \setminus F_i$ by F_i .

$$(3.2) \quad \tau(T, S) = r_2(T) + r_1(S) - R + \min_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} \{r_{\mathcal{M}_1/S}(F_1) + \gamma(F_1^c, F_2^c) + r_{\mathcal{M}_2/T}(F_2)\}.$$

Observe from (3.1) that

$$(3.3) \quad \tau(T, S) = \min_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} \{r_1(F_1 \cup S) + \gamma(F_1^c, F_2^c) + r_2(F_2 \cup T)\} - R.$$

Our proof hinges upon Schrijver's characterization of linking functions (F1), (F2), and (F3) and his generalization of Edmond's intersection theorem:

THEOREM (Schrijver). Let $\mathcal{N}_i = (E_i, J_i)$ $i = 1, 2$ be matroids with rank functions ρ_i and let (E_1, E_2, \mathcal{G}) be a linking system with linking function γ . Then the maximum cardinality of sets $F_1 \in J_1$ and $F_2 \in J_2$ such that $(F_1|F_2) \in \mathcal{G}$ equals

$$(3.4) \quad \min_{\substack{H_1 \subset E_1 \\ H_2 \subset E_2}} \{\rho_1(H_1) + \gamma(H_1^c, H_2^c) + \rho_2(H_2)\}.$$

THEOREM 1. τ satisfies (F1), (F2), and (F3). Hence τ is the linking function of a linking system.

Proof. Clearly, from (3.3) τ is nondecreasing in T and S , so (F2) is satisfied.

To show $\tau(\emptyset, \emptyset) = 0$, we note from (3.3) that

$$\tau(\emptyset, \emptyset) = \min_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} \{r_1(F_1) + \gamma(F_1^c, F_2^c) + r_2(F_2)\} - R.$$

We conclude from the hypothesis that the largest F_1, F_2 such that $F_1 \in \mathcal{F}_1, F_2 \in \mathcal{F}_2$ and $(F_1|F_2) \in \mathcal{G}$ are both bases in $\mathcal{M}_1, \mathcal{M}_2$ respectively. Hence we conclude $\tau(\emptyset, \emptyset) = 0$ from (3.4) and $r_1(\mathcal{M}_1) = r_2(\mathcal{M}_2) = R$.

If we evaluate the right-hand expression in (3.3) for $F_1 = E_1, F_2 = \emptyset$ and apply $r_1(E_1) = R$, we obtain $\tau(T, S) \leq r_2(T)$. Similarly, $\tau(T, S) \leq r_1(S)$. Hence

$$(3.5) \quad \tau(T, S) \leq \min \{r_2(T) \cdot r_1(S)\}.$$

This stronger result establishes (F1).

Finally, we prove the submodularity property (F3) of τ for given $T', T'' \subset E_2$ and $S', S'' \subset E_1$. Suppose F'_1 and F'_2 attain the minimum (3.3) for T', S' , and F''_1 and F''_2

attain the minimum (3.3) for T'', S'' . Then

$$\begin{aligned} \tau(T', S') + \tau(T'', S'') &= r_1(F'_1 \cup S') + \gamma(F_1^c, F_2^c) + r_2(F'_2 \cup T') - R \\ &\quad + r_1(F''_1 \cup S'') + \gamma(F_1^c, F_2^c) + r_2(F''_2 \cup T'') - R \\ &\cong r_1((F'_1 \cap F''_1) \cup (S' \cap S'')) \\ &\quad + \gamma((F'_1 \cap F''_1)^c, (F'_2 \cup F''_2)^c) + r_2((F'_2 \cup F''_2) \cup (T' \cup T'')) - R \\ &\quad + r_1((F'_1 \cup F''_1) \cup (S' \cup S'')) \\ &\quad + \gamma((F'_1 \cup F''_1)^c, (F'_2 \cap F''_2)^c) + r_2((F'_2 \cap F''_2) \cup (T' \cap T'')) - R \\ &\cong \tau(T' \cup T'', S' \cap S'') + \tau(T' \cap T'', S' \cup S''). \end{aligned}$$

The first inequality follows from the submodularity of r_1, r_2 and γ . The second inequality follows from (3.3). QED.

Let \mathcal{T} be the set of pairs

$$(3.6) \quad \mathcal{T} = \{(Y|X) \mid Y \in \mathcal{J}_2, X \in \mathcal{J}_1 \text{ and there exist } F_1 \subset E_1, F_2 \subset E_2 \text{ such that } Y \cap F_2 = X \cap F_1 = \emptyset, Y \cup F_2 \in \mathcal{B}_2, X \cup F_1 \in \mathcal{B}_1, \text{ and } (F_1|F_2) \in \mathcal{G}\}.$$

THEOREM 2. $\tau(Y, X) = |Y| = |X|$ if and only if $(Y|X) \in \mathcal{T}$.

Proof. Suppose $(Y|X) \in \mathcal{T}$. Then $r_2(Y) = r_1(X) = |X| = |Y|$, and by (3.4)

$$\min_{\substack{G_1 \subset E_1 \\ G_2 \subset E_2}} \{r_{\mathcal{M}_1/X}(G_1) + \gamma(G_1^c, G_2^c) + r_{\mathcal{M}_2/Y}(G_2)\} \cong |F_1| = |F_2|$$

where F_1, F_2 are from (3.6). Since $|F_1| = |F_2| = R - |X|$ we know $\tau(Y, X) \cong |X| + |Y| - R + R - |X| = |Y|$. By (F1), $\tau(Y, X) = |X| = |Y|$.

Now suppose $\tau(Y, X) = |Y| = |X|$. Then by (3.5), $X \in \mathcal{J}_1$ and $Y \in \mathcal{J}_2$. Hence

$$\min_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} \{r_{\mathcal{M}_1/X}(F_1) + \gamma(F_1^c, F_2^c) + r_{\mathcal{M}_2/Y}(F_2)\} = R - |X|.$$

By (3.4) there exist F_1, F_2 with $(F_1|F_2) \in \mathcal{G}$, F_1 is independent in \mathcal{M}_1/X , F_2 is independent in \mathcal{M}_2/Y with $|F_1| = |F_2| = R - |X|$. Hence $F_1 \cap X = F_2 \cap Y = \emptyset$ and $F_1 \cup X, F_2 \cup Y$ are respectively bases in $\mathcal{M}_1, \mathcal{M}_2$. We conclude $(Y|X) \in \mathcal{T}$. QED.

4. Proofs of the coordinatization results. Throughout this chapter, assume G is generic, that is, assume (2.5). Let h_{F_1} and k_{F_2} denote products of the indeterminates corresponding to $F_1 \subset E_1$ and $F_2 \subset E_2$. The consequence of the assumption and the Cauchy–Binet theorem is that

$$\det M_B, GM'_{B_2}(W|U) = \sum_{\substack{F_1 \subset E_1 \\ F_2 \subset E_2}} \det M_{B_1}(W|F_1) \det G'(F_1|F_2) \det M_{B_2}(U|F_2) h_{F_1} k_{F_2}$$

is nonzero if and only if for some $F_1 \subset E_1$ and $F_2 \subset E_2$, F_1 is a base in $\mathcal{M}_1/(B_1 \setminus W)$, F_2 is a base in $\mathcal{M}_2/(B_2 \setminus U)$ and $(F_1|F_2) \in \mathcal{G}$.

Assume G coordinatizes \mathcal{G} and $\mathcal{M}_1, \mathcal{M}_2$ respectively are coordinatized by cycle spaces $\mathcal{C}_1, \mathcal{C}_2$.

LEMMA. *The Bott–Duffin inverse problem (2.3)–(2.4) has a unique solution v for all $i_0 \in \mathcal{E}_1$ if and only if \mathcal{M}_1 and \mathcal{M}_2 have equal rank R and there are bases $B_1 \in \mathcal{B}_1$ and $B_2 \in \mathcal{B}_2$ such that $(B_1|B_2) \in \mathcal{G}$ (i.e. $G(B_1|B_2)$ is nonsingular).*

Proof. Let B'_1, B'_2 be arbitrary bases in $\mathcal{M}_1, \mathcal{M}_2$ respectively and $M_{B'_1}, M_{B'_2}$ be

fundamental cocycle matrices (§ 2). For every $v \in \mathcal{C}_2^\perp$ there is a unique ρ for which

$$(4.1) \quad v = M_{B_2}^t \rho.$$

Hence (2.3)–(2.4) have a unique solution if and only if

$$(4.2) \quad M_{B_1} G M_{B_2}^t \rho - M_{B_1} i_0 = 0$$

has a unique solution ρ .

Suppose (4.2) has a unique solution ρ for all i_0 . Since (4.2) has a solution for all i_0 , $\text{image}(M_{B_1}) = \text{image}(M_{B_1} G M_{B_2}^t)$ and so $r_1(E_1) = \text{rank}(M_{B_1}) = \text{rank}(M_{B_1} G M_{B_2}^t) \leq \text{rank}(M_{B_2}) = r_2(E_2)$. By uniqueness, $\ker(M_{B_1} G M_{B_2}^t) = (0)$ so $r_1(E_1) = |B_1| \cong |B_2| = r_2(E_2)$. Hence $M_{B_1} G M_{B_2}^t$ is a square, nonsingular matrix. The Cauchy–Binet theorem implies there are bases B_1, B_2 as claimed.

Conversely, suppose the hypotheses about $\mathcal{M}_1, \mathcal{M}_2$ and \mathcal{G} are true. Then (because G is generic) $M_{B_1} G M_{B_2}^t$ is nonsingular. Hence (4.2) has a unique solution. QED.

THEOREM 3. *When G is generic, (2.3)–(2.4) have a unique solution $v = T i_0$ for all $i_0 \in \mathcal{C}_2$ if and only if the condition for the existence of the Bott–Duffin constrained inverse linking system (E_2, E_1, \mathcal{T}) for $\mathcal{M}_1, \mathcal{M}_2$ and \mathcal{G} is true. When the condition holds, T coordinatizes (E_2, E_1, \mathcal{T}) .*

Proof. When the condition in the lemma is true, the unique constrained inverse matrix T is clearly

$$(4.3) \quad T = M_{B_2}^t (M_{B_1} G M_{B_2}^t)^{-1} M_{B_1}.$$

B_1, B_2 are arbitrary bases in $\mathcal{M}_1, \mathcal{M}_2$ respectively.

We must show $T(Y|X)$ is nonsingular if and only if $(Y|X) \in \mathcal{T}$.

Suppose $T(Y|X)$ is nonsingular. X cannot contain a circuit in \mathcal{M}_1 because otherwise there would be a cycle $i_0 \in \mathcal{C}_1$ whose support (set of nonzero coordinates) is a nonempty subset of X . Then $T i_0 = 0$ so ${}^1 T(Y|X) i_0(X) = 0$ with $i_0(X) \neq 0$. Likewise, Y cannot contain a circuit in \mathcal{M}_2 . Otherwise, there would be a nonzero cycle c in \mathcal{C}_2 with support in Y , so $c M_{B_2}^t = c(Y) M_{B_2}^t (Y|B_2) = 0$ so $c(Y) T(Y|X) = 0$ with $c(Y) \neq 0$. Hence there are bases $B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2$ such that $X \subset B_1$, and $Y \subset B_2$.

Let us use these bases in (4.3). We get

$$(4.4) \quad T(Y|X) = (M_{B_1} G M_{B_2}^t)^{-1} (Y|X)$$

because $M_{B_1}(B_1|B_1)$ and $M_{B_2}(B_2|B_2)$ are identity matrices. By Jacobi’s theorem,

$$(4.5) \quad (M_{B_1} G M_{B_2}^t)(B_1 \setminus X | B_2 \setminus Y)$$

is nonsingular. Therefore, by the Cauchy–Binet theorem there are bases F_1 in \mathcal{M}_1/X and F_2 in \mathcal{M}_2/Y such that $(F_1|F_2) \in \mathcal{G}$. Hence $(Y|X) \in \mathcal{T}$.

Conversely, suppose $(Y|X) \in \mathcal{T}$. Let $B_1 = X \cup F_1$ and $B_2 = Y \cup F_2$ be the bases from (2.2). Hence $T(Y|X)$ is given by (4.4). The Cauchy–Binet theorem implies (4.5) is nonsingular because $(F_1|F_2) \in \mathcal{G}$ and G is generic. Jacobi’s theorem implies $T(Y|X)$ is nonsingular. QED.

Acknowledgment. The author wishes to thank the referee for simplifications of some of the proofs.

¹ $i_0(X)$ is the “submatrix” of the column i_0 with rows X .

REFERENCES

- [1] R. BOTT AND R. J. DUFFIN, *On the algebra of networks*, Trans. Amer. Math. Soc., 74 (1953), pp. 99–109.
- [2] J. P. S. KUNG, *Bimatroids and invariants*, Advances in Math., 30 (1978), pp. 238–249.
- [3] C. R. RAO AND S. K. MATRA, *Theory and application of constrained inverses of matrices*, SIAM J. Appl. Math., 24 (1973), pp. 473–488.
- [4] A. SCHRIJVER, *Matroids and linking systems*, Math. Centre Tract 88, Mathematical Centre, Amsterdam, 1978.
- [5] ———, *Matroids and linking systems*, J. Combin. Theory B, 26 (1979), pp. 349–369.
- [6] D. J. A. WELSH, *Matroid Theory*, Academic Press, New York, 1976.

A CLASS OF M -MATRICES WITH TREE GRAPHS*

CHARLES R. JOHNSON,[†] D. D. OLESKY[‡] AND P. VAN DEN DRIESSCHE[§]

Abstract. A class of matrices is considered for which the directed graphs have a longest simple circuit of length two; for an irreducible matrix this means that its undirected graph is a tree. A matrix in this class which is both positive stable and inverse nonnegative is proved to be an M -matrix. A characterization is given of those inverse M -matrices for which the corresponding M -matrix lies in this class. The results are related to known theorems on tridiagonal matrices.

1. Introduction. For the class of tridiagonal matrices many results are known regarding inverses and eigenvalues. Some of these depend crucially on the circuit structure of the graph of the matrix. For example, Johnson [6] has recently demonstrated that some bounds for spectral radii depend on the length of the longest simple circuit. The aim of our work is to similarly extend known results for tridiagonal matrices to the class of matrices having a longest simple circuit of length two. Before stating these in more detail, some notation from matrix theory and graph theory is essential.

Throughout this paper $A \equiv [a_{ij}]$ is a square nonsingular matrix of order n . If $a_{ij} \leq 0$ for all $i \neq j$, then A belongs to class \mathcal{L} . If $A^{-1} \equiv [\alpha_{ij}]$ has all elements $\alpha_{ij} \geq 0$, then A is inverse nonnegative (monotone), and we write $A^{-1} \geq 0$. The class of matrices which have the \mathcal{L} sign pattern and are inverse nonnegative is the well-known class of nonsingular M -matrices, which we denote by \mathcal{M} ; for properties of these matrices see [2]. Several recent papers (e.g. [4], [9], [12]) are concerned with characterizations and properties of those nonnegative matrices which have inverses in \mathcal{M} . If $A \in \mathcal{M}$ we call A^{-1} an inverse M -matrix. We are also interested in matrices which are both inverse nonnegative and positive stable (that is, have all eigenvalues in the right half plane), but without the \mathcal{L} sign pattern restriction. We call a monotone positive stable matrix an N -matrix and use \mathcal{N} to denote the class of all such matrices. N -matrices were studied in [11], where it was shown that among tridiagonal matrices the classes \mathcal{N} and \mathcal{M} are identical.

The (undirected) graph of A , denoted by $G(A)$, has vertex set $\mathcal{V} = \{1, 2, \dots, n\}$ and edge set $\mathcal{E} = \{(i, j): i \neq j, \text{ and at least one of } a_{ij} \text{ and } a_{ji} \text{ is nonzero}\}$. The valence (degree) of a vertex denotes the number of edges incident to that vertex. Vertices with valence greater than one are called interior vertices.

The directed graph of A , $D(A)$, has the same vertex set as $G(A)$, but has a directed edge from vertex i to vertex $j \neq i$, denoted by (i, j) , if and only if $a_{ij} \neq 0$. A path from i to j is a sequence of edges $(i_1, i_2), (i_2, i_3), \dots, (i_{p-1}, i_p)$, with $i_1 = i$ and $i_p = j$. The graph $D(A)$ is strongly connected if there is a path from i to j for each ordered pair i, j . It is well known that a matrix A is irreducible if and only if $D(A)$ is strongly connected. A simple circuit in $D(A)$ is a path for which $i_p = i_1$ but

* Received by the editors August 2, 1982, and in revised form December 1, 1982.

[†] Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The work of this author was supported by Air Force Wright Aeronautical Laboratory contract F33615-81-K-3224 and the National Science Foundation under grant MCS-80-01611.

[‡] Department of Computer Science, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada. The work of this author was partially supported by the Natural Sciences and Engineering Research Council under grant A-8214.

[§] Department of Mathematics, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada. The work of this author was partially supported by the Natural Sciences and Engineering Research Council under grant A-4645.

i_1, i_2, \dots, i_{p-1} are distinct vertices. The length of the longest simple circuit in $D(A)$ is denoted by $m(A)$; for example, an irreducible tridiagonal matrix A has $m(A) = 2$.

With the notation now established, we can focus on the main results of our paper. If A is an irreducible matrix in class \mathcal{N} with $m(A) = 2$, then we prove in § 2 that A is in class \mathcal{M} . Obviously \mathcal{M} is a subset of \mathcal{N} , so for the matrices specified these two classes are identical. We also give a characterization of irreducible inverse M -matrices for which the corresponding M -matrix has a certain graph structure. In § 3 the reducible case is investigated, while § 4 includes some examples and a discussion of our work in relation to tridiagonal matrices.

2. Irreducible matrices. We are concerned with irreducible matrices A having $m(A) = 2$, and we denote the class of all such matrices by \mathcal{G} . This class of matrices may be defined using other graph theoretic terms; for example, Maybee [10] notes that $A \in \mathcal{G}$ if and only if A is combinatorially symmetric (i.e., $a_{ij} \neq 0$ implies $a_{ji} \neq 0$) and $G(A)$ is a tree. Our results are also related to those of [7], where the term “tree-diagonal matrix” is used. If $G(A)$ is a tree, then A is called a tree-diagonal matrix; if $A \in \mathcal{G}$, it is clear that A is tree-diagonal.

We now prove our main result.

THEOREM 1. *Let A be a nonsingular irreducible matrix with $m(A) = 2$. Then the following are equivalent:*

- (i) A is an M -matrix;
- (ii) A is an N -matrix;
- (iii) $A^{-1} = [\alpha_{ij}] \geq 0$, and $\alpha_{ii}\alpha_{jj} - \alpha_{ij}\alpha_{ji} > 0$ for all (i, j) in the edge set of $G(A)$.

Proof. We first prove the equivalence of (i) and (iii), and then complete the proof by showing that (i) is equivalent to (ii).

If A is an irreducible M -matrix, then it is well known that $A^{-1} > 0$, and since all principal minors of an inverse M -matrix are positive (see [4]), we have (i) implies (iii). To prove the converse, we use [7, Thm. 1], which specifies the entries of the inverse of a matrix A which has a tree graph. Specifically, for $a_{ij} \neq 0$ and $i \neq j$,

$$(1) \quad \alpha_{ij} = -\frac{a_{ij} \det A(i, j|i, j)}{\det A},$$

where $\det A(i, j|i, j)$ denotes the determinant of the principal submatrix of A obtained by deleting rows and columns i and j from A . By a well-known formula for determinants of inverse matrices (see e.g. [3, p. 21]),

$$\frac{\det A(i, j|i, j)}{\det A} = \alpha_{ii}\alpha_{jj} - \alpha_{ij}\alpha_{ji}.$$

Thus it follows from (1) and condition (iii) that $A \in \mathcal{X}$ and consequently $A \in \mathcal{M}$.

It is clear that (i) implies (ii); however, the proof of the converse requires several steps. We first show that A is sign symmetric and then that A can be diagonally symmetrized. From this it follows that all principal minors of A are positive, and this enables us to show that $A \in \mathcal{X}$.

To accomplish these steps, we begin by noting that since A is irreducible with $m(A) = 2$, $G(A)$ is a tree and A is combinatorially symmetric (see [10]). If we assume that $a_{pq} \neq 0$, $p \neq q$, the combinatorial symmetry of A then implies that $a_{qp} \neq 0$. Let $\hat{A} = [\hat{a}_{ij}]$ be defined by

$$\hat{a}_{ij} = \begin{cases} 0 & (i = p \text{ and } j = q) \text{ or } (i = q \text{ and } j = p), \\ a_{ij} & \text{otherwise.} \end{cases}$$

Then there exists a permutation matrix P such that $\hat{B} = P\hat{A}P'$ is the direct sum of two square submatrices B_{11} and B_{22} . In terms of the graph, this is equivalent to removing the edge between vertices p and q , reducing $G(A)$ to two disconnected subtrees. Now

$$B = PAP' = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where B_{11} and B_{22} are as above; B_{12} and B_{21} each contain only one nonzero element. Since B is irreducible, it now follows from the lemma in [5] that B^{-1} cannot be nonnegative if $B_{12} \leq 0$ and $B_{21} \geq 0$. Thus $A^{-1} = P'B^{-1}P$ cannot be nonnegative unless $a_{pq}a_{qp} > 0$.

With the sign symmetry of A established, we can diagonally symmetrize A (see [10, Thm. 3]), and in fact there exists a positive diagonal matrix D such that $C = D^{-1}AD$ is symmetric. As A is assumed to be positive stable, matrix C is positive definite; thus both C and A have all their principal minors positive.

To complete the proof, we use the relationship (1) above, in which it is assumed that $a_{ij} \neq 0$ and $i \neq j$. From our assumption that $A^{-1} \geq 0$ and the fact that all principal minors of A are positive, we obtain $a_{ij} < 0$. Thus we have established that $A \in \mathcal{X}$, and hence A is an M -matrix. \square

We note that $A \in \mathcal{N}$ implies only that $A^{-1} \geq 0$ and that the inverse of an irreducible N -matrix may contain zeros (see [11]). However, it follows from Theorem 1 that if A is an irreducible N -matrix with $m(A) = 2$, then $A^{-1} > 0$.

We now give a restatement of our theorem as it relates to nonnegative matrices.

COROLLARY 1. *Let A be a nonsingular irreducible matrix with $m(A) = 2$ and $A^{-1} \equiv [\alpha_{ij}] \geq 0$. Then the following are equivalent:*

- (i)' $A \in \mathcal{X}$;
- (ii)' A is positive stable;
- (iii)' $\alpha_{ii}\alpha_{jj} - \alpha_{ij}\alpha_{ji} > 0$ for all (i, j) in the edge set of $G(A)$.

3. Reducible matrices. If we relax the condition of irreducibility while retaining the other assumptions, then the full strength of Theorem 1 no longer holds. If A is a reducible M -matrix, then $A \in \mathcal{N}$, and conditions (iii) are true. However, the example

$$\begin{bmatrix} 2 & -1 & -3 \\ -1 & 2 & 1.5 \\ 0 & 0 & 3 \end{bmatrix},$$

which satisfies (iii), is in \mathcal{N} , but is not in \mathcal{M} , illustrates that the converses are no longer true. Any reducible matrix is permutation similar to a block triangular matrix, say A . If this matrix is in \mathcal{N} , then clearly the main diagonal irreducible submatrix blocks are also in \mathcal{N} , and hence if $m(A) = 2$ are in \mathcal{M} . But the off-diagonal submatrix entries are not necessarily nonpositive, as illustrated by the above example.

In spite of this example, the following result extends Theorem 1 to certain reducible matrices A , which have either $m(A) = 2$ or have no simple circuits.

THEOREM 2. *Let A denote a matrix obtainable by setting to zero any number of off-diagonal elements of a nonsingular irreducible matrix \tilde{A} with $m(\tilde{A}) = 2$. Then (i), (ii) and (iii) of Theorem 1 are equivalent for A .*

Proof. By virtue of Theorem 1, only the case that A is reducible needs to be considered, and we proceed as in that proof.

It is clear that (i) implies (iii). The proof of the converse parallels that for the irreducible case, the form of A insuring that $G(A)$ is either a tree or a forest, so that [7, Thm. 1] may again be applied to show that $A \in \mathcal{X}$.

That (i) implies (ii) is also clear. To prove the reverse implication, we may assume without loss of generality that

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where A_{11} and A_{22} are square submatrices and A_{12} contains at most one nonzero element, since this form may be attained by a permutation similarity transformation. If matrix $A \in \mathcal{N}$, then A_{11} and A_{22} are in \mathcal{N} ; so $A^{-1} \geq 0$, $A_{11}^{-1} \geq 0$ and $A_{22}^{-1} \geq 0$ imply that $A_{12} \leq 0$. If either A_{11} or A_{22} is irreducible, then it is an M -matrix by Theorem 1. If either submatrix is reducible, the above argument may be repeated (as many times as necessary) in order to show that all off-diagonal elements of A are nonpositive, implying that $A \in \mathcal{M}$. \square

Note that if the matrix \tilde{A} in Theorem 2 is an M -matrix, then $A \in \mathcal{X}$ and $A - \tilde{A} \geq 0$, so that A is also an M -matrix.

4. Discussion. Our results show that all irreducible (and some reducible) matrices A in class \mathcal{N} with a certain circuit structure must be M -matrices and thus inherit all the properties known for this class. However, in general, \mathcal{M} is a proper subset of \mathcal{N} ; for example, there exist order 3 triangular N -matrices, and N -matrices with $m(A) = 3$, which are not M -matrices (see [11]). For an irreducible matrix $A \in \mathcal{N}$ with $m(A) = 2$, it is true that $A^p \in \mathcal{N}$ for any positive integer p , but, in general, A^p is not an M -matrix.

Irreducible tridiagonal matrices are a well studied set of matrices with a longest simple circuit of length two. Another set of matrices with this property consists of those matrices with “star” graphs. An example is

$$\begin{bmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1n} \\ s_{21} & s_{22} & & & \\ s_{31} & & s_{33} & 0 & \\ \vdots & 0 & & \ddots & \\ s_{n1} & & & & s_{nn} \end{bmatrix}.$$

In our proof of the final implication in Theorem 1, matrix A is shown to be similar to a symmetric matrix; thus all eigenvalues of A are real and positive. This, together with eigenvalue simplicity, is well known for tridiagonal matrices, but the eigenvalues of a matrix with a star graph need not be simple. Note that the matrices characterized by Theorem 1 are invariant under permutation similarity transformations, a property not shared by tridiagonal M -matrices.

For tridiagonal matrices, Theorem 1 may be combined with a result of Lewin [8, Thm. 2] to obtain the following:

THEOREM 3. *Let A be a nonsingular irreducible tridiagonal matrix. Then the following are equivalent:*

- (i) A is an M -matrix;
- (ii) A is an N -matrix;
- (iii) $A^{-1} = [\alpha_{ij}] \geq 0$,
and $\alpha_{ii}\alpha_{jj} - \alpha_{ij}\alpha_{ji} > 0$ for all (i, j) in the edge set of $G(A)$;
- (iv) A^{-1} is oscillatory.

The restriction of Theorem 1 to tridiagonal matrices also yields a simple inverse M -matrix characterization. On modifying condition (iii) to reflect the fact that $(i, j) \in \mathcal{E}$ if and only if $j = i + 1$, we obtain the following:

COROLLARY 2. *Let A be a nonsingular irreducible tridiagonal matrix. Then A is an M -matrix if and only if $A^{-1} = [\alpha_{ij}] \geq 0$ and*

$$\alpha_{ii}\alpha_{i+1,i+1} - \alpha_{i,i+1}\alpha_{i+1,i} > 0 \quad \text{for } i = 1, 2, \dots, n-1.$$

We note that this corollary may also be proved directly using the triangle property of Barrett [1].

In conclusion, a subclass of matrices in \mathcal{N} , namely those with tree graphs, has been identified with those in \mathcal{M} ; and we have given a characterization of those nonnegative matrices which are their inverses. The circuit structure is vitally important to these results and sheds some light on the broader question of when the two classes \mathcal{M} and \mathcal{N} are identical.

Acknowledgment. The authors thank M. Lewin for constructive suggestions on the presentation of these results.

REFERENCES

- [1] W. W. BARRETT, *A theorem on inverses of tridiagonal matrices*, *Linear Algebra Appl.*, 27 (1979), pp. 211–217.
- [2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
- [3] F. R. GANTMACHER, *The Theory of Matrices I*, Chelsea, New York, 1959.
- [4] C. R. JOHNSON, *Inverse M -matrices*, *Linear Algebra Appl.*, 47 (1982), pp. 195–216.
- [5] ———, *Sign patterns of inverse nonnegative matrices*, *Linear Algebra Appl.* (1983), to appear.
- [6] ———, *Principal submatrices and bounds for the spectral radius of a sparse matrix*, preprint (1982).
- [7] D. J. KLEIN, *Treediagonal matrices and their inverses*, *Linear Algebra Appl.*, 42 (1982), pp. 109–117.
- [8] M. LEWIN, *Totally nonnegative, M -, and Jacobi matrices*, *SIAM J. Alg. Disc. Meth.*, 1 (1980), pp. 419–421.
- [9] M. LEWIN AND M. NEUMANN, *On the inverse M -matrix problem for $(0, 1)$ -matrices*, *Linear Algebra Appl.*, 30 (1982), pp. 41–50.
- [10] J. S. MAYBEE, *Combinatorially symmetric matrices*, *Linear Algebra Appl.*, 8 (1974), pp. 529–537.
- [11] D. D. OLESKY AND P. VAN DEN DRIESSCHE, *Monotone positive stable matrices*, *Linear Algebra Appl.* (1983), to appear.
- [12] R. A. WILLOUGHBY, *The inverse M -matrix problem*, *Linear Algebra Appl.*, 18 (1977), pp. 75–94.

THE ALGEBRAIC GEOMETRY OF STRESSES IN FRAMEWORKS*

NEIL L. WHITE[†] AND WALTER WHITELEY[‡]

Abstract. A bar-and-joint framework, with rigid bars and flexible joints, is said to be generically isostatic if it has just enough bars to be infinitesimally rigid in some realization in Euclidean n -space. We determine the equation that must be satisfied by the coordinates of the joints in a given realization in order to have a nonzero stress, and hence an infinitesimal motion, in the framework. This equation, called the pure condition, is expressed in terms of certain determinants, called brackets. The pure condition is obtained by choosing a way to tie down the framework to eliminate the Euclidean motions, computing a bracket expression by a method due to Rosenberg and then factoring out part of the expression related to the tie-down. A major portion of this paper is devoted to proving that the resulting pure condition is independent of the tie-down chosen. We then catalog a number of small examples and their pure conditions, along with the geometric conditions for the existence of a stress which are equivalent to the algebraic pure conditions. We also explain our methods for calculating these conditions and determining their factorization. We then use the pure conditions to investigate stresses and tension-compression splits in 1-overbraced frameworks. Finally we touch briefly upon some of the problems arising when multiple factors occur in the pure condition.

The statics of bar-and-joint frameworks have been studied by mathematicians and engineers for over a century. Two divergent traditions of analysis have evolved: a) direct arithmetic calculations based on the specific positions of the joints and the bars and b) general synthetic geometric algorithms. With the rise of the computer and the decline of geometry, the arithmetic calculations became the dominant method of understanding static behavior in frameworks.

However, there has been a recent revival of interest in underlying projective geometric patterns of points and lines which “explain” the behavior of all the particular arithmetic examples based on the same underlying graph [2], [4], [21], [22]. Numerous types of graphs in space have been analyzed using synthetic geometry and certain more abstract patterns also emerge in the form of these geometric explanations.

Summarized in naive geometric terms, a count of “conditions” and “choices” is made, based only on the number of joints and bars in certain subgraphs. For example, in the plane a framework with V joints and E bars has, if $k = E - (2V - 2)$, at least a $(k + 1)$ -dimensional space of static stresses (if $k \geq 0$) for any position of the joints, i.e., at least k choices of a stress, up to scalar multiple, or at most $-k$ conditions on the positions of the joints for a stress to exist (if $k < 0$) [4]. Up until now such “meta-theorems” remained intuitive guidelines rather than precise theorems. One of our major goals in this paper is to give precision of form and of proof to these statements in the cases $k = -1$ and $k = 0$. Other cases will be investigated in later papers.

Because of the weakness of our geometric traditions, as well as the expectation that two approaches are better than one, it is helpful to develop the middle ground between synthetic geometry and the arithmetic algorithms—the algebraic geometry and geometric algebra of frameworks.

As suggested by the projective invariance of static stresses and infinitesimal motions [12, Thm. 5.10], the algebraic language which works best is the language of projective geometric invariant theory—the language of brackets (§ 2). For some of the more geometric discussions we extend this language to the Grassmann algebra or Cayley algebra, languages including brackets along with algebraic operations roughly corresponding to the geometric intersection and join.

* Received by the editors February 19, 1982.

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

[‡] Department of Mathematics, Champlain Regional College, St. Lambert, Quebec, Canada.

The methods we use are primarily algebraic, and invariant theoretic, but the questions and motivations are geometric—based on problems arising with frameworks. In addition to the pleasures of a clean, more easily communicated foundation for some standard practices in the study of frameworks, our motivation in undertaking this study was an unsolved problem in the theory of tensegrity frameworks. For a variety of reasons, in the theory of both the infinitesimal and the finite rigidity of frameworks with cables it is important to know how the form (and signs) of the static stress changes in the frameworks as the positions of the vertices are varied continuously through space [3], [12]. This prior preoccupation may explain some of the topics studied in §§ 5 and 6.

These geometric questions, and the algebra which results, are not unique to the study of engineering frameworks. The same geometry (and algebra) has arisen in the fields of scene analysis and of satellite geodesy. In scene analysis, the basic problem is to recognize the correct projections of 3-dimensional polyhedral objects—a task which is equivalent to the detection of stresses in planar frameworks [13], [14], [21]. In satellite geodesy, the basic problem is to take certain earth-satellite measurements and then calculate all the additional distances in the configuration—a calculation which breaks down if this configuration, formed as a framework with bars for the measured lengths, has even an infinitesimal motion [1]. The analysis given here also extends and clarifies the current mathematical analysis of critical configurations in geodesy [15].

As we have pursued the basic questions arising in these fields into algebraic geometry over the reals, we have encountered a steadily growing array of interesting problems. Thus there is much more work to be done.

1. Preliminaries on frameworks. Our work is motivated by the study of one essential property of bar-and-joint frameworks. This property can be described in two equivalent ways—as infinitesimal rigidity (the absence of velocities assigned to the joints which infinitesimally deform the structure) or as static rigidity (the ability of the framework to absorb all suitable external forces). The essential information for both concepts is condensed in a single rigidity matrix for the framework. However, to study the algebraic geometry of this matrix, we must step back to a more abstract level of the underlying graph and related polynomial domains.

A *graph* G is a finite set $V = \{a, b, \dots, f\}$ of *vertices* together with a collection E of two element subsets of V called *edges*.

A *bar-and-joint* framework in dimension n is a *coordinatization* of a graph G by a function $\alpha : a \rightarrow (a_1, \dots, a_n, 1)$ for every $a \in V$, where a_1, \dots, a_n are elements of a polynomial domain $R = k[x_1, \dots, x_r]$. (For most applications $k = \mathbb{R}$.) In a coordinatization the edges are called *bars* and the points $\alpha(a)$ are called *joints* of the frameworks.

The coordinates $(a_1, a_2, \dots, a_n, 1)$ may be regarded as a vector in the vector space R^{n+1} , or as special homogeneous coordinates in $PG(R, n)$ a projective space of dimension n , and we will frequently alternate between these points of view. It is no problem that we employ vector spaces and projective spaces over an integral domain R instead of a field; the process is essentially equivalent to working over the field of fractions of R but we use the integral domain to allow nontrivial homomorphisms of R . While most authors use simple Euclidean coordinates for the joints of a framework, the underlying geometry is projective [21]. Thus projective coordinates are essential to the algebraic geometry we will study.

A *real framework* or a *realization* of a graph G in $\dim n$ is a coordinatization of G with $R = \mathbb{R}$.

The rigidity matrix $M(G(\alpha))$ of a framework $G(\alpha)$ in dimension n is an $|E| \times n|V|$ matrix:

$$\begin{array}{l} \text{edge } \{a, b\} \\ \text{edge } \{a, f\} \\ \dots \\ \text{edge } \{e, f\} \end{array} \begin{array}{ccc} \text{vertex } a & \text{vertex } b & \text{vertex } f \\ \left[\begin{array}{ccc} a_1 - b_1, \dots, a_b - b_n & b_1 - a_1, \dots, b_n - a_n & \dots & 0, \dots, 0 \\ a_1 - f_1, \dots, a_n - f_n & 0, \dots, 0 & \dots & f_1 - a_1, \dots, f_n - a_n \\ \dots & \dots & \dots & \dots \\ 0, \dots, 0 & 0, \dots, 0 & \dots & f_1 - e_1, \dots, f_n - e_n \end{array} \right] \end{array}.$$

Thus for edge $\{d, f\}$ the matrix has the row with $d_1 - f_1, \dots, d_n - f_n$ in the columns of $d, f_1 - d_1, \dots, f_n - d_n$ in the columns of f and 0 elsewhere.

In the vocabulary of infinitesimal kinematics a solution to the homogeneous system of equations $M(G(\alpha))X = 0$ is called an infinitesimal or *instantaneous motion*. Such a motion is viewed as an n -vector for each joint $(m(a), m(b), \dots, m(f))$ where the equation for bar $\{a, b\}$ becomes

$$(m(a) - m(b)) \cdot (a_1 - b_1, \dots, a_n - b_n) = 0$$

a record of the condition that the velocities preserve the length $(a_1 - b_1)^2 + \dots + (a_n - b_n)^2 = \text{constant}$. This system of equations always has a nontrivial solution space since the rigid motions of space (rotations, translations and their combinations) always provide the *trivial motions*. A framework is *infinitesimally rigid* if the space of instantaneous motions is exactly the space of trivial motions of the joints. If the joints of the framework span at least an affine space of dimension $n - 1$ (a *full* framework), then the trivial motions form a space of dimension $\binom{n+1}{2}$. For such full frameworks infinitesimal rigidity is equivalent to the statement that

$$\text{rank}(M(G(\alpha))) = n|V| - \binom{n+1}{2}.$$

For example, in dimension 2 we need

$$\text{rank}(M) = 2|V| - 3.$$

Thus a triangle ($V = 2, E = 3$) is infinitesimally rigid provided the rows of the matrix are independent—a requirement which translates to the geometric statement that the triangle is not collinear. If the triangle is collinear then a nontrivial instantaneous motion exists, with a zero velocity at two of the joints, while the third joint has a velocity orthogonal to the line of the triangle.

In dimension 3, the condition for infinitesimal rigidity is $\text{rank}(M) = 3|V| - 6$. For example, a tetrahedron ($V = 4, E = 6$) is infinitesimally rigid if and only if the rows of the matrix are independent—or equivalently if and only if the joints are not coplanar. By a similar count, any triangulated sphere has $E = 3|V| - 6$ and will be infinitesimally rigid if and only if the rows of the rigidity matrix are independent [23].

In the vocabulary of statics, we directly investigate the row space of the rigidity matrix. We write F_{ab} for the row corresponding to a bar $\{a, b\} \in E$, or F_{cd} for the corresponding vector for any pair of joints $\{c, d\}$ (even if $\{c, d\}$ is not a bar). These latter vectors are read as special *static loads*—forces or n -vectors assigned to the joints of the framework. If we define the *equilibrium loads* on a framework as the space of vectors orthogonal to the rigid or trivial motions, then a framework is *statically rigid* if and only if the row space of the rigidity matrix (the space of the $F_{ab}, \{a, b\} \in E$) coincides with the space of equilibrium loads. An equivalent characterization [23,

Thm. 2] says that: a framework is statically rigid in dimension n if and only if either there are $|V| \leq n$ joints which span an affine space of $\dim |V| - 1$ and the framework coordinatizes a complete graph or there are $|V| > n$ joints which affinely span the n -space and all loads F_{cd} lie in the row space of the rigidity matrix.

As noted above, the trivial motions of a full framework form a space of $\dim \binom{n+1}{2}$ so the equilibrium loads form a subspace of dimension $n|V| - \binom{n+1}{2}$. Static rigidity, for a full framework, is equivalent to the statement that $\text{rank}(M) = n|V| - \binom{n+1}{2}$. Static and infinitesimal rigidity are clearly equivalent for full frameworks, and this equivalence also holds for smaller frameworks [12, Thm. 4.3].

Still within the language of statics, a linear dependence of a set of rows is called a stress.

$$\sum \lambda_{ab} F_{ab} = 0 \quad (\text{sum over bars}).$$

These scalars give a set of tensions ($\lambda_{ab} < 0$) and compressions ($\lambda_{ab} > 0$) in the bars, and the equations, rewritten for each joint, describe a static equilibrium of the corresponding forces,

$$\sum \lambda_{ab}(a - b) = 0 \quad (\text{fixed } a, \text{ sum over } \{a, b\} \in E).$$

A minimal statically rigid framework on a set of joints—a statically rigid framework with no static stresses—is called *isostatic*. For an isostatic framework the rows of the rigidity matrix form a basis for the equilibrium loads on the joints, and these frameworks are the basic objects of study in the next three chapters.

Given a framework $F = (G, \alpha)$, the *coordinatization matrix* A has a row $\alpha(a)$ for each joint $a \in V$,

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n & 1 \\ b_1 & b_2 & \cdots & b_n & 1 \\ \cdots & & & & \\ e_1 & e_2 & \cdots & e_n & 1 \end{bmatrix}.$$

If the entries $\{a_1, a_2, \dots, e_n\}$ are distinct algebraically independent elements of R (in which case we simply regard them to be distinct indeterminates over k), the framework F is a *generic coordinatization* of the graph G . If this generic coordinatization is an isostatic framework, we say that the graph G is *generically isostatic in dimension n* .

The small or flat frameworks (ones for which the joints do not even span an $(n - 1)$ -dimensional affine space) are well understood, as mentioned previously. However such flat frameworks would constantly clutter up the rest of our algebra in this paper, so we will *assume for the rest of the paper that the frameworks have $|V| \geq n$* which implies, in the generic case, that the framework is full.

A full framework is isostatic if and only if it has $E = n|V| - \binom{n+1}{2}$ bars, and there is no static stress (the bars are independent). The traditional way to check that the rows of a matrix form an appropriate basis is by taking determinants, but this is easier when the matrix is square. Our first task is to extend the rigidity matrix to a square matrix by adding $\binom{n+1}{2}$ independent rows—called a *tie-down*—so that the framework is isostatic if and only if this extended matrix has determinant $\neq 0$. There are many possible arrangements for the tie-down—but our objective is to introduce these rows in a natural format, as additional bars, and to later factor this extension out of the algebra (§ 3).

A *tie-down* of a framework $G(\alpha)$ in dimension n is a set of $\binom{n+1}{2}$ tie-down bars $\{a, x\}$, $a \in V$, $x \notin V$, where x has $m(x) = 0$ for any infinitesimal motion and adds the

row (nonzero only in the columns of a)

$$(a_1 - x_1, \dots, a_n - x_n, 0, \dots, 0)$$

to the extended rigidity matrix $M(G(\alpha), T)$.

We anticipate that, for an isostatic framework in dimension n , some correctly chosen set of $\binom{n+1}{2}$ tie-down bars will give an invertible matrix $M(G(\alpha), T)$. In the vocabulary of kinematics, these bars must block the trivial motions or, in the language of statics, they must generate the nonequilibrium loads. We begin with a simple proof that such tie-downs exist.

PROPOSITION 1.1. *A full framework in dimension n is isostatic if and only if there is a tie-down T of $\binom{n+1}{2}$ bars which produces an invertible rigidity matrix.*

Proof. Assume the tied-down framework has an invertible matrix—and hence no nonzero motions (solutions to the homogeneous system). Removing the $\binom{n+1}{2}$ tie-downs will introduce a space of infinitesimal motions of dimension $\binom{n+1}{2}$. Since this removal also introduces the rigid motions (a space of dimension $\binom{n+1}{2}$), there are no additional infinitesimal motions and the smaller framework is isostatic.

Conversely, assume that the full framework is isostatic in n -space. The rows of the rigidity matrix form an independent set of $n|V| - \binom{n+1}{2}$ vectors in the vector space $R^{n|V|}$. We can extend this independent set with $\binom{n+1}{2}$ vectors from the standard basis $(1, 0, 0, \dots, 0), \dots, (0, \dots, 0, 1)$ to form a basis for the entire space and an invertible matrix.

For each standard vector chosen we define a tie-down as follows: if the standard vector has 1 in the column for b_i , then the tie-down bar is $\{b, y\}$ and $\alpha(y) = (b_1, \dots, b_i - 1, \dots, b_n, 1)$. This choice gives the desired standard vector as a row of the extended rigidity matrix and thus is a “correct” tie-down. Q.E.D.

For any two isostatic frameworks on the same joints, the row spaces are the same, and thus the correct tie-downs will be the same. There is a geometric characterization of the correct tie-downs: when $\binom{n+1}{2}$ bars connect two infinitesimally rigid objects in n space (e.g., the framework and the ground), then the new unit is infinitesimally rigid if and only if the lines of the bars are independent as line segments of projective n -space (or equivalently, the Plucker coordinates for the lines are linearly dependent) [9, p. 659] or [21, Thm. 5.1, Corollary 5.3]. We will build this observation into the following proposition about the form of a static stress (row dependence) in the extended matrix.

PROPOSITION 1.2. *Given a framework $G(\alpha)$ in dimension n , with $|E| = n|V| - \binom{n+1}{2}$ bars, and tie-down T of $\binom{n+1}{2}$ bars, then the extended rigidity matrix $M(G(\alpha), T)$ has a row dependence if and only if either the tie-down bars lie on dependent lines on the projective space or there is a row dependence omitting the tie-down bars (a nontrivial stress on the original framework).*

Proof. Assume there is a nonzero motion in the tied-down framework (i.e., a row dependence in the square matrix). Either this motion is a rigid motion of the framework (excluding the ground) or the original framework is not infinitesimally rigid.

In the first case the tie-down bars did not block all rigid motions and this remaining rigid motion requires the dependence of the tie-down bars in projective space.

In the second case the framework has more than an $\binom{n+1}{2}$ -dimensional space of infinitesimal motions and the lower rank for the rigidity matrix without tie-downs gives the desired row dependence.

Conversely, if we assume a row dependence omitting the tie-downs, then $M(G, T)$ has a row dependence. If we assume the tie-down bars are dependent line segments in projective space, then the original framework has a rigid motion relative to the

ground which does not alter any of the tie-down bars (instantaneously). The square matrix $M(G, T)$ has a nonzero determinant—and a row dependence. Q.E.D.

Remark. In § 3, we will give a new combinatorial characterization of correct generic tie-downs. Proposition 3.5 gives the details about these combinatorially good arrangements.

2. The bracket ring and Cayley algebra. While we know that a framework in dimension n , with $|E| = n|V| - \binom{n+1}{2}$ bars, is infinitesimally rigid if and only if for some tie-down T , $\det(M(G(\alpha), T)) \neq 0$, we also recognize that this rigidity was determined by the rigidity matrix $M(G(\alpha))$. Our essential problem is to extract from $\det(M(G(\alpha), T)) \neq 0$, for some T , an algebraic condition which is independent of T (§ 3).

First, however, we must introduce the language of brackets, the classical language of projective geometric invariants, which is the most suitable for efficient expression and manipulation of $\det(M(G(\alpha), T))$. This language has been employed in the projective theory of frameworks [18], [21] and has reappeared in several nonprojective studies of the rigidity matrix [11], [15].

For example, in [11], Rosenberg gives a direct combinatorial decomposition of $\det(M(G, T))$ in the case $n = 2$. When generalized to n dimensions the basic units of his formulae are the *brackets* $[a, b, \dots, d]$ which represent the volume of the n simplex with $n + 1$ vertices a, b, \dots, d , a volume which is equivalent to an $(n + 1) \times (n + 1)$ determinant using the affine coordinates of the points as rows of a square matrix (i.e., an $(n + 1) \times (n + 1)$ subdeterminant of A). In Rosenberg's expansion the determinant of $M(G(\alpha), T)$ is the sum of products of such brackets in the joints. In each product of the sum, a joint occurs exactly (the valence of the vertex) $+ 1 - n$ times, so the expansions belong to the language of brackets, homogeneous in occurrences of symbols for the joints.

If a, b, \dots, d are $n + 1$ joints in V , the element of R obtained as the determinant of the corresponding $n + 1$ rows of A is a *bracket* $[a, b, \dots, d]$. The brackets satisfy the following well-known relations, called *syzygies*.

- 1) $[x_0, x_1, \dots, x_n] = 0$ if $x_i = x_j$ for some i, j with $i \neq j$, or if x_0, x_1, \dots, x_n are affinely dependent.
- 2) $[x_0, x_1, \dots, x_n] = \text{sign}(\sigma)[x_{\sigma 0}, x_{\sigma 1}, \dots, x_{\sigma n}]$ for any permutation σ of $0, 1, \dots, n$.
- 3) $[x_0, x_1, \dots, x_n][y_0, y_1, \dots, y_n] = \sum_{i=0}^n [y_i, x_1, \dots, x_n] \times [y_0, y_1, \dots, y_{i-1}, x_0, y_{i+1}, \dots, y_n]$.

Let B be the subring of R generated by all $(n + 1) \times (n + 1)$ determinants of A . B is called the *bracket ring* on V . If A is a generic coordinatization, then B is isomorphic to the bracket ring of the uniform matroid of rank $n + 1$ on V , as defined in [17], according to Hodge and Pedoe [10, p. 315, Thm. 1].

The commutative ring B is clearly an integral domain, since it is a subring of the integral domain R . We now wish to show that the generic bracket ring B has certain unique factorization properties. We first need the following result on factorization of invariants in R . Let A be generic. An element $f(a_1, \dots, e_n)$ of R is called an *invariant* if there exists an integer $s \geq 0$ such that for each nonsingular linear transformation S of the row space of A to itself, if $S(x)$ denotes the image of x , normalized by a scalar multiple so that $S(x)_{n+1} = 1$, then $f(S(a)_1, S(a)_2, \dots, S(e)_n) = (\det S)^s f(a_1, a_2, \dots, e_n)$. The integer s is the *degree* of the invariant f . Now the invariants in R are precisely the elements of B which are homogeneous in total degree, by the first fundamental theorem of invariant theory [7, Thm. 1].

Remark. Although we are working with a generic coordinatization of G , any coordinatization of G may be realized as a specialization of the generic one, by assigning values to the indeterminates. The question which will concern us in the following sections is, for a given graph G , which specializations of the generic coordinatization induce a stress or stresses in the framework.

We now adjoin generic vertices z_1, z_2, \dots, z_{n+1} distinct from a, b, \dots, e , letting $V' = V \cup \{z_1, z_2, \dots, z_{n+1}\}$. Letting A' be the matrix for V' analogous to A , R' the polynomial ring and B' the bracket ring, we note that R and B are subrings of R' and B' .

THEOREM 2.1. *Let f be an invariant element of R , where A is generic. Then any polynomial which is a factor of f in R is also invariant.*

Proof. We define a linear transformation S on the row space of A' by

$$S(a) = ([a, z_2, \dots, z_{n+1}], [z_1, a, z_3, \dots, z_{n+1}], \dots, [z_1, z_2, \dots, z_n, a]).$$

We note that $\det S = [z_1, \dots, z_{n+1}]^n$, since $\det(S(z_1), S(z_2), \dots, S(z_{n+1})) = [z_1, \dots, z_{n+1}]^{n+1}$. Now let f be an invariant element of R , and suppose that f factors in R as

$$f(a, b, \dots, e) = \prod_{j=1}^l g_j(a, b, \dots, e)^{r_j},$$

where the $g_j(a, b, \dots, e)$ are irreducible in R . By the definition of invariant, f is also invariant in R' and

$$f(Sa, Sb, \dots, Se) = [z_1, \dots, z_{n+1}]^{ns} f(a, b, \dots, e)$$

or

$$(*) \quad \prod_{j=1}^l g_j(Sa, Sb, \dots, Se)^{r_j} = [z_1, \dots, z_{n+1}]^{ns} \prod_{j=1}^l g_j(a, b, \dots, e)^{r_j}.$$

Now, $g_j(Sa, Sb, \dots, Se)$ is a polynomial in the coordinates of Sa, Sb, \dots, Se , but each such coordinate is a bracket, by our choice of S . Thus $g_j(Sa, Sb, \dots, Se)$ is a polynomial in B' , hence, is invariant. Furthermore, $[z_1, z_2, \dots, z_{n+1}]$ is an irreducible polynomial in R' , as is well known (see [6, Lemma A], where the same argument works even though the last row of our matrix consists of 1's). Thus from (*) we see

$$g_j(Sa, Sb, \dots, Se) = [z_1, z_2, \dots, z_{n+1}]^k h_j(a, b, \dots, e),$$

where $h_j(a, b, \dots, e)$ has no occurrences of any of the coordinates z_{ij} , $1 \leq i \leq n$, $1 \leq j \leq n+1$, that is, $h_j(a, b, \dots, e) \in R$.

Now $h_j(a, b, \dots, e) = g_j(Sa, Sb, \dots, Se) / [z_1, z_2, \dots, z_{n+1}]^k$ is a nonconstant invariant, as may be seen by applying an arbitrary linear transformation S' , hence h_j is in B . Now

$$f(a, b, \dots, e) = \prod_{j=1}^l g_j(a, b, \dots, e)^{r_j} = \prod_{j=1}^l h_j(a, b, \dots, e)^{r_j}$$

provides two factorizations of f , each involving the same number of nontrivial factors, with the $g_j(a, b, \dots, e)$ irreducible in R . But R is a unique factorization domain, hence for every j , $g_j(a, b, \dots, e) = \alpha_j h_j(a, b, \dots, e)$ for some i and some scalar α_j in k . Thus each irreducible factor of f is invariant. **Q.E.D.**

COROLLARY 2.2. *B is an integral domain in which each homogeneous element has a unique factorization into irreducible elements, and furthermore, the irreducible elements involved are homogeneous.*

Another algebraic structure to which we will frequently refer is the Cayley algebra. We will give here a brief and very informal introduction to this algebra, referring the reader to [7] for details.

We begin with a vector space U , which for our purpose we will take to be the row space of the matrix A considered earlier. The Cayley algebra is an extension of U with the usual operations of addition and scalar multiplication and two additional operations, join and meet, denoted \vee and \wedge . If u, v, \dots, w are m vectors, $m \leq n + 1$, then $u \vee v \vee \dots \vee w$, also denoted $uv \dots w$, is called an *extensor of step m* . Computationally, $uv \dots w$ may be identified with the vector of Plücker coordinates of the subspace $\text{span}(u, v, \dots, w)$, that is, the sequence of $m \times m$ minors of the $m \times (n + 1)$ matrix whose rows are u, v, \dots, w . The m -dimensional subspace $\text{span}(u, v, \dots, w)$ is also called the *support* of the extensor $uv \dots w$, assuming that u, v, \dots, w are linearly independent. An extensor $uv \dots w$ of step $n + 1$ is denoted as a bracket $[u, v, \dots, w]$, and may be identified with the brackets discussed previously. The meet of an extensor E of step m with an extensor E' of step l is an extensor of step $m + l - n - 1$, provided $n + 1 \leq m + l$, and its support is the intersection of the supports of E and E' , provided the union of those supports spans U . Thus the join and meet in the Cayley algebra correspond to the lattice operations on subspaces of U , provided the subspaces are independent in the case of join or are sufficiently large in the case of meet.

The condition in Proposition 1.2 that the tie-down bars $\{a, x\}, \{b, y\}, \dots, \{c, z\}$ be on dependent lines in projective space may now be restated as the condition that the 2-extensors ax, by, \dots, cz are linearly dependent in the Cayley algebra.

We will denote by $U^{(m)}$ the subspace of the Cayley algebra spanned by all extensors of step m from U . It also makes sense to use a Cayley algebra over a projective space $PG(R, n)$, by regarding it as the Cayley algebra over the corresponding vector space of dimension $n + 1$ over R . Again, $PG(R, n)^{(m)}$ denotes the space of step- m extensors in this case.

3. The pure condition for a stress in an isostatic framework. It follows immediately from the discussion in § 1 that for a generically isostatic graph G with an independent set T of $\binom{n+1}{2}$ tie-down bars specified, the condition for the existence of a stress in a specialization α of the generic coordinatization is that the $n|V| \times n|V|$ rigidity matrix $M(G(\alpha), T)$ has determinant equal to zero.

Let T be a tie-down consisting of ax, by, \dots, cz , where $a, b, \dots, c \in V$ and are not necessarily distinct and $x, y, \dots, z \notin V$ are distinct. Let $x_1, x_2, \dots, x_n, y_1, \dots, z_n$, the coordinates of x, y, \dots, z , be distinct indeterminants not involved in the coordinatization of V . Then we say that T is a *generic tie-down*.

LEMMA 3.1. *If G is generically isostatic with a set T of $\binom{n+1}{2}$ independent tie-down bars, then the determinant of the rigidity matrix $M(G, T)$ equals an element $C(G, T)$ of the bracket ring B on the set of vertices of $G \cup T$.*

Proof. Assume that G is given a generic coordinatization and that T is also generic. Since the entries of $M(G, T)$ are linear combinations of the coefficients of the vertices, its determinant is an element of $R = k[a_1, a_2, \dots, a_n, b_1, \dots, e_n, w_1, w_2, \dots, w_n, x_1, \dots, y_n]$, where a, b, \dots, e are the vertices of G and w, x, \dots, y the vertices of T which are not vertices of G . We wish to show that $\det M$ is an invariant of degree v , where $v = |V|$, by induction on v .

Let N be an arbitrary $n \times n$ minor in the first n columns of $M(G, T)$. If $\det N \neq 0$, each row of N corresponds to a bar incident to a . Let us denote the bars involved as

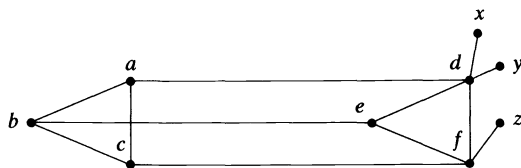
ap, aq, \dots, ar , which may be either bars of $G(\alpha)$ or tie-down bars. Then

$$\begin{aligned} \det N = \det \begin{bmatrix} a_1 - p_1 & \cdots & a_n - p_n \\ a_1 - q_1 & \cdots & a_n - q_n \\ \cdots & & \\ a_1 - r_1 & \cdots & a_n - r_n \end{bmatrix} &= \det \begin{bmatrix} a_1 - p_1 & \cdots & a_n - p_n & 0 \\ a_1 - q_1 & \cdots & a_n - q_n & 0 \\ \cdots & & & \\ a_1 - r_1 & \cdots & a_n - r_n & 0 \\ a_1 & \cdots & a_n & 1 \end{bmatrix} \\ &= \pm \det \begin{bmatrix} p_1 & \cdots & p_n & 1 \\ q_1 & \cdots & q_n & 1 \\ \cdots & & & \\ r_1 & \cdots & r_n & 1 \\ a_1 & \cdots & a_n & 1 \end{bmatrix} = \pm [p, q, \dots, r, a]. \end{aligned}$$

Let S be an arbitrary nonsingular transformation applied to $a, b, \dots, e, w, x, \dots, y$. If we now compute the determinant of $SM(G, T)$ by a Laplace expansion by the first n columns, each term in the expansion is an $n \times n$ minor times an $n(v-1) \times n(v-1)$ minor, with S applied to each entry of both minors. Thus, compared to the corresponding term in the expansion of $\det M(G, T)$ the $n \times n$ minor has been multiplied by $\det S$, since such a minor is a bracket by the preceding paragraph. By the induction hypothesis, the $n(v-1) \times n(v-1)$ minor is multiplied by $(\det S)^{v-1}$. Thus $C(G, T) = \det M(G, T)$ is an invariant of degree v and hence an element of B . The lemma now follows by specializing the generic coordinatization and the tie-down. Q.E.D.

The remainder of this section is devoted to showing that the bracket condition $C(G, T)$ factors as $C(G, T) = C(G)C(T)$ in B , where $C(G)$ is independent of the choice of the tie-down bars T and $\alpha C(G)$ is zero in a given specialization α of the coordinatization of V if and only if $G(\alpha)$ has a stress. We call $C(G)$ the *pure condition* for G . We prove these facts via a series of lemmas, after illustrating with an example.

Example.



Let G be the generically isostatic graph shown with 6 vertices and 9 edges in the plane ($n = 2$). Let us adjoin tie-down bars dx, dy, fz . We have chosen a tie-down that will give a particularly simple form to $C(T)$ (see Lemma 4.1). By Rosenberg’s method [11] we may compute

$$C(G, T) = \pm [dxy][dfz][abc][def]([adb][ecf] - [ade][bcf]).$$

The sign of $C(G, T)$ depends upon the order in which we list the bars of $G \cup T$ to index the rows of $M(G, T)$. We note parenthetically that computation of $C(G, T)$ can lead to other bracket expressions which are equivalent to the given one via the syzygies. We also note that this is a polynomial of degree 12 in 18 variables, thus we avoid expanded notation whenever possible.

Now we note that this tied-down framework has a trivial motion whenever dxy or dfz is collinear. Thus the irreducible polynomials $[dxy]$ and $[dfz]$ should be factors of $C(T)$, and indeed it can be shown that $C(T) = [dxy][dfz]$.

Thus $C(G) = \pm[abc][def]([adb][ecf] - [ade][bcf])$. From this we can recognize that $G(\alpha)$ has a stress in any coordinatization α in which abc is collinear or def is collinear. The third factor corresponds to the Cayley algebra expression $ad \wedge be \wedge cf$ and is 0 whenever the lines ad, be and cf are concurrent or parallel. We will consider factoring of the pure condition and the corresponding geometric meaning in § 4.

LEMMA 3.2. *Let T be a generic tie-down of $\binom{n+1}{2}$ bars. Then the dependence of the step-two extensors ax, by, \dots , and cz is a bracket condition $C(T) = 0$, where $C(T)$ is a factor of $C(G, T)$.*

Proof. Each of the step-two extensors ax, by, \dots, cz may be represented by its vector of Plücker coordinates, that is, by the sequence of 2×2 minors of the $(n + 1) \times 2$ matrix with columns a and x in the case of ax , etc. Each vector of Plücker coordinates is of length $\binom{n+1}{2}$, but we have exactly $|T| = \binom{n+1}{2}$ of them, hence we have a square matrix N , and the dependence of the step-two extensors in $PG(R, n)^{(2)}$ is equivalent to $\det N = 0$.

The entries of N are all polynomials in the coordinates of $G \cup T$. It can be shown directly by applying elementary linear transformations that $\det N$ is an invariant, hence an element of B .

Let $C(T) = \det N$. Let \bar{k} be the algebraic closure of k , and let us temporarily work in $\bar{k}[a_1, a_2, \dots, z_n]$. We know from Proposition 1.2 that whenever the step-two extensors ax, by, \dots, cz are dependent in $PG(R, n)^{(2)}$ then $M(G, T)$ has dependent rows. Thus $\alpha C(G, T) = 0$ for any specialization α for which $\alpha C(T) = 0$. Then, by Hilbert's Nullstellensatz, $C(T) | (C(G, T))^r$ for some integer r . But $C(T)$ is at most linear in each coefficient of the vectors x, y, \dots, z , hence $C(T)$ has no multiple factors, hence $C(T) | C(G, T)$.

Since $C(T)$ and $C(G, T)$ both have integer coefficients, we must also have $C(T) | C(G, T)$ in the polynomial ring $k[a_1, a_2, \dots, z_n]$. Q.E.D.

We must now characterize the independent generic tie-downs. This is done in Proposition 3.5, after some preliminary lemmas.

LEMMA 3.3. *Let w_1, \dots, w_k be distinct vectors in a vector space or projective space W of dimension n over $R = k[x_1, \dots, x_r]$, where we assume a standard basis e_1, \dots, e_n of W has been fixed. If $\Phi: R \rightarrow R'$ is a k -algebra homomorphism, then we denote by $\check{\Phi}: W \rightarrow W'$ the map obtained by applying Φ coordinatewise. Then w_1, \dots, w_k is linearly independent if and only if there exists $\check{\Phi}$ such that $\check{\Phi}w_1, \dots, \check{\Phi}w_k$ is linearly independent with distinct elements.*

Proof. The "only if" statement is trivial. Now we suppose that $\check{\Phi}w_1, \dots, \check{\Phi}w_k$ is linearly independent with distinct elements, and we extend it to a basis $B' = \check{\Phi}w_1, \dots, \check{\Phi}w_k, e'_{k+1}, \dots, e'_n$ where (after reindexing) e'_i is from the standard basis of W' . If we reindex e_1, \dots, e_n similarly so that $e'_i = \check{\Phi}e_i$, then $B = w_1, \dots, w_k, e_{k+1}, \dots, e_n$ is a preimage of B . But $\Phi(\det B) = \det B' \neq 0$, hence $\det B \neq 0$ and w_1, \dots, w_k is linearly independent. Q.E.D.

The maps Φ and $\check{\Phi}$ of the previous lemma are what we have previously referred to as "specialization" maps.

LEMMA 3.4. *Let v_1, \dots, v_{n+1} be any distinct linearly independent points of $PG(R, n)$. Then the $\binom{n+1}{2}$ lines $v_i v_j, i \neq j$, are a basis of $PG(R, n)^{(2)}$.*

We will now show that certain generic tie-downs are independent by applying a specialization map from the tie-down to the $\binom{n+1}{2}$ lines of Lemma 3.4. This specialization is no longer a tie-down in the usual sense since we are using only vertices of

the original framework, but this is an easy configuration to use as a standard specialization.

PROPOSITION 3.5. *Let $F = (G, \alpha)$ be a generic framework with $|V| = m \geq n$ and T a generic tie-down of $\binom{n+1}{2}$ bars. For $v_i \in V$, let α_i be the number of tie-down bars incident to v_i , and assume that we have reindexed so that $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_m$. Then $C(T) \neq 0$ if and only if*

$$(*) \quad \sum_{i=1}^k \alpha_i \leq nk - \binom{k}{2} \quad \text{for all } k, \quad 1 \leq k \leq n-1.$$

Proof. We know $C(T) \neq 0$ if and only if the step-two extensors of the bars in T are linearly independent in $PG(R, n)^{(2)}$. Let K be the subspace of $PG(R, n)$ spanned by v_1, v_2, \dots, v_k . Since P is generic, $\dim K = k-1$. It is well known (and easy to verify) that the subspace of $PG(R, n)^{(2)}$ consisting of all 2-extensors corresponding to lines which intersect K has dimension $n + (n-1) + (n-2) + \dots + (n-k+1) = nk - \binom{k}{2}$. Thus $\sum_{i=1}^k \alpha_i \leq nk - \binom{k}{2}$ for all j_1, j_2, \dots, j_k distinct, and in particular, $\sum_{i=1}^k \alpha_i \leq nk - \binom{k}{2}$. We note that $\dim PG(R, n)^{(2)} = \binom{n+1}{2}$ and $\sum_{i=1}^m \alpha_i = \binom{n+1}{2} = n^2 - \binom{n}{2}$.

Conversely, suppose that the α_i 's satisfy (*). To prove that the 2-extensors of T are linearly independent, it suffices to find a specialization in which they are linearly independent, by Lemma 3.3. Since the vertices in V as well as the tie-down vertices (i.e., vertices on the "ground") all have distinct indeterminants as coordinates, we may specialize so that any pair of vertices that we wish is identified, by specifying a homomorphism Φ that maps the coordinates of one to the coordinates of the other. First we will identify various vertices in V so that exactly $n+1$ distinct vertices remain. The fact that new stresses are thus introduced in F need not concern us, since we are presently concerned only with the independence of the tie-down bars. If we started with $|V| = n$, we add a dummy vertex to V , having no bars, so again $|V| = n+1$. We then identify each tie-down vertex with a vertex of V , hence identifying each tie-down bar with a line between two vertices in V . We will show that this can be done in such a way that distinct tie-down bars are identified with distinct such lines, and hence by Lemma 3.4, the specialized bars are independent, and we are done.

Now suppose that $m > n+1$. We identify v_{n+2} with v_{n+1} to form a new vertex with $\alpha_{n+1} + \alpha_{n+2}$ incident tie-down bars. We choose l such that $\alpha_{l-1} > \alpha_{n+1} + \alpha_{n+2} \geq \alpha_l$, $1 \leq l \leq n+1$. Let $\alpha'_i = \alpha_i$ if $i < l$, $\alpha'_i = \alpha_{i-1}$ if $l < i \leq n+1$, $\alpha'_i = \alpha_{i+1}$ if $n+2 \leq i \leq m-1$ and $\alpha'_i = \alpha_{n+1} + \alpha_{n+2}$. Then α'_i is the number of the tie-down bars incident to the i th vertex after the identification, correctly indexed so that $\alpha'_1 \geq \alpha'_2 \geq \dots \geq \alpha'_{m-1}$. If we show that the α'_i satisfy (*), then by induction we may assume that $m = n+1$ and that (*) is still satisfied. It suffices to consider k such that $l \leq k \leq n$.

Case 1. If $\sum_{i=1}^k \alpha_i < nk - \binom{k}{2} - \alpha_{n+2}$, we are done, for $\sum_{i=1}^k \alpha'_i = (\sum_{i=1}^{k-1} \alpha_i) + \alpha_{n+1} + \alpha_{n+2} \leq (\sum_{i=1}^k \alpha_i) + \alpha_{n+2} < nk - \binom{k}{2}$.

Case 2. $\sum_{i=1}^k \alpha_i \geq nk - \binom{k}{2} - \alpha_{n+2}$. Then $\sum_{i=k+1}^{n+2} \alpha_i = \sum_{i=1}^{n+2} \alpha_i - \sum_{i=1}^k \alpha_i \leq \binom{n+1}{2} - nk + \binom{k}{2} + \alpha_{n+2}$. Since the α_i are decreasing, $\alpha_{n+1} + \alpha_{n+2} \leq 2$ (average $\{\alpha_i\}_{i=k+1}^{n+2}$), hence

$$(n-k+2)(\alpha_{n+1} + \alpha_{n+2}) \leq 2 \left[\binom{n+1}{2} - nk + \binom{k}{2} + \alpha_{n+2} \right],$$

$$(n-k+1)(\alpha_{n+1} + \alpha_{n+2}) \leq (n-k+2)\alpha_{n+1} + (n-k)\alpha_{n+2} \leq 2 \left[\binom{n+1}{2} - nk + \binom{k}{2} \right]$$

$$= (n-k+1)(n-k),$$

so $\alpha_{n+1} + \alpha_{n+2} \leq n-k$. Now $\sum_{i=1}^k \alpha'_i = (\sum_{i=1}^{k-1} \alpha_i) + \alpha_{n+1} + \alpha_{n+2} \leq n(k-1) - \binom{k-1}{2} + n-k = nk - \binom{k}{2} - 1$. Thus we may assume that $m = n+1$.

Let us now form a bipartite graph on the set B consisting of copies of v_i for all i , $B = \{v_{11}, v_{12}, \dots, v_{1\alpha_1}, v_{21}, \dots, v_{n+1, \alpha_{n+1}}\}$, and the set E of pairs $v_i v_j$, $i < j$, by letting v_{ik} be adjacent to $v_i v_j$ and $v_j v_i$ for all i, j, k . We will now show by Hall's marriage theorem (see, for example, [8, Thm. 5.1.11]) that a complete matching of B into E exists, that is, that it is possible to assign to each v_{ik} an adjacent $v_i v_j$ in one-to-one fashion. We will then be done, for we may specialize the k th tie-down bar at v_i to the line $v_i v_j$, obtaining the independent specialization required.

If $U \subseteq B$, let $R(U) = \{e \in E : \text{for some } u \in U, e \text{ is adjacent to } u\}$. By Hall's marriage theorem, it suffices to show that $|U| \leq |R(U)|$ for all $U \subseteq B$. For $U \subseteq B$, let $I = \{i : 1 \leq i \leq n + 1, \text{ and for some } j, v_{ij} \in U\}$ and $U^* = \{v_{ik} : i \in I\}$. Then $|U| \leq |U^*| = \sum_{i \in I} \alpha_i \leq n|I| - \binom{|I|}{2} = |R(U^*)| = |R(U)|$, completing the proof. Q.E.D.

Let us say that the tie-down T is *saturated* at v_k if $\sum_{i=1}^k \alpha_i = nk - \binom{k}{2}$ and *unsaturated* if it is unsaturated at v_k for all k , $1 \leq k \leq n - 1$. In showing that we could collapse down to the case $m = n + 1$, we actually showed that if we started with an unsaturated tie-down, then it remains unsaturated after the collapse to $m = n + 1$. By a virtually identical proof we could have actually collapsed one step further to $m = n$. However, we then automatically get a tie-down which is saturated at v_n and which can easily be saturated at other v_k as well, even if the original tie-down was unsaturated. We did not need this further collapse for the remainder of the proof, but neither do we need the information about unsaturation. However, we will consider unsaturated tie-downs further in Proposition 3.12.

COROLLARY 3.6. *The dependence or independence of the bars of a generic tie-down of a generic framework is determined solely by the unordered list (with repetition) of the $\binom{n+1}{2}$ vertices to which the tie-down bars are incident.*

COROLLARY 3.7. *If $n = 2$, for the generic tie-down T and the generic framework $F = (G, \alpha)$, $C(T) = 0$ if and only if all three bars of T are incident to the same vertex of G . If $n = 3$, $C(T) = 0$ if and only if, of the six bars of T , at least 4 are incident to one vertex, or 3 are incident to each of two vertices.*

Remark. The situation is much more complicated if we take a nongeneric tie-down. For example, in dimension 3, four bars which determine lines on a common regulus are dependent. The indeterminate (generic) endpoints of our tie-down bars prevent any such special position from occurring. However, we may still state the following for nongeneric tie-downs.

COROLLARY 3.8. *Condition (*) of Proposition 3.5 is a necessary condition for an arbitrary tie-down T to satisfy $C(T) \neq 0$.*

LEMMA 3.9. *Let $I \subseteq \{1, 2, \dots, n\}$, $I \neq \emptyset$ and $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_{n+1}$ nonnegative integers such that $\sum_{i=1}^k \alpha_i \leq nk - \binom{k}{2}$ for all k , $1 \leq k \leq n + 1$. Let p be minimal such that $p \notin I$, $1 \leq p \leq n + 1$ and let q be maximal such that $q \in I$, $1 \leq q \leq n + 1$. If $\sum_{i \in I} \alpha_i = n|I| - \binom{|I|}{2}$, then $\alpha_p < \alpha_q$. Furthermore, $q = p - 1$ and $I = \{1, 2, \dots, q\}$.*

Proof.

$$\begin{aligned} \alpha_p &= \sum_{i \in I \cup \{p\}} \alpha_i - \sum_{i \in I} \alpha_i \leq \sum_{i=1}^{|I|+1} \alpha_i - \sum_{i \in I} \alpha_i \\ &\leq n(|I| + 1) - \binom{|I| + 1}{2} - n|I| + \binom{|I|}{2} = n - |I|, \end{aligned}$$

and

$$\begin{aligned} \alpha_q &= \sum_{i \in I} \alpha_i - \sum_{i \in I - \{q\}} \alpha_i \geq \sum_{i \in I} \alpha_i - \sum_{i=1}^{|I|-1} \alpha_i \\ &\geq n|I| - \binom{|I|}{2} - n(|I| - 1) + \binom{|I| - 1}{2} = n - |I| + 1 \end{aligned}$$

hence $\alpha_p < \alpha_q$, and the rest follows immediately. Q.E.D.

LEMMA 3.10. *Let T be a generic tie-down of $\binom{n+1}{2}$ bars of a generic framework F , with α_i tie-down bars incident to the vertex v_i , $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_m$, $m = |V| \geq n$. Let T^* be the tie-down obtained from T by specializing one tie-down bar $\{v_j, x\}$ at v_j to the line $v_j v_l$, for $j < l$. Let Φ denote the specialization map such that $\tilde{\Phi}: x \rightarrow v_l$, where $\tilde{\Phi}$ fixes all vertices other than x . Then $\Phi C(T) \neq 0$ if and only if $C(T) \neq 0$.*

Proof. If $\Phi C(T) \neq 0$ then $C(T) \neq 0$ by Lemma 3.3. If $C(T) \neq 0$, then $\sum_{i=1}^k \alpha_i \leq nk - \binom{k}{2}$ for all k . To prove that $\Phi C(T) \neq 0$, we proceed as in the proof of Lemma 3.5, until we form the bipartite graph. Since $\{v_j, x\}$ has already been assigned the line $v_j v_l$, we eliminate $\{v_j, x\}$ from T to obtain a tie-down with incidence numbers $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_j - 1 \geq \dots \geq \alpha_m$. We also eliminate v_{j1} , from B and $v_j v_l$ from E in our bipartite graph, letting $B' = B - \{v_{j1}\}$, $E' = E - \{v_j v_l\}$. Now if $U \subseteq B'$ and $R(U)$, I and U^* are defined as before but in terms of the new graph, then if $j \notin I$ and $l \notin I$, then $|U| < |R(U)|$ as before. If $j \in I$, then

$$|U| \leq |U^*| = \left(\sum_{i \in I} \alpha_i \right) - 1 \leq n|I| - \binom{|I|}{2} - 1 = |R(U)|.$$

Finally, if $l \in I$ and $j \notin I$, then by Lemma 3.9, since $j < l$, $\sum_{i \in I} \alpha_i < n|I| - \binom{|I|}{2}$. Thus

$$|U| \leq |U^*| = \sum_{i \in I} \alpha_i \leq n|I| - \binom{|I|}{2} - 1 = |R(U)|,$$

and we are done. Q.E.D.

We will call a tie-down T *nondegenerate* if T has $\binom{n+1}{2}$ bars and $C(T) \neq 0$.

Let us return now to the problem of factoring $C(G, T)$ for G generically isostatic. By Lemma 3.2, $C(T)$ is a factor, so we obtain a factorization $C(G, T) = C(T)C_T(G)$. It remains to be shown that $C_T(G)$ is independent of T and is therefore the pure condition $C(G)$ which we seek. Let $F = (G, \alpha)$ be generic, with $|V| \geq n$.

LEMMA 3.11. *Let T' be a generic tie-down of F obtained from the generic tie-down T by replacing a bar ax by dx , where ad is an edge of G . Assume T and T' are nondegenerate. Then $C_T(G) = C_{T'}(G)$.*

Proof. Let us first specialize x to x^* , a point in general position on the line ad (e.g., $x_i^* = \beta a_i + (1 - \beta)d_i$ for $i = 1, 2, \dots, n$, where β is an indeterminate). Let T^* and T'^* denote the sets of tie-down bars obtained from T and T' (resp.) by specializing x to x^* , and Φ the specialization map $\tilde{\Phi}: x \rightarrow x^*$. Since both $C(T) \neq 0$ and $C(T') \neq 0$, Lemma 3.10 implies that either $\Phi C(T) \neq 0$ or $\Phi C(T') \neq 0$, depending on which of a or d has more incident tie-down bars. But $\Phi C(T) \neq 0$ if and only if $\Phi C(T') \neq 0$, since T^* determines the same set of lines as T'^* . Thus $\Phi C(T) \neq 0$ and $\Phi C(T') \neq 0$.

Now let us examine the rows of the rigidity matrix for $G \cup T^*$ corresponding to the bars ad and ax^* . These rows have nonzero entries only in columns corresponding to the vertices a and d , namely:

$$\begin{array}{cc} & a & & d \\ ad & (a_1 - d_1, a_2 - d_2, \dots, a_n - d_n, & d_1 - a_1, d_2 - a_2, \dots, d_n - a_n) \\ ax^* & (a_1 - x_1^*, a - x_2^*, \dots, a_n - x_n^*, & 0, \quad 0, \dots, 0 \end{array}.$$

Since a , d and x^* are collinear, the scalar $1 - \beta$ satisfies $1 - \beta = (a_i - x_i^*) / (a_i - d_i)$ for $i = 1, 2, \dots, n$. If we subtract $1 - \beta$ times the row ad from the row ax^* , we have

$$\begin{array}{cc} ad & (a_1 - d_1, a_2 - d_2, \dots, a_n - d_n, & d_1 - a_1, d_2 - a_2, \dots, d_n - a_n) \\ ax^* & (0, \quad 0, \dots, \quad 0, & a_1 - x_1^*, a_2 - x_2^*, \dots, a_n - x_n^*) \end{array}.$$

If $\lambda = (1 - \beta)/(-\beta)$, $\lambda = (a_i - x_i^*)/(d_i - x_i^*)$ for $i = 1, 2, \dots, n$, so we now have

$$ad \begin{pmatrix} a_1 - d_1, & a_2 - d_2, & \dots, & a_n - d_n, & & d_1 - a_1, & d_2 - a_2, & \dots, & d_n - a_n \\ 0, & 0, & \dots, & 0, & \lambda(d_1 - x_1^*), & \lambda(d_2 - x_2^*), & \dots, & \lambda(d_n - x_n^*) \end{pmatrix}$$

and we have used only row operations which leave the determinant of the rigidity matrix unchanged. But these are the rows corresponding to the bars ad and dx^* , except for the factor λ in the rigidity matrix for $G \cup T^*$. Thus $\Phi C(G, T) = \lambda \Phi C(G, T')$. Now the vector of Plücker coordinates for ax^* is also λ times the vector for dx^* , as may be easily verified. Hence $\Phi C(T) = \lambda \Phi C(T')$.

It then follows, since $\Phi C(T) \neq 0$ and $\Phi C(T') \neq 0$, that $C_{T^*}(G) = C_{T'^*}(G)$. Thus $C_T(G)$ and $C_{T'}(G)$ are elements of the polynomial ring $k[a_1, \dots, z_n]$ whose images are equal under the specialization map $\Phi: x_i \rightarrow x_i^*$, $i = 1, 2, \dots, n$. But x_1, x_2, \dots, x_n do not appear in $C_T(G)$ or $C_{T'}(G)$, hence $C_T(G) = C_{T'}(G)$. Q.E.D.

PROPOSITION 3.12. *Let G be generically isostatic in dimension n , with $|V| = m \geq n$. Then any two nondegenerate generic tie-downs T' and T'' satisfy $C_{T'}(G) = C_{T''}(G)$.*

Proof. By Lemma 3.11, we may move a tie-down bar of a nondegenerate generic tie-down T from any vertex to an adjacent vertex, keeping $C_T(G)$ fixed, provided we begin and end such a move with a nondegenerate generic tie-down. We will call such a move an *edge move*, and we will show that we can transform T' to T'' by a sequence of edge moves.

We know that a generic tie-down T with incidence numbers $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_m$ is nondegenerate if and only if

$$\sum_{i=1}^k \alpha_i \leq kn - \binom{k}{2},$$

for all k , $1 \leq k \leq n - 1$, and T has $\binom{n+1}{2}$ bars. Thus T is unsaturated if and only if T is nondegenerate, and if a single tie-down bar is moved from any vertex to any other vertex, the resulting tie-down is also nondegenerate.

Now we show that if T is nondegenerate, then T may be made unsaturated by a sequence of edge moves. Since G is generically isostatic, G is connected. (In fact, it is not difficult to show that G must be n -connected.) Suppose that r is maximal so that $\alpha_1 = \alpha_2 = \dots = \alpha_r$. Then it is possible to find a path P from some $v \in \{v_1, v_2, \dots, v_r\}$ to v_n such that the path contains no other vertex besides v from $\{v_1, v_2, \dots, v_r\}$. We may reindex so that $v = v_r$. We now successively do an edge move along each edge of the path P . We must show that all the intermediate tie-downs are nondegenerate. But at each step along P , the net effect is to have moved a single tie-down bar from v_r to some vertex v_s , where $s > r$. But the sequence $\alpha_1, \alpha_2, \dots, \alpha_r - 1, \dots, \alpha_s + 1, \dots, \alpha_m$ satisfies (*) if $\alpha_1, \dots, \alpha_n$ does. Thus the intermediate tie-downs are nondegenerate, and by valid edge moves, we have transformed T to a tie-down T_1 with incidence numbers $\beta_1 \geq \dots \geq \beta_m$, where $\beta_r = \alpha_r - 1$, $\beta_t = \alpha_n + 1$ for some $t \leq n$ and $\beta_i = \alpha_i$ otherwise. But $t < n$ only if $\alpha_t = \alpha_{t+1} = \dots = \alpha_n$, whence by Lemma 3.9, $\sum_{i=1}^k \beta_i = \sum_{i=1}^k \alpha_i < nk - \binom{k}{2}$ for $t \leq k \leq n - 1$ and similarly for $1 \leq k \leq r - 1$. For $r \leq k \leq t - 1$, $\sum_{i=1}^k \beta_i < \sum_{i=1}^k \alpha_i$, and so T_1 is unsaturated.

Thus we may assume that T' and T'' are unsaturated. Let $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_m$ be the incidence numbers of T' and $\beta_1, \beta_2, \dots, \beta_m$ the incidence numbers of T'' . Thus the vertex v_i has α_i incident tie-down bars in T' and β_i in T'' . We may assume that if $\alpha_i = \alpha_{i+1}$, then $\beta_i \geq \beta_{i+1}$. We proceed now by induction on $\sum_{i=1}^m |\beta_i - \alpha_i| = q$. If $q = 0$, $\alpha_i = \beta_i$ for all i and T' is isomorphic to T'' and $C_{T'}(G) = C_{T''}(G)$. If $q > 0$, we consider two cases.

Case 1. There exist $j < l$ such that $\beta_j < \alpha_j > \alpha_l < \beta_l$. Then we choose any path in G from v_j to v_l and do a sequence of edge moves along that path for the tie-down T' . Since T' is unsaturated, at each step along the path the resulting tie-down is nondegenerate, and the result is a tie-down T'' with incidence numbers $\alpha_1, \dots, \alpha_j - 1, \dots, \alpha_l + 1, \dots, \alpha_m$, where we may have to reindex to keep decreasing order, say $\alpha'_1 \geq \dots \geq \alpha'_m$ but if so, we reindex the β_i in the same way.

Now we must check that T'' is also unsaturated. If $\alpha_j - 1 \geq \alpha_l + 1$, we cannot have increased the partial sums $\sum_{i=1}^k \alpha'_i$. If $\alpha_j - 1 < \alpha_l + 1$, then $\alpha_j = \alpha_l + 1$ and $\alpha_i = \alpha'_i$ for all i , although α_i and α'_i need not refer to the same vertex. Therefore the partial sums $\sum_{i=1}^k \alpha_i$ remain unchanged, and T'' is unsaturated since T' is. But now T'' and T'' satisfy $C_{T''}(G) = C_{T''}(G)$ by the induction hypothesis, hence $C_T(G) = C_{T''}(G)$.

Case 2. There exists p such that $\beta_i \geq \alpha_i$ for all $i < p$ and $\beta_i \leq \alpha_i$ for all $i \geq p$. Thus $\beta_i \geq \alpha_i \geq \alpha_j \geq \beta_j$ if $i < p \leq j$. Let us now reindex so that $\beta_1 \geq \beta_2 \geq \dots \geq \beta_m$ and reindex the α_i in the same way. Note that the original $\beta_1, \dots, \beta_{p-1}$ remain the first $p - 1$ entries in perhaps different order. Since $q > 0$, and $\sum_{i=1}^m \alpha_i = \sum_{i=1}^m \beta_i = \binom{n+1}{2}$, there exist j, l such that $\alpha_j < \beta_j$ and $\alpha_l > \beta_l$. But then $j < p$ and $l \geq p$ and since $\alpha_j \geq \alpha_l, \beta_j > \beta_l$. Thus $\alpha_j < \beta_j > \beta_l < \alpha_l$ and we apply the argument of Case 1 with α and β interchanged. Q.E.D.

THEOREM 3.13. *If G is a generically isostatic graph in dimension n with $|V| \geq n$, then there exists an element $C(G)$ of the bracket ring on the vertices of G such that for any specialization α of the generic coordinatization of G given by a specialization map Φ , $G(\alpha)$ has a stress if and only if $\Phi C(G) = 0$.*

Proof. Choose a nondegenerate generic tie-down T for $G(\alpha)$. Then by Proposition 1.2, $G(\alpha)$ has a stress if and only if $G(\alpha) \cup T$ has a stress, if and only if $\Phi C(G, T) = 0$, if and only if $\Phi C_T(G) = 0$, since $C(T) \neq 0$. Since $C_T(G)$ is independent of the choice of T , we take $C(G) = C_T(G)$. Q.E.D.

COROLLARY 3.14. *Let G be an isostatic framework with $|V| > 3$ and $n = 2$ or 3 . With the generic coordinatization on G , let T^* be an arbitrary tie-down. Then if $C(G, T^*)$ is the determinant of the rigidity matrix and $C(T^*)$ the bracket condition for the dependence of the tie-down bars (now specialized to T^* instead of a generic tie-down), we still have $C(G, T^*) = C(T^*)C(G)$.*

Proof. To the polynomial ring $k[a_1, a_2, \dots, a_n, \dots, z_n]$ apply the specialization map Φ (a ring homomorphism) taking $x_1, x_2, \dots, x_n, \dots, z_n$ to the coordinates of the corresponding endpoints of the bars in T^* . The equation $C(G, T) = C(T)C(G)$ for a generic tie-down T is preserved under the homomorphism but $C(G)$ is independent of $x_1, x_2, \dots, x_n, \dots, z_n$, hence the result. Q.E.D.

There are a number of unsolved problems regarding the pure condition. Can two distinct isostatic frameworks on the same set of vertices have identical pure conditions? Given a bracket expression, what frameworks have a pure condition with the given expression as a factor? Other problems relating to the factoring of the pure condition are discussed in the following sections.

4. Factoring pure conditions. We know that, for a graph with the correct count $|E| = n|V| - \binom{n+1}{2}$, the infinitesimal rigidity of the framework in n -space is equivalent to the pure condition being nonzero at that coordinatization. What do these pure conditions actually look like? Since there is no generally available collection of graphs and their pure conditions, we begin with two tables of examples—one for the plane and one for 3-space. Since the pure condition may change sign if the order of the edges is changed, all conditions are given up to a global sign.

These tables require some explanation and raise certain obvious questions: How does one determine these pure conditions? Do algebraic patterns, such as the factoring,

reflect underlying patterns in the graph? How do we know that the factors shown are irreducible?

The answers to all these questions are tied up together, since knowledge of the factoring is often used to determine the pure condition given in the tables. We will summarize the techniques we used under four headings. The first three subsections will derive pure conditions, using certain patterns in the graph and in the factoring of the conditions, while the fourth subsection outlines techniques to show that the given factors are irreducible. Along the way we will explain the conditions given in Tables 1 and 2.

TABLE 1
Plane frameworks

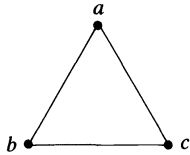
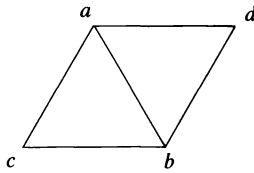
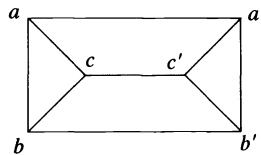
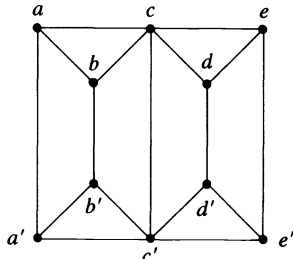
Name and graph	Pure condition and geometric condition
1.1. Triangle 	Points a, b, c , collinear $[abc]$
1.2. Two triangles 	One of the triangles abc or abd is collinear $[abc][abd]$
1.3. Triangular prism 	Triangles $abc, a'b'c'$ are perspective from a line \equiv Triangle abc or $a'b'c'$ is collinear or the two triangles are perspective from a point $[abc][a'b'c']([abb'][a'c'c'] - [a'bb'][ac'c'])$
1.4. Edge linked prisms 	Either one of the triangles is collinear or one of the triples aa', bb', cc' or cc', dd', ee' is concurrent $[abc][a'b'c'][cde][c'd'e']$ $\cdot ([abb'][a'c'c'] - [a'bb'][ac'c'])$ $\cdot ([cdd'][c'e'e'] - [c'dd'][ce'e'])$

TABLE 1 (cont.)

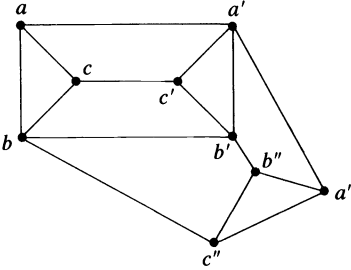
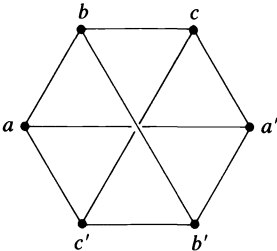
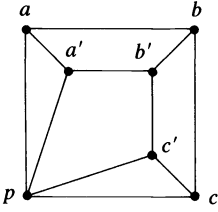
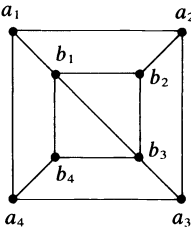
Name and graph	Pure condition and geometric condition
1.5. 3 vertex linked prisms	Either one of the triangles is collinear or one of the triples aa' , bb' , cc' or $a'a''$, $b'b''$, bc'' is concurrent
	$[abc][a'b'c'][a''b''c'']$ $\cdot ([abb'] [a'c'c] - [a'bb'] [ac'c])$ $\cdot ([a'b'b''] [a''c''b] - [a''b'b''] [a'c''b])$
1.6. $K_{3,3}$	The six joints lie on a plane conic
	$[abc][ab'c'] [a'b'c] [a'bc']$ $- [a'bc] [a'b'c'] [ab'c] [abc']$
1.7.	Either one of the triangles is collinear or the three points $ab \wedge a'b'$, $bc \wedge b'c'$ and p are collinear
	$[paa'] [pcc'] ([aba'] [bcb'] [b'c'p] - [abb'] [bcb'] [a'c'p])$ $+ [abb'] [bcc'] [a'b'p])$
1.8. Cube with 1 bar	Either one of the triangles is collinear or the three points $a_1b_1 \wedge a_2b_2$, $a_3b_3 \wedge a_4b_4$ and $a_1a_4 \wedge a_2a_3$ are collinear
	$[b_1b_2b_3][b_1b_3b_4]([a_1a_2b_1][a_2a_3b_2][a_3a_4b_3][a_4a_1b_4]$ $- [a_1a_2b_2][a_2a_3b_3][a_3a_4b_4][a_4a_1b_1])$

TABLE 2
Spatial frameworks

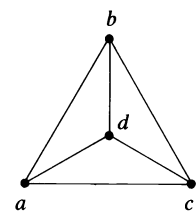
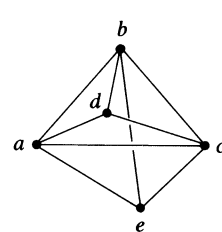
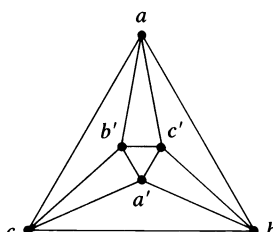
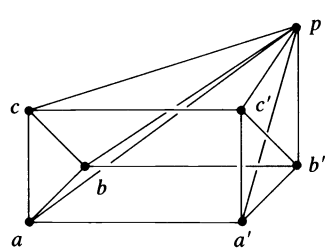
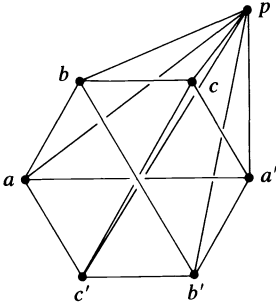
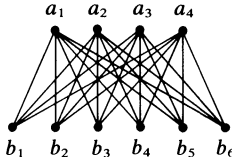
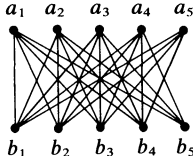
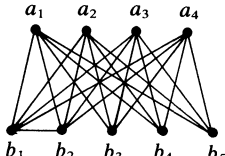
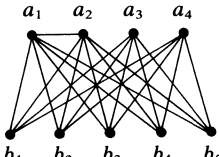
Name and graph	Pure condition and geometric condition
<p>2.1. Tetrahedron</p> 	<p>The four points are coplanar</p> $[abcd]$
<p>2.2. Triangular bipyramid</p> 	<p>One of the quadruples $abcd$ or $abce$ is coplanar</p> $[abcd][abce]$
<p>2.3. Octahedron</p> 	<p>Four alternate face planes abc, $ab'c'$, $a'bc'$, $a'b'c$ are concurrent in a point</p> $[abc'a'] [bca'b'] [cab'c'] + [abc'b'] [bca'c'] [cab'a']$
<p>2.4. 1-point cone on the prism</p> 	<p>Projected from p onto a plane, the prism appears perspective from a line</p> $[abcp][a'b'c'p]([abb'p][a'cc'p] - [a'bb'p][acc'p])$

TABLE 2 (cont.)

Name and graph	Pure condition and geometric condition
2.5. 1-point cone on $K_{3,3}$	Projected from p onto a plane, the $K_{3,3}$ appears to lie on a conic
	$[abc p][ab'c'p][a'b'cp][a'bc'p]$ $-[a'bc p][a'b'c'p][ab'cp][abc'p]$
2.6. $K_{4,6}$	Either $a_1 a_2 a_3 a_4$ are coplanar or the 10 points lie on a quadric surface
	$[a_1 a_2 a_3 a_4]^2 Q(a_1 \cdots a_4, b_1 \cdots b_6)$
2.7. $K_{55} - \{a_5, b_5\}$	Either $a_1 a_2 a_3 a_4$ or $b_1 b_2 b_3 b_4$ are coplanar or the ten points lie on a quadric surface
	$[a_1 a_2 a_3 a_4][b_1 b_2 b_3 b_4] Q(a_1 \cdots a_5, b_1 \cdots b_5)$
2.8. $K_{4,5} + 1$ edge	Either $a_1 a_2 a_3 a_4$ are coplanar or the nine points $a_1 \cdots a_4 b_1 \cdots b_5$ and the line of the added edge lie on a quadric surface
	$[a_1 a_2 a_3 a_4] Q(a_1 \cdots a_4, b_1 \cdots b_5, (b_1 + b_2)/2)$
	$[a_1 a_2 a_3 a_4] Q(a_1 \cdots a_4, b_1 \cdots b_5, (a_1 + a_2)/2)$

4.1. Direct calculation of pure conditions. It is possible to directly decompose $\det(M(G, T))$ as a bracket expression, using a Laplace expansion which we will return to at the end of this section. In such a direct calculation, it is desirable to have the tie-down condition $C(T)$ appear as an immediate factor.

LEMMA 4.1. *If an n -isostatic graph includes the vertices v_1, \dots, v_n and the generic tie-down T is given with $\alpha_1 = n, \alpha_2 = n - 1, \dots, \alpha_n = 1$ and tie-down bars $\{v_i, x_{i,j}\}, i < j \leq n + 1$, then the tie-down factor is*

$$C(T) = [v_1 x_{1,2} \cdots x_{1,n+1}] [v_1 v_2 x_{2,3} \cdots x_{2,n+1}] \cdots [v_1 v_2 \cdots v_n x_{n,n+1}].$$

Proof. The tie-down factor is independent of the n -isostatic graph which includes the joints v_1, \dots, v_n . For convenience we will use the complete graph on these n joints and reshuffle the edges to give the order

$$\{v_1, x_{1,2}\}, \dots, \{v_1, x_{1,n+1}\}, \{v_1, v_2\}, \{v_2, x_{2,3}\}, \dots, \{v_{n-1}, v_n\}, \{v_n, x_{n,n+1}\}.$$

We now do a Laplace expansion on the last n -columns—the columns of v_n : The only nonzero term uses the last n -rows and is the bracket $[v_1, \dots, v_n, x_{n,n+1}]$ times the corresponding minor. We now do a Laplace expansion of this minor by the n -columns for v_{n-1} —giving only one nonzero term with the bracket $[v_1, \dots, v_{n-1}, x_{n-1,n}, x_{n-1,n+1}]$. We continue this process, finding the last factor (using the n columns for v_1) $[v_1, x_{1,2}, \dots, x_{1,n+1}]$. Thus $C(G, T) = [v_1, x_{1,2}, \dots, x_{1,n+1}] \cdots [v_1, \dots, v_n, x_{n,n+1}] = C(T)$.

The graph G has the pure condition 1 because such a complete graph on n joints on n -space is isostatic if and only if the joints span an affine $n - 1$ space—and this is true whenever $C(T) \neq 0$.

We conclude that $C(T)$ has the desired form. **Q.E.D.**

In any reasonable decomposition of $\det M(G, T)$, using this standard tie-down, the given brackets will appear as factors of each monomial of the decomposition. Thus no energy or ingenuity need be expended in pulling out this tie-down factor, and we always choose this tie-down in actual calculations of pure conditions.

The proof of Lemma 4.1 gives the following corollary. An n -simplex is the complete graph on $n + 1$ vertices.

COROLLARY 4.2. *The pure condition for an $n - 1$ simplex in n -space is 1.*

The pure condition for an n simplex in n -space is $[v_1, \dots, v_{n+1}]$.

Proof. The $n - 1$ simplex was directly given in the proof of Lemma 4.1.

To obtain the n simplex from the $n - 1$ simplex we add one vertex, v_{n+1} , and n edges $\{v_i, v_{n+1}\}, 1 \leq i \leq n$. We add n columns for v_{n+1} , and these n rows at the bottom of the matrix used in Lemma 4.1. A Laplace expansion by the last n columns gives the brackets $(\pm)[v_1 \cdots v_{n+1}]$ times the cofactor which is $C(T)$. **Q.E.D.**

Remark. It is clear from this analysis that if any graph G' is built from an n -isostatic graph G by adding a new n -valent vertex p with edges $\{p, a_i\}, 1 \leq i \leq n$, then the pure condition has the form $C(G') = [p, a_1, \dots, a_n]C(G)$.

In general $\det(M, T)$ can always be decomposed by taking a series of Laplace expansions on the n columns for each vertex in turn (Rosenberg, [11]). Such an expression will produce a sum of monomials, each of which has the form

$$\prod_i [a_i, b_{i,1}, \dots, b_{i,n}],$$

where the rows $\{a_i, b_{ij}\}, 1 \leq j \leq n$ were used for the columns of a_i in this term. The following useful property follows from this expression by a simple counting argument (using the simple tie-down).

LEMMA 4.3. *The pure condition for an n -isostatic graph G is homogeneous of degree $k + 1 - n$ in each vertex of valence k in the graph.*

4.2. Factors determined by decomposition of the graph. A second source of pure conditions lies in certain decompositions of the graph. We begin with the simplest result of this type.

PROPOSITION 4.4. *If G is an n -isostatic graph and H is an n -isostatic subgraph with at least $n + 1$ vertices, then $C(G) = C(H) \cdot C'$ for some factor C' .*

Proof. We attach the simple tie-down T to n vertices in H . The tie-down rows, plus the rows corresponding to edges in H now give a square submatrix, with all other entries in these rows zero, and a simple Laplace expansion, using these rows as a block, gives

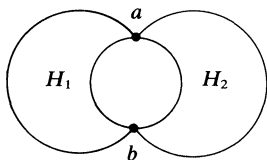
$$C(G, T) = C(H, T) \cdot C' = C(T) \cdot C(H) \cdot C'. \quad \text{Q.E.D.}$$

This result explains the illustrated factoring for examples such as the prism (Table 1, 1.3), the combinations of prisms (Table 1, 1.4 and 1.5), other planar graphs with triangles (Table 1, 1.7, 1.8) and the 1-point cone on a prism (Table 2, 2.4).

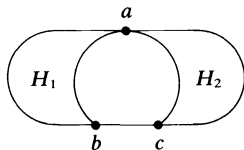
The form of C' depends on the pattern of the rest of the graph. When the number of edges or vertices of attachment to H is small, then we can give more details about C' . A number of such examples are illustrated in Table 3.

TABLE 3a
Plane

3.1 If H_1 is 2-isostatic then $C(G) = C(H_1) \cdot C(H_2 + \{a, b\})$. If neither H_1 nor H_2 is 2-isostatic $C(G) \equiv 0$



3.2 If H_1 and H_2 are 2-isostatic then $C(G) = C(H_1) \cdot [abc] \cdot C(H_2)$. Otherwise $C(G) \equiv 0$



3.3 If H_1 and H_2 are 2-isostatic then $C(G) = C(H_1) \cdot ([a_1 b_1 a_2] \cdot a_3 b_3 b_2] - [a_1 b_1 b_2][a_3 b_3 a_2]) \cdot C(H_2)$. Otherwise $C(G) \equiv 0$

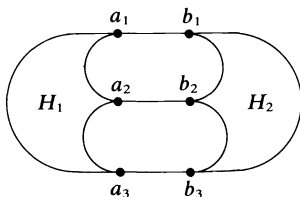
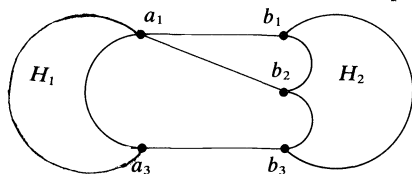


TABLE 3a (cont.)

3.4 If H_1 and H_2 are 2-isostatic then $C(G) = C(H_1) \cdot [a_1 b_1 b_2] \cdot [a_3 a_1 b_3] \cdot C(H_2)$



3.5 If H is 2-isostatic $C(G) = C(H) \cdot ([a_1 a_2 b_1] \cdots [a_k a_1 b_k] + (-1)^{k+1} [a_1 a_2 b_2] \cdots [a_k a_1 b_1])$

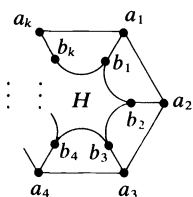
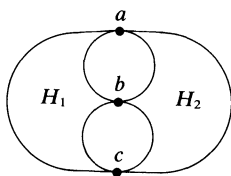
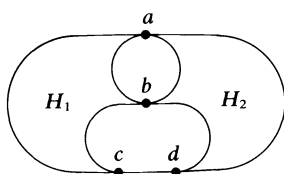


TABLE 3b
Space

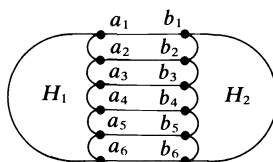
3.6 If H_1 is 3-isostatic then $C(G) = C(H_1) \cdot C(H_2 \cup \{a, b, \{b, c\}, \{c, a\}\})$



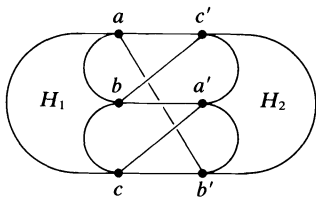
3.7 If H_1 is 3-isostatic then $C(G) = C(H_1) \cdot [abcd] \cdot C(H_2 \cup \{a, b\})$.
If neither H_1 nor H_2 is 3-isostatic then $C(G) = 0$



3.8 If H_1 and H_2 are 3-isostatic then $C(G) = C(H_1) \cdot C(H_2) \cdot C(S)$, where $C(S)$ is the "tie down" factor for $a_1 b_1, \dots, a_6 b_6$.



3.9 If H_1 and H_2 are 3-isostatic then $C(G) = C(H_1) \cdot C(H_2) \cdot ([bca'b'] [cab'c'] [abc'a'] + [cab'a'] [abc'b'] [bca'c'])$. If one of H_1, H_2 is not 3-isostatic then $C(G) = 0$.



The simplest examples in Table 3 cover two n -isostatic subgraphs tied together by $\binom{n+1}{2}$ edges (Table 3, 3.3, 3.4, 3.8, 3.9). Such a tie-together S gives a factoring

$$C(G) = C(H_1) \cdot C(S) \cdot C(H_2),$$

where $C(S)$ is a specialization of the tie-down factor due to any identification of the ends of the tie-together edges. If we tie down H_1 , then H_1 functions as a “ground” while S gives a tie-down for H_2 .

A second property of a graph which is reflected in the pure condition concerns 1 point cones (Table 2, 2.4, 2.5).

DEFINITION 4.1. Given a graph $G = \langle V, E \rangle$ the 1-point cone $G * p$ is the graph with vertices $V \cup \{p\}$ and edges $E \cup \{\{p, v_i\} | v_i \in V\}$.

PROPOSITION 4.5. *If G is an n -isostatic graph with condition $C(G)$, then the 1 point cone $G * p$ is an $(n + 1)$ -isostatic graph with pure condition $C(G * p) = C(G) * p$, where $(L) * p$ means extending each bracket in L by inserting an $(n + 1)$ st entry p .*

Proof. We take the standard tie-down T_n on G , at vertices a_1, \dots, a_n , and obtain $M(G, T_n)$ and take the standard tie-down T_{n+1} of $G * p$, at vertices p, a_1, \dots, a_n to obtain $M(G * p, T_{n+1})$.

We obtain $\det M(G, T_n) = C(T_n) \cdot C(G)$ by a series of Laplace expansions, each on the n columns of a vertex of G , and similarly $\det M(G * p, T_{n+1}) = C(T_{n+1}) \cdot C(G * p)$ by a series of Laplace expansions each on the $n + 1$ columns of a vertex of $G * p$. The term for the columns of p is part of the tie-down factor, $[py_1 \cdots y_{n+1}]$, and for each other vertex v_i the nonzero terms will involve n rows of G , plus the row for $\{p, v_i\}$. Otherwise, some term has no occurrence of p —another term has two occurrences of p —and we have the zero term. Since these two expansions give the form $C(G * p, T_{n+1}) = [py_i \cdots y_{n+1}] \cdot (C(G, T_n) * p)$ and $C(T_{n+1}) = [py_i \cdots y_{n+1}] \cdot (C(T_n) * p)$ we have the desired result. Q.E.D.

Remark. This proposition reflects the geometric theorem that a framework realizing a 1-point cone in $(n + 1)$ -space with apex p has a static stress if and only if the projection of the framework from p into an n -space has a stress in the n -space [19, § 10]. The geometric process of projection is expressed in the brackets as a reduction by p —placing p in each bracket as the last entry—and then deleting the p 's, thus moving from brackets of length $n + 2$ to brackets of length $n + 1$.

4.3. Factors and pure conditions by geometry. Other direct analyses of the infinitesimal and static behavior of frameworks have produced projective geometric statements of sufficient (and necessary) conditions for nontrivial motions or stresses in frameworks with various graphs [19], [20], [21], [22]. Either by the direct presentation, or by a simple translation, such projective conditions for n -isostatic graphs reduce to a single polynomial $F(G)$ such that $F(G) = 0$ is sufficient for a realization of G to not be isostatic.

LEMMA 4.6. *If a polynomial F in the vertices of an n -isostatic graph G has the property that $F(G) = 0 \Rightarrow C(G) = 0$, then each irreducible factor of F is a bracket expression which is a factor of $C(G)$.*

Proof. The analysis of $C(G)$, and of the sufficient conditions $F(G) = 0$, are done over the complex numbers. By Hilbert's Nullstellensatz we have

$$(F(G) = 0 \Rightarrow C(G) = 0) \Rightarrow A \cdot F(G) = (C(G))^r,$$

where A is some bracket expression and r is a positive integer. Clearly each irreducible factor of F is a factor of $(C(G))^r$ —and thus of $C(G)$. Q.E.D.

Lemma 4.6 can give us some factors of $C(G)$, based on our geometric analyses. When combined with Lemma 4.3, which limits the total occurrences of each vertex in $C(G)$, it is possible to count how many, if any, occurrences of joints remain after this factoring. In many cases these two lemmas give a complete description of the pure condition.

For example, the factors used in the condition for the prism (Table 1, 1.3), the $K_{3,3}$ (Table 1, 1.5), and other plane examples (Table 1, 1.7, 1.8), as well as the octahedron in space (Table 2, 2.3) were originally calculated by a direct geometric analysis. In each case these known factors contain all available occurrences of the joints—so they must be the pure condition, provided they are irreducible. In § 4.4 we will verify this irreducibility.

A more surprising class of examples includes the bipartite framework $K_{4,6}$ in 3-space (Table 2, 2.6). We recall that a bipartite graph $K_{m,n}$ has vertices $V = \{a_1, \dots, a_m, b_1, \dots, b_n\}$ and edges $\{a_i, b_j\}$, $1 \leq i \leq m$, $1 \leq j \leq n$. It is a simple counting argument to check that $K_{n+1,m}$, where $m = \binom{n+1}{2}$, counts to be an n -isostatic graph—or at least give a square matrix $M(G, T)$.

PROPOSITION 4.7. *The pure condition in n -space for the bipartite graph $K_{n+1,m}$, where $m = \binom{n+1}{2}$, is*

$$[a_1 \cdots a_{n+1}]^d \cdot Q(a_1, \dots, a_{n+1}, b_1, \dots, b_m),$$

where $d = (n + 1)(n - 2)/2$ and $Q(a_1, \dots, b_m)$ is the bracket expression for all the joints to lie on a quadric surface in n -space.

Proof. By the analysis of [22] if the joints lie on a quadric surface, then there is an infinitesimal motion. The expression for the joints to be on a quadric surface is a projectively invariant equation of degree 2 in each joint, since picking values for all but 1 joint must leave a general quadric equation. (This expression is irreducible, as we will see in Proposition 4.9.) When this factor Q is removed from the pure condition, we are only left with the vertices a_1, \dots, a_{n+1} , which still occur to degree $d = \binom{n+1}{2} - (n + 1)$. The only possible nonzero n -bracket formula with $n + 1$ vertices is $[a_1 \cdots a_{n+1}]$, so this occurs d times. (The factor is nonzero, since the graph does have n -isostatic realizations.) Q.E.D.

Remark 1. The factor $[a_1 \cdots a_{n+1}]$ ($n \geq 3$) was also predicted by the geometric analysis, since [22, Thm. 1.1] guarantees a nontrivial infinitesimal motion whenever all points a_1, \dots, a_{n+1} lie in a quadric surface of a hyperplane in n -space. In fact $[a_1 \cdots a_{n+1}] = 0$ guarantees that these joints lie on d such quadric surfaces of the hyperplane, thus giving d motions and d stresses to match the d identical factors. We return to this “coincidence” in Chapter 6.

Remark 2. Since, for example, $K_{4,6}$ has 2 stresses when $[a_1 \cdots a_4] = 0$, we also know that removing one bar in that case will still leave at least 1 stress, regardless of the position of the b_i . Removing $\{a_4, b_6\}$ also leaves b_6 as a 3-valent joint, which will not participate in the dependence unless b_6 is in the plane of $a_1 a_2 a_3$. Thus the condition $[a_1 \cdots a_4] = 0$ actually is sufficient for a stress in the 1-underbraced frameworks realizing $K_{4,5}$. As a result, any graph G containing this $K_{4,5}$ as a subgraph must have $[a_1 \cdots a_4]$ as a factor of its pure condition. By a similar argument, we find $K_{n+1, n+2}$ can be a strongly $(d - 1)$ underbraced graph for $n > 3$ which still induces a factor $[a_1 \cdots a_{n+1}]$ in the pure condition of any n -isostatic graph containing it.

Remark 3. We have previously observed (§ 4.2) that the presence of an n -simplex gives the factor $[a_1 \cdots a_{n+1}]$. We have now found that a bipartite framework with none of these edges present can give the same factor. There is no simple correlation between factors and subgraphs. Similar factors give a similarity in the geometric

conditions under which a framework is critical (not-isostatic). However we conjecture that if two graphs give the same pure condition, then the frameworks are the same.

By an argument similar to the proof of Proposition 4.7, we conclude that the graph $K_{5,5} - \{a_5, b_5\}$ has the pure condition given in Table 2, 2.7. The simplest known bracket expression for $Q(a_1, \dots, b_5)$ is given in [16, p. 266] as the sum of 240 bracket monomials! The same paper gives expressions for a quadric through 9 points and 1 line (e.g., $Q(a_1, \dots, a_4, b_1, \dots, b_5, (a_1 + a_2)/2)$ in Table 2, 2.8) as the sum of 6 monomials. The pure condition for this last example follows from the geometric analysis of [22, Thm. 4.1] by an analogous argument.

4.4. Irreducibility of factors of the pure condition. In the tables and the preceding discussion we have offered many “irreducible” factors. However these were often enormous polynomials in, say 40 variables (for $K_{4,6}$), and it requires some proof to see that such expressions are irreducible.

At present we have two main tools for proving this irreducibility—symmetries of the graph, or the corresponding geometric conditions, which would impose symmetry on the factoring, and specializations of the coordinatization which reduce the expression to a simpler form, which is either known to be irreducible or else factors only in a way which is incompatible with the original symmetries or geometry of the condition. Without being exhaustive, we illustrate these techniques on the simple examples given in the tables.

PROPOSITION 4.8. *The tie-down factor $C(T)$ for $T = \{\{a_i, x_i\} | 1 \leq i \leq \binom{n+1}{2}\}$ is irreducible if a_i, x_i are distinct.*

Proof. We view T as a tie-down of a rigid body—and recognize that this factor is independent of the n -isostatic graph used to connect the a_i . We take a specialization Φ which identifies the a_i with appropriate joints b_j of the standard tie-down and gives a homomorphism of the polynomial

$$\Phi(C(T)) = [b_1 y_{1,2} \cdots y_{1,n+1}] [b_1 b_2 y_{2,3} \cdots y_{2,n+1}] [b_1 \cdots b_n y_{n,n+1}].$$

From this factoring of $\Phi(C(T))$ we conclude, for example that $\Phi^{-1}(y_{1,2}), \dots, \Phi^{-1}(y_{1,n+1})$ (some of the x_i) must be in the same factor of $C(T)$. But this identification was arbitrary—so all the x_i lie in the same factor. Similarly, $\Phi^{-1}(b_n)$ and $\Phi^{-1}(y_{n,n+1})$ must be in the same factor of $C(T)$. However $\Phi^{-1}(b_n)$ is an arbitrary a_i —so we conclude that all a_i and x_i are in the same factor. Since $C(T)$ is of first degree in all a_i and x_i , this requires that $C(T)$ is irreducible. Q.E.D.

If we apply this result to the tie-together of 2 bodies, we have shown the irreducibility of the factors in the examples in Table 1, 1.3, 1.4, 1.5. The result for 1-point cones also gives the factoring of $C(G * p)$, so we have also explained the factors of Table 2, 2.4.

PROPOSITION 4.9. *The bracket condition Q_n for $\binom{n+2}{2}$ points to lie on a quadric surface in n -space is irreducible.*

Proof. Assume that the condition Q_n factors as $f \cdot g$.

Case 1. f is of first degree in some point a . Since Q is of degree 2, g is also of first degree in a . By a suitable choice of real position for all the other points (which define, in general, a unique quadric) the condition Q can be specialized to $Q(a) = a_1^2 + a_2^2 = f(a) \cdot g(a)$. Since $a_1^2 + a_2^2$ is irreducible, we have a contradiction.

Case 2. F is of degree 2 in some set of points a, \dots, c , while g is of degree 2 in the remaining points d, \dots, f . However the geometric condition is symmetric in all points, so if a and c share a factor, so must a and d , etc. We conclude that all points appear in the factor f , and $g = 1$. Q.E.D.

If the factor being examined is small (a sum of monomials with ≤ 3 brackets), then any further factoring must include a factor which is a sum of single brackets. However the factors are homogeneous, so there must be a single bracket as a factor—and thus some set of $n + 1$ joints which, if coplanar, would induce a stress. In many cases this possibility can be eliminated by a direct inspection of the geometry.

Consider example Table 1, 1.7. The third factor is reducible if and only if some *triple of joints in this factor* being collinear causes a stress. Since we know the geometric condition (see the table), it is simple to check that any such triple can be collinear while the factor is $\neq 0$. We conclude that this third factor is irreducible.

The octahedron (Table 2, 2.3) also gives an irreducible factor. By Cauchy’s theorem and its extensions [23] any set of 4 joints such as $abcc'$ or $aba'b'$ can be made coplanar without making the condition $= 0$. By the symmetries of the graph the same is true for all quadruples, so the condition is an irreducible polynomial.

If we take any planar graph ($|V| - |E| + |F| = 2$) with $|E| = 2|V| - 3$, $|V| \geq 3$, a simple counting argument shows that there must be at least one triangle. Thus a planar graph with more than 3 vertices must have factors in its pure condition. Only nonplanar graphs, such as $K_{3,3}$, can have irreducible pure conditions.

The situation for 3-isostatic graphs is much more complicated. While the construction of general 3-isostatic graphs is an unsolved problem, some classes, such as triangulated spheres, are well known [23].

5. Overbraced frameworks. So far we have been considering isostatic frameworks, which may be described as maximal frameworks which have no stress in generic position. Let $I(v) = nv - \binom{n+1}{2}$. Then we know that an isostatic framework has exactly $I(v)$ bars. We now consider *overbraced* frameworks, that is, frameworks with more than $I(v)$ bars. Such frameworks always have a stress. Let G be such a framework. Then every subframework G' of G having exactly $I(v)$ bars is either generically isostatic, in which case we have the pure bracket condition $C(G')$ for the existence of a stress of G' , or else G' generically has a stress, in which case the rigidity matrix of G' has dependent rows, and we may set $C(G') = 0$.

If G is a 1-overbraced framework, that is, a framework with exact $I(v) + 1$ bars, we can actually compute the coefficients of a stress of G , using brackets.

THEOREM 5.1. *Let G be a 1-overbraced framework with bars E_0, E_1, \dots, E_I such that for some j , $G - E_j$ is isostatic. Then there exists a nonzero stress on G whose value on the bar E_i is $(-1)^i C(G - E_i)$, where the bars of $G - E_i$ are taken in order of subscript in computing $C(G - E_i)$.*

Proof. Choose a nondegenerate generic tie-down T , and let M be the generic rigidity matrix of $G \cup T$. Since M is an $(nv + 1) \times nv$ matrix, its rows are linearly dependent and the coefficients of any such dependence are the values of a stress. Let F_i be the i th row of M , $i = 0, 1, \dots, nv + 1$, and M_i the matrix M with the row F_i deleted. Then, by Cramer’s rule, $\sum_{i=0}^{nv+1} (-1)^i (\det M_i) F_i = 0$. If $i > I + 1$, so that F_i is a row corresponding to one of the tie-down bars, then since the framework $G \cup T$ with one tie-down bar removed is clearly stressed, $M_i = 0$. Thus we have

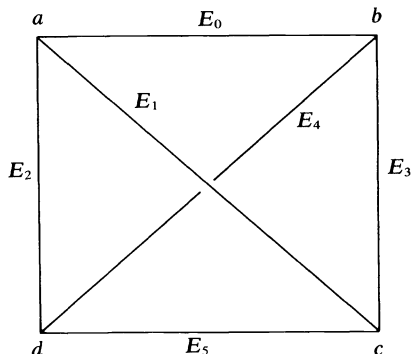
$$\sum_{i=0}^{I+1} (-1)^i C(G - E_i, T) F_i = 0.$$

Now each term has the same tie-down factor $C(T)$, and we chose T to be nondegenerate, so $C(T) \neq 0$. Thus $C(T)$ may be factored out, leaving

$$\sum_{i=0}^{I+1} (-1)^i C(G - E_i) F_i = 0.$$

Finally, we note that since $G - E_j$ is isostatic by hypothesis, the coefficient $C(G - E_j)$ is generically nonzero, so we have the desired stress. Q.E.D.

Example 5.2.



The tetrahedral framework in the plane is a 1-overbraced framework. The stress coefficients $(-1)^i C(G - E_i)$ are tabulated below.

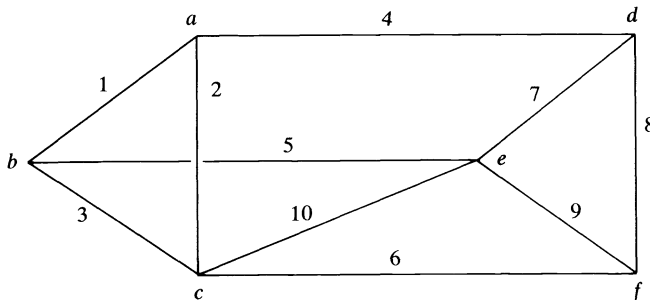
E_0	$[acd][bcd]$	E_3	$[abd][acd]$
E_1	$-[abd][bcd]$	E_4	$-[abc][acd]$
E_2	$[abc][bcd]$	E_5	$[abc][abd]$

The determination of the signs of the coefficients directly from the definition of $C(G - E_i)$ is tricky; in any case the best we can do is determine the relative signs of two coefficients. One way to determine the signs is to recall that by the definition of a stress, if S_{ab} is the value of a stress on a bar ab , then for each $a \in V, \sum_{b \in V, ab \in G} S_{ab} \cdot \mathbf{ab} = 0$, where \mathbf{ab} is the vector $a - b$. This is equivalent to the Cayley algebra equation $\sum_{b \in V, ab \in G} (-1)^i C(G - E_i) ab = 0$, where ab here is a step-two extensor. But this equation must be a syzygy in the Cayley algebra, which can be determined directly. For example,

$$[acd][bcd]ab - [abd][bcd]ac + [abc][bcd]ad = 0$$

is a syzygy in the Cayley algebra, and this gives the correct relative signs for E_0, E_1 and E_2 . Alternatively, we may specialize coordinates to any particular coordinatization whose stress coefficients have known signs in order to determine the correct signs generically.

Example 5.3.



This framework G is the triangular prism with one additional bar ce , in the plane. The stress coefficient $(-1)^i C(G - E_i)$ is given for each bar E_i , now denoted simply i .

$$\begin{aligned}
 1 & [acd][bce][cef][def] \\
 2 & -[abd][bce][cef][def] \\
 3 & [abe][acd][cef][def] \\
 4 & [abc][bce][cef][def] \\
 5 & -[abc][acd][cef][def] \\
 6 & [abc][ade][bce][def] \\
 7 & [abc][adf][bce][cef] \\
 8 & -[abc][ade][bce][cef] \\
 9 & [abc][ade][bce][cdf] \\
 10 & [abc][def]([abd][cef] - [ade][bcf]).
 \end{aligned}$$

Now that we have a bracket formulation for a stress on a 1-overbraced framework in generic position, the stress for any particular coordinatization (or realization) of the framework in real n -space may be computed by simply plugging in the values for the brackets for that particular coordinatization. One especially important piece of information which can be obtained from the stress values is the split (or partition) of tension members versus compression members. This is obtained simply by observing which bars have a positive value in the stress, and which have negative. The fact that we have determined only the relative signs of the stress coefficients is no problem, since the split between tension and compression members is reversible.

Example 5.3 (continued). Let us fix the positions of a , c , d , e and f in the framework G of Example 5.3 and think of b as moving around. The irreducible factors which involve b and which occur in the stress coefficients are $[abc]$, $[abd]$, $[abe]$, $[bce]$ and $([abd][cef] - [ade][bcf])$. The locus of points which makes each of these factors 0 is a curve; in this case each is a line, namely, ac , ad , ae , ce and $(ad \cap cf)e$, respectively (note that the last factor listed is equivalent to the Cayley algebra expression $ad \wedge be \wedge cf$). We will call these curves *switching curves for b* (or *switching surfaces for b* for examples in 3-space).

We could more generally consider the 12-dimensional space of all affine realizations of $\{a, b, c, d, e, f\}$ and consider all of the irreducible factors of the stress coefficients, obtaining various *switching hypersurfaces*.

For an arbitrary framework, if a vertex b lies on one of the switching surfaces, then some of the stress coefficients are zero and the support of the stress is a proper subframework of G . If we move b from one side of the switching surface to the other, the corresponding factor switches sign (assuming that we are doing our factoring over \mathbb{R}), thus all members having that factor to an odd power switch from tension to compression, or vice-versa, while the remaining members do not switch. Thus a crucial question is when an irreducible factor of a pure condition can occur to a higher power than one, or, more precisely, what are the relative powers of a given factor in the stress coefficients.

We illustrate in Fig. 1 the switching curves for the vertex b in one realization of G . We also illustrate the tension compression split for several of the plane regions determined by the switching curves. Once the split has been determined in one region,

it may quickly be determined in all others by successively crossing switching curves, one at a time and switching members accordingly. We show switching curves by dotted lines, tension members by dashed lines and compression members by cross-hatched lines.

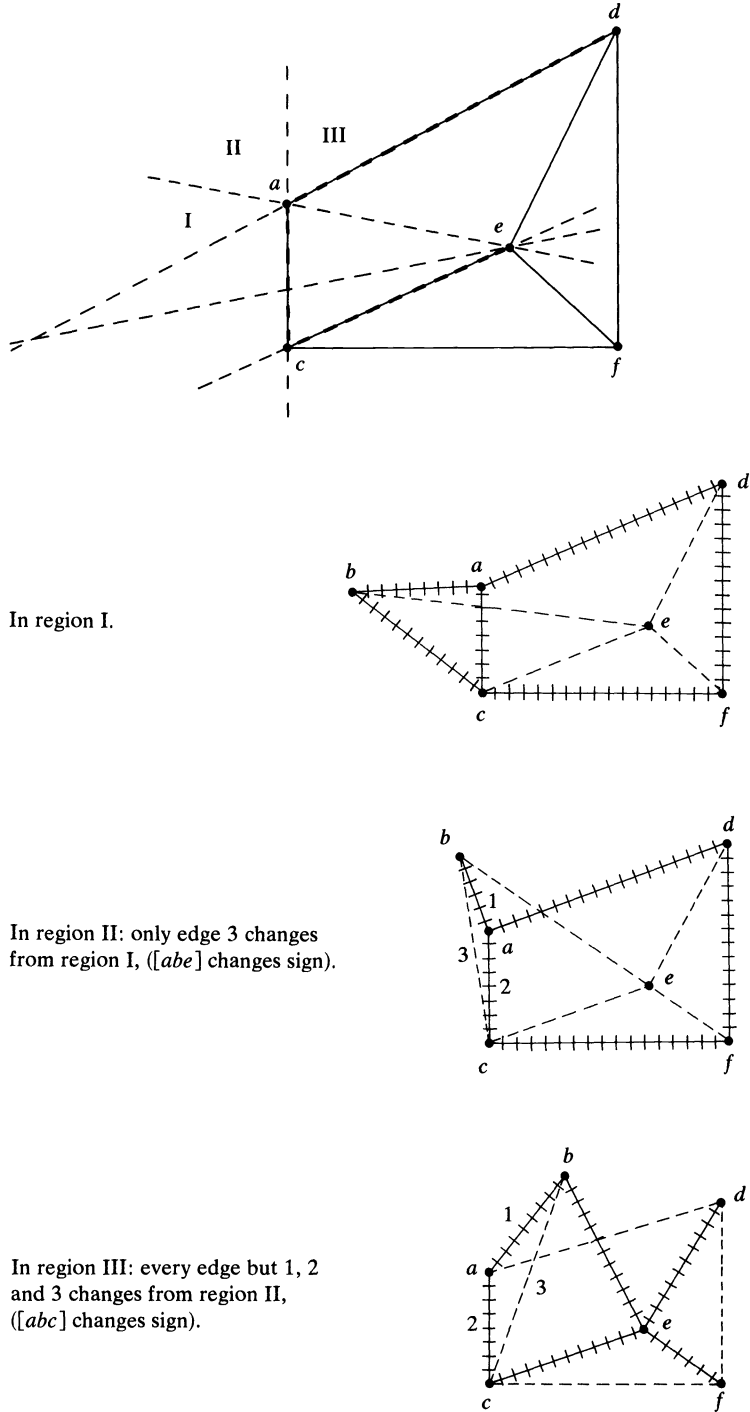


FIG. 1

6. Multiple stresses. In our current list of examples, we have no example where a factor occurs to a k th power in one coefficient of a stress and a power $< k - 1$ in some other coefficient. This would seem to indicate that every factor of a pure condition gives a switching surface for every 1-overbraced framework containing this subgraph, unless this factor occurs to the same power in all coefficients (for example, $K_{5,5}$).

The only examples we have of a factor occurring to higher than the first power are in the bipartite frameworks. In these examples the multiple occurrences of a factor are associated with a multiple stress. The following result generalizes this observation.

PROPOSITION 6.1. *If, for some irreducible factor H of the pure condition of an n -isostatic graph G , all coordinatizations α with $H(\alpha(G)) = 0$ give at least r stresses, then H^r is a factor of $C(G)$.*

Proof. For $r = 1$, this is Lemma 4.6. We proceed by induction on r , with the additional assumption that some α with $H(\alpha(G)) = 0$ give joints which span the space. If, on the contrary, all such α gave flat coordinatizations, then $H = 0$ implies $[a \cdots d] = 0$ for each $(n + 1)$ -tuple a, \cdots, d , so H would be precisely such a single bracket. This, together with an over-all flatness in G requires that G is the n -simplex, where we know $C(G) = [a \cdots d]$ and $r = 1$.

Since $\alpha(G)$ is not flat for a generic coordinatization with $H(\alpha(G)) = 0$, and the complete graph is statically rigid in such coordinatizations, we can find an extra bar E which is independent in $\alpha(G + E)$. We now examine the stress equation for the 1-overbraced framework $G + E$:

$$\sum (-1)^i C(G + E - E_i) F_i + C(G) F_E = 0.$$

We assume $H(\alpha(G)) = 0$ gives an $(r + 1)$ -tuple stress. This must also give a r -tuple stress in $\alpha(G + E - E_i)$ and by our induction hypothesis H is an r -fold factor of all these coefficients. We factor H^r out of the stress equation, leaving:

$$\sum (C_i) F_i + C' F_E = 0.$$

Since E is independent for the generic α with $H(\alpha(G)) = 0$, we have $H = 0$ implies $C' = 0$. Over the complex numbers this gives, via Hilbert's Nullstellensatz, $AH = (C')^s$ for some integer s . Since H is irreducible, H divides C' .

We conclude that H is an $(r + 1)$ -fold factor of $C(G)$. Q.E.D.

Is the converse of this proposition true? If an irreducible factor H is an r -fold factor of $C(G)$ do all coordinatizations α with $H(\alpha(G)) = 0$ give r stresses? This problem remains a basic block to a good analysis of the behavior of a stress at a "switching surface".

REFERENCES

- [1] G. BLAHA, *Investigations of critical configurations for fundamental range networks*, Dept. of Geodetic Science Report 150, Ohio State Univ., Columbus, OH, 1971.
- [2] E. BOLKER AND B. ROTH, *When is a bipartite graph a rigid framework?*, Pacific J. Math., 90 (1980), pp. 27–44.
- [3] R. CONNELLY, *Rigidity and energy*, Preprint, Dept. Mathematics, Cornell University, Ithaca, NY, 1980.
- [4] H. CRAPO, *Structural Rigidity*, Structural Topology, 1 (1979), pp. 26–45.
- [5] H. CRAPO, *A combinatorial perspective on algebraic geometry*, International Colloquium on Combinatorial Theory, Vol. II, Academia Nazionale dei Lincei, Rome, 1976, pp. 343–357.
- [6] J. DÉSARMÉNIEN, J. KUNG AND G. C. ROTA, *Invariant theory, Young bitableaux and combinatorics*, Advances in Math., 27 (1978), pp. 63–92.
- [7] P. DOUBLET, G. C. ROTA AND J. STEIN, *On the foundations of combinatorial theory: IX combinatorial methods in invariant theory*, Studies in Appl. Math., 53 (1974), pp. 185–216.

- [8] M. HALL, *Combinatorial Theory*, Blaisdell, Waltham, MA, 1967.
- [9] L. HENNEBERG, *Die Graphische Statik der Starren Systeme*, Leipzig, 1911 (Johnson Reprint, 1968).
- [10] W. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, Vol. I, Cambridge Univ. Press, Cambridge, 1968.
- [11] I. ROSENBERG, *Structural rigidity in the plane*, preprint C.M.R. 510, University of Montreal, Montreal, Quebec.
- [12] B. ROTH AND W. WHITELEY, *Tensegrity frameworks*, Trans. Amer. Math. Soc., 265 (1981), pp. 419–446.
- [13] K. SUGIHARA, *A unifying approach to descriptive geometry and mechanisms*, preprint RMI-81-01, Faculty of Eng., Nagoya Univ., Furō-chō, Chikusa-ku, Nagoya, Japan, 1981.
- [14] K. SUGIHARA, *Mathematical structures of line drawings of polyhedra—toward man-machine communication by means of line drawings*, Research Note RNS 81-02, Faculty of Eng., Nagoya Univ., Furō-chō, Chikusa-ku, Nagoya, Japan, 1981.
- [15] E. TSIMIS, *Critical configurations (determinantal loci) for range and range difference satellite networks*, Dept. Geodetic Science Report 191, Ohio State Univ., Columbus, OH, 1973.
- [16] H. TURNBULL AND A. YOUNG, *Linear invariants of ten quaternary quadrics*, Trans. Camb. Phil. Soc., 23 (1926), pp. 265–301.
- [17] N. WHITE, *The bracket ring of a combinatorial geometry I*, Trans. Amer. Math. Soc., 202 (1975), pp. 79–95.
- [18] W. WHITELEY, *Logic and invariant theory II: homogeneous coordinates, the introduction of higher quantities, and structural geometry*, J. Algebra, 50 (1978), pp. 380–394.
- [19] ———, *Introduction to structural geometry I: infinitesimal motions and infinitesimal rigidity*, preprint, Champlain Regional College, St. Lambert, Quebec, 1977.
- [20] ———, *Introduction to structural geometry II: statics and stresses*, preprint, Champlain Regional College, St. Lambert, Quebec, 1978.
- [21] ———, *Motion, stresses and projected polyhedra*, preprint, Champlain Regional College, St. Lambert, Quebec, 1981.
- [22] ———, *Infinitesimal motions of a bipartite framework*, preprint, Champlain Regional College, St. Lambert, Quebec, 1981.
- [23] ———, *Infinitesimally rigid polyhedra I: frameworks*, to appear.

EFFICIENT OPTIMIZATION OF MONOTONIC FUNCTIONS ON TREES*

YEHOASHUA PERL[†] AND YOSSI SHILOACH[‡]

Abstract. The problem of optimizing weighting functions over all the k -subtrees (subtrees with k vertices) of a given tree is considered. A general algorithm is presented that finds an optimal k -subtree of a given tree whenever the weighting function is what we call monotonic. Monotonicity is a very natural property, satisfied by most of the functions that one can think of.

The problem is solved for cases of both rooted and undirected trees. On the other hand, even simple extensions of it to general graphs are NP-hard.

Key words. optimal subtrees, monotonic weighting functions, dynamic programming, polynomial algorithms, NP completeness

1. Introduction. Let T be a rooted tree.

Henceforth a *subtree* of T will always mean a rooted subtree of T , and a k -*subtree* will mean a subtree with k vertices. A *complete* subtree of T is one that contains all the descendants of its root.

Let W be a weighting function that assigns a real number (called weight) to every subtree of T .

A k -subtree T' is *maximal (minimal)* with respect to W if it is the heaviest (lightest) k -subtree of T . The word *optimal* will later be used for both maximal and minimal. Let T, T_1 and T_2 be rooted trees, and let v be a vertex of T . Let $T'_1 (T'_2)$ be the tree obtained by hooking $T_1 (T_2)$ on T at v . We say that W is a *monotonic* weighting function if for any such triple T, T_1, T_2 and any vertex v of T ,

$$W(T_1) \leq W(T_2) \text{ implies } W(T'_1) \leq W(T'_2).$$

Optimization of monotonic weighting functions for tree partitioning is discussed in [BP].

In this paper we present a general and efficient algorithm for finding optimal k -subtrees of a given tree, which is good for any monotonic weighting function. Since most of the weighting functions that one can think of, including the following several examples, are monotonic, this algorithm might be proved very applicable. Moreover, in many cases this algorithm can be applied directly to corresponding problems in undirected trees. Furthermore, it will be shown that even some nonmonotonic functions, like the diameter, can still be optimized if monotonic auxiliary functions are properly used.

Examples of several natural weighting functions. Let $w: V \rightarrow \mathcal{R}$ and $l: E \rightarrow \mathcal{R}$ be two real-valued functions that assign weights to the vertices and lengths to the edges of T respectively.

Let $T' = (V', E')$ be given subtree of T rooted at r' . Consider the following weighting functions:

1. $W1(T') = \sum_{v \in V'} w(v)$,
2. $W2(T') = \sum_{e \in E'} l(e)$,
3. $W3(T') = \min_{v \in V'} w(v)$,
4. $W4(T') = \min_{e \in E'} l(e)$,
5. $W5(T') =$ number of terminal vertices in T' .

* Received by the editors November 19, 1980, and in revised form December 27, 1982.

† Department of Computer Sciences, Rutgers University, New Brunswick, New Jersey 08903 and Bar-Ilan University, Ramat-Gan, Israel.

‡ IBM Israel Scientific Center, Technion City, Haifa, Israel.

Defining the *distance* $d(u, v)$ from a vertex u to its descendant v as the sum of lengths of the edges along the path connecting them, we can further define:

$$6. \text{ Internal path length: } W6(T') = \sum_{v \in V'} d(r', v),$$

$$7. \text{ Height: } W7(T') = \max_{v \in V'} d(r', v).$$

Similar functions can be defined for undirected trees and undirected graphs as well. For almost all these functions, though, the corresponding problems of finding optimal k -subgraphs or even optimal k -subtrees in general graphs are NP-hard.

In the next section we describe the algorithm and analyze its complexity.

In § 3, the corresponding problem for undirected trees is defined and solved with the aid of the algorithm below. It is also shown how the nonmonotonic diameter function is maximized by utilizing an auxiliary monotonic function.

Section 4 contains a brief discussion on the NP-hardness of the problems of optimizing the weighting functions above over k -subgraphs and k -subtrees of general graphs.

2. The algorithm. The algorithm is based on dynamic programming. The common bottom-up approach requires exponential time, and therefore an efficient variation of it is used.

Let T be a tree rooted at r , and let r_1, \dots, r_s be r 's sons. Let T_1, \dots, T_s be T 's complete subtrees rooted at r_1, \dots, r_s respectively. Given a monotonic weighting function W and an integer k , assume that optimal k -subtrees T'_1, \dots, T'_s of T_1, \dots, T_s , respectively, have already been found. The optimal k -subtree of T is either the optimal subtree among T'_1, \dots, T'_s or another k -subtree rooted at r . Thus, there remains to be found an optimal k -subtree among those rooted at r . Such a subtree is called an *r -optimal* subtree. It turns out, however, that in order to find an r -optimal k -subtree, one has to find r_j -optimal i -subtrees for all $1 \leq i \leq k$ in each of the subtrees T_j , $1 \leq j \leq s$. Let $T'_j(i)$ denote an r_j -optimal i -subtree of T_j for all $1 \leq j \leq s$ and $1 \leq i \leq k$. Adopting a well-known dynamic programming technique, we assume that these trees are already known and proceed to find r -optimal i -subtrees for all $1 \leq i \leq k$. At this point, the straightforward approach requires that for each sequence i_1, \dots, i_s such that $i_1 + \dots + i_s = k-1$, a corresponding sequence of $T'_j(i_j)$, $1 \leq j \leq s$, will be considered. Since the number of such sequences is $O(k^{\deg(r)})$, this approach is intractable when r has many sons. It turns out, however, that a left-to-right propagation will save us this time. To this end let \underline{T}_j , $1 \leq j \leq s$, be the r -rooted subtree of T , spanned by T_1, \dots, T_j and r . As expected, let $\underline{T}'_j(i)$ denote an r -optimal i -subtree of \underline{T}_j for all $1 \leq j \leq s$ and $1 \leq i \leq k$. Since \underline{T}_1 is the subtree spanned by T_1 and r , it follows from the monotonicity that $\underline{T}'_1(i)$ can be taken as the subtree spanned by $T'_1(i-1)$ and r , $1 \leq i \leq k$. It should now be shown that the trees $\underline{T}'_{j+1}(i)$, $1 \leq i \leq k$, can be efficiently obtained from $\underline{T}'_j(i)$ and $T'_{j+1}(i)$, $1 \leq i \leq k$. Let i_0 be a fixed integer between 1 and k . For all i , $1 \leq i \leq i_0$, denote by $\underline{T}T'_{j+1}(i, i_0)$ the i_0 -subtree rooted at r that is spanned by $\underline{T}'_j(i)$ and $T'_{j+1}(i_0-i)$ (see Fig. 1).

The following simple theorem both motivates the algorithm below and establishes its validity.

THEOREM. $\underline{T}'_{j+1}(i_0)$ can be chosen as one of the subtrees $\underline{T}T'_{j+1}(i, i_0)$, $1 \leq i \leq i_0$, that optimizes W .

Proof. Follows immediately from the monotonicity of W . \square

In order to translate the ideas above into a more formal algorithm, let us associate a sequence $S(v)$ of length k with each vertex $v \in T$. $S(v)[i]$, the i th element of $S(v)$, should be by the end of the algorithm a v -optimal i -subtree. Let v be given, let v_1, \dots, v_s be its sons and let \underline{T}_j be defined as above for $1 \leq j \leq s$. Denote by $S_j(v)$ a sequence whose i th element is a v -optimal i -subtree of \underline{T}_j rather than T .

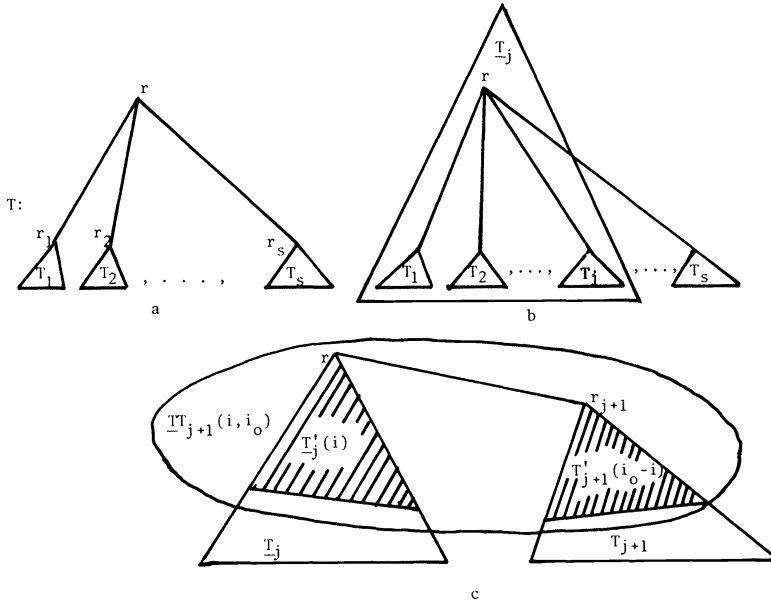


FIG. 1

If the sequences $S(v_1), \dots, S(v_s)$ have already been constructed, then v 's sequence can be obtained by:

CONSTRUCT $S(v)$.

Step 1. $S_1(v)[i] \leftarrow$ the subtree spanned by $S(v_1)[i-1]$ and r ; $j \leftarrow 1$.

Step 2. While $j < s$, obtain $S_{j+1}(v)$ from $S_j(v)$ and $S(v_{j+1})$ as described in the theorem above; $j \leftarrow j + 1$.

Step 3. $S(v) \leftarrow S_s(v)$.

Finding an r -optimal k -subtree of T may now look like:

OPT (r, k) .

Step 1. (Initialization.) For each terminal vertex $v_t \in T$, $S(v_t) \leftarrow (\{v\}, \emptyset, \dots, \emptyset)$. (The i th empty set indicates that v_t has no i -subtree, $2 \leq i \leq k$.)

Step 2. Process the internal vertices of T in end-order (or any other order in which each vertex follows its sons) constructing $S(v)$ for each internal vertex v of T .

Step 3. Output the k th element of $S(r)$ as an r -optimal k -subtree of T .

Since OPT (r, k) constructs the sequences $S(v)$ for all $v \in V$, we know the v -optimal k -subtrees for all $v \in V$. Thus one can easily obtain an optimal k -subtree simply by taking the best k -subtree out of all the v -optimal k -subtrees. Thus, substituting the following Step 3' in place of Step 3 in OPT (r, k) yields OPT (k) that finds an optimal k -subtree of T .

Step 3. Find a best v -optimal k -subtree over all $v \in V$, and output it as an optimal k -subtree of T .

Note that OPT (r, k) yields the r -optimal i -subtrees for all i , $1 \leq i \leq k$. Hence the optimal i -subtrees for all $1 \leq i \leq k$ can also be easily found.

Complexity. The complexity of the algorithm strongly depends on the time required to compute $W(T)$ for a given tree T . In order to eliminate this factor from

our evaluation, it is taken as $O(1)$. In any other case, our complexity should be multiplied by the appropriate factor in order to obtain the right figure.

Assume that T has n vertices and that merging several disjoint subtrees into one tree takes $O(n)$ time. Let us first analyze the complexity of CONSTRUCT $S(v)$. Obviously Step 2 is the most time-consuming. For a fixed j , obtaining $S_{j+1}(v)$ from $S_j(v)$ and $S(v_j)$ requires $O(k^2n)$ time. This yields time of $O(\deg(v) \times k^2n)$ for the whole of Step 2 and thus for CONSTRUCT $S(v)$ too. Summing the last expression over all $v \in T$ yields a total time of $O(k^2n^2)$ for the entire OPT(r, k) algorithm. The same time bound applies to OPT(k) too.

It should be noted that many weighting functions, including all the examples above, can be computed without complete information on the tree's structure. In most of the cases, only the weights of the optimal subtrees $\underline{T}'_j(i)$ and $T'_{j+1}(i)$, $1 \leq i \leq k$, together with some additional information on the root and the edges connecting it to its sons, is really required. In such cases, one can store in $S(v)$ just the weights of the appropriate optimal subtrees. Additional pointers that would enable us to recover the desired final tree should also be maintained. Both time and space can be reduced in these cases by a factor of n . This yields an $O(k^2n)$ time bound for all the examples above and for many other functions as well.

For example, let us take the Internal Path Length function $W6$. The basic step in the construction of $S(v)$, namely Step 2 of CONSTRUCT $S(v)$, has in this case the form:

Choose $\underline{T}'_{j+1}(i_0)$ as one of the subtrees of the form $\underline{T}'_{j+1}(i, i_0)$ for which $W6(\underline{T}'_j(i)) + W6(T'_{j+1}(i_0 - i)) + (i_0 - i) \times l(v, v_j)$ is optimized. This formula shows that in this case, as in many others, we need the weights of the appropriate optimal trees rather than the trees themselves.

3. Applications to optimization problems in undirected trees. In this section, T is an undirected tree and "subtrees" are undirected subtrees unless otherwise specified. Let v be a vertex of T , and let $T(v)$ denote the rooted tree obtained by hooking T on v . In order to apply the algorithm above to undirected trees, we would first like to extend the notion of a monotonic function to weighting functions that are defined on undirected trees. Fortunately, there is a natural way to do so. If W is initially defined only for undirected trees, it can easily be extended to rooted trees by defining the weight of a rooted tree as that of its underlying tree. W is *ud-monotonic* if its extension to rooted trees is monotonic according to the first definition. Note that if T' is any subtree of T , then rooting T at any vertex v turns T' to a rooted subtree of T , say $T'(u)$, for some u in T' . This observation yields the following undirected modification of OPT(k) for *ud-monotonic* functions:

UD-OPT(k).

Step 1. Choose a vertex v of T .

Step 2. Apply OPT(k) to $T(v)$ yielding a k -optimal rooted subtree $T'(u)$ of $T(v)$.

Step 3. Return T' , the underlying undirected tree of $T'(u)$, as an optimal k -subtree.

It turns out that optimization of nonmonotonic functions can sometimes be carried out by the aid of monotonic "middle" functions. This is the case of the following two functions:

$$\text{the diameter: } W8(T') = \max_{u, v \in V'} d(u, v),$$

$$\text{the radius: } W9(T') = \min_{c \in V'} \max_{v \in V'} d(c, v).$$

Both functions are not *ud*-monotonic, as one can easily verify.

In the following optimal and OPT stand for maximal.

Since UD-OPT(k) calls for OPT(k), and OPT(k) is solved via OPT(r, k), we have only to hook T on an arbitrary root r and show that OPT(r, k) can be solved for $T(r)$. The “middle” function in both cases is the monotonic height function $W7$. Since both cases are quite similar, only $W8$ will be discussed. Again, we restrict ourselves to Step 2 of CONSTRUCT $S(v)$, which is the heart of the algorithm.

As before, let $\underline{T}'_j(i)$ denote an r -optimal i -subtree of \underline{T}_j , and let $T'_{j+1}(i)$ be an r -optimal i -subtree of T_{j+1} , $1 \leq i \leq k$, where optimality is taken with respect to the height. Similarly, let $\underline{T}''_j(i)$ be an r -optimal i -subtree of \underline{T}_j with respect to the diameter. Assuming that the weights of all these trees are already known, and an integer i_0 , $1 \leq i_0 \leq k$, is given, the next tree to be computed, namely $\underline{T}''_{j+1}(i_0)$ is either $\underline{T}\underline{T}_{j+1}(i, i_0)$ for some i , $1 \leq i \leq i_0$, that optimizes $W7(\underline{T}'_j(i)) + W7(T'_{j+1}(i_0 - i)) + l(r, r_{j+1})$, or $\underline{T}''_j(i_0)$, if $W8(\underline{T}''_j(i_0))$ is even better.

The validity proof for this way of choosing $\underline{T}''_{j+1}(i_0)$ is straightforward.

4. Complexity of similar optimization problems in general graphs. A k -subgraph of a given undirected graph is a connected subgraph with k vertices.

In this section we consider the complexity of finding optimal k -subgraphs and k -subtrees of a given graph. For most of the weighting functions mentioned above, the corresponding decision problems are NP-complete even in their simple 0–1 forms. For each problem claimed to be NP-complete, a corresponding NP-complete problem which is reducible to it is listed. For the exact definitions of the source NP-complete problems, see [GJ].

1. *Maximizing W over k -subgraphs.*

W1: The unit length Steiner tree problem.

W2: The maximum clique problem.

W5: The maximum leaf spanning tree problem.

W8, W9: The longest path problem with unit lengths.

W3, W4: Can be solved in polynomial time.

2. *Minimizing W over k -subgraphs.*

W1, W2: The unit length Steiner tree problem.

W7: The longest path problem with unit lengths.

W8: The maximum clique problem.

W3, W4, W9: Have polynomial solutions.

3. *Maximizing W over k -subtrees.*

W1, W7, W8, W9: Same reductions as for k -subgraphs.

W2: The unit length Steiner tree problem.

W3, W4: Have polynomial solutions.

4. *Minimizing W over k -subtrees.*

W1, W2, W7: Same reductions as for k -subgraphs.

W3, W4, W8, W9: Can be solved in polynomial time.

REFERENCES

- [BP] R. I. BECKER AND Y. PERL, *Shifting algorithms for tree partitioning with general weighting functions*, J. Algorithms, to appear.
- [GJ] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability; A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

CANONICAL FORMS AND SOLVABLE SINGULAR SYSTEMS OF DIFFERENTIAL EQUATIONS*

STEPHEN L. CAMPBELL† AND LINDA R. PETZOLD‡

Abstract. In this paper we investigate the relationship between solvability and the existence of canonical forms for the linear system of differential equations $E(t)x'(t) + F(t)x(t) = f(t)$. We show that if E, F are analytic on the interval $[0, T]$, then the differential equation is solvable if and only if it can be put into a certain canonical form. We give examples to show that this is not true if E, F are only differentiable.

1. Introduction. Linear systems of differential equations of the form

$$(1) \quad E(t)x'(t) + F(t)x(t) = f(t)$$

with $E(t)$ a singular $n \times n$ matrix occur in a wide variety of circuit and control applications. Many of these applications are described in some detail in [3], [4], see also [2]. The constant coefficient case is now fairly well understood. However, the theory for the time varying case is still incomplete. This note has several purposes. One is to clear up some of the misconceptions and confusion in the current literature. A second is to give some new results.

We shall say (1) is *analytically solvable* on the interval $[0, T]$ if for any sufficiently smooth (C^n will do) f there exist solutions to (1), and solutions when they exist, are defined on all of $[0, T]$ and are uniquely determined by their value at any $t_0 \in [0, T]$. It is useful to note that a system fails to be analytically solvable if it has any *turning points* in $[0, T]$ (where by turning point we mean a point where the dimension of the manifold of solutions changes), since at these points solutions fail either to exist or to be unique.

If (1) is in the form

$$(2a) \quad y_1' + C(t)y_1 = f_1,$$

$$(2b) \quad N(t)y_2' + y_2 = f_2,$$

where $N(t)$ is nilpotent and lower (or upper) triangular, the system is said to be in *standard canonical form*, SCF [7]. If, in addition, N is constant, then the system is in *strong standard canonical form*, SSCF. The SSCF is the one considered in the work of Petzold and Gear [8], [9], [11].

We shall consider transformations of the form

$$(3) \quad x = Q(t)y, \text{ and left multiplication of the equation by } P(t),$$

where P, Q are invertible on $[0, T]$ and are as smooth as E, F . Clearly, if (1) can be put into SCF, it is analytically solvable. Recently some authors have suggested that analytic solvability implies SSCF except at a finite number of isolated points. In § 2 we shall give a series of examples that show this is not the case unless the matrices $E(t), F(t)$ in (1) are analytic functions of t . In § 3 we prove that if E, F are analytic

* Received by the editors November 9, 1982, and in revised form January 12, 1983.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650. The research of this author was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, under grant AFOSR-81-0052A.

‡ Sandia National Laboratories, Applied Mathematics Division, Livermore, California 94550. The research of this author was supported in part by the U.S. Department of Energy Office of Basic Energy Sciences.

on $[0 T]$, then analytic solvability implies (1) can be transformed by (3) to SCF everywhere on $[0 T]$.

2. Examples.

Example 1. Let $\eta(t)$ be an infinitely differentiable function defined on $[0 T]$ so that $\eta = 0$ on $[2^{-k-1} 2^{-k}]$ for k even and $\eta > 0$ otherwise. Consider the system

$$(4) \quad \begin{bmatrix} 0 & 0 \\ \eta & 0 \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

which has the solution

$$(5a) \quad x_1 = f_1,$$

$$(5b) \quad x_2 = f_2 - \eta f'_1.$$

Clearly (4) is solvable on $[0 1]$ and already in SCF. However any P, Q putting (4) into SSCF will have discontinuities at $\{2^{-k}; k \text{ even}\} \cup \{0\}$.

This example is also interesting from another point of view. Suppose $\eta(t)$ is analytic on $[0 T]$, $\eta(t^*) = 0$, $t^* \in [0 T]$ and nonzero on $[0 T] \setminus \{t^*\}$. Then (4) is transformable to SSCF on any closed subinterval of $[0 T] \setminus \{t^*\}$. The point t^* is not a turning point as we defined it in § 1. However, the system (4) does have, in some sense, a structure change at t^* since the coefficient of x' changes rank and index at t^* .

Example 2. Let

$$(6) \quad N(t) = t^3 \begin{bmatrix} \sin(t^{-1}) \\ \cos(t^{-1}) \end{bmatrix} [\cos(t^{-1}), -\sin(t^{-1})], \quad N(0) = 0.$$

Note that $N'(0) = 0$ and $N^2 \equiv 0$. Thus there is an interval containing zero so that

$$(7) \quad Nx' + x = f$$

is solvable on that interval [5]. However, if $\psi(t)$ is a vector valued function so that $N\psi \equiv 0$, $\psi(0) \neq 0$, then $\psi(t)$ is a multiple of $[\sin(t^{-1}), \cos(t^{-1})]$ and hence is discontinuous at $t = 0$. In particular any P, Q putting (6), (7) into SCF must be discontinuous at zero.

A slight modification of Example 2, along the lines of Example 1, can be used to construct a solvable system such that any P, Q putting the system into SCF would have an infinite number of singularities in a finite interval.

3. Analytic coefficients. The essential problem in Example 2 is that if $\text{rank}(A(t)) \leq r < n$ for all t and A is $n \times n$, then there need not exist any piecewise smooth, nonzero vectors $\psi(t)$ so that $A\psi \equiv 0$. However, it is a not generally known, and nontrivial fact, that such a ψ exists if A is analytic. The version of the result we shall need is the following theorem from [13, p. 335].

THEOREM 1. *If $A(t)$ is real analytic on $[0 T]$ and $r \cong \text{rank}(A(t))$ for all t , then there exists real analytic $P(t), Q(t)$ so that*

$$(8) \quad PAQ = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}$$

and A_1 is $r \times r$.

An infinite dimensional version of Theorem 1 appears in [1, Thm. 2.2]. See also [10].

We are now in a position to prove the main result of this section.

THEOREM 2. *If E, F are analytic on $[0, T]$, and (1) is analytically solvable, then there exist analytic P, Q so that the transformations (3) put (1) into SCF.*

The key to proving Theorem 2 is to first show that while E may have variable rank, solvability forces E to be always singular or always nonsingular.

LEMMA 1. *If (1) is analytically solvable on $[0, T]$, then E is either always singular or always nonsingular on $[0, T]$.*

Proof of Lemma 1. Suppose for purposes of contradiction that $E(t_0)$ is nonsingular and $E(t_1)$ is singular. Then for any f , there are n linearly independent solutions of (1) at t_0 . Let ψ be a vector so that $\psi^T E(t_1) = 0$. Now multiply (1) by ψ^T and evaluate at t_1 to observe that $\psi^T F(t_1)x(t_1) = \psi^T f$. The case $f = \psi$ implies $\psi^T F(t_1) \neq 0$. Hence all solutions for $f = 0$ satisfy $\psi^T F(t_1)x(t_1) = 0$ and are not linearly independent at t_1 which contradicts analytic solvability. \square

Proof of Theorem 2. Suppose Theorem 2 is not true and that E, F are analytic on $[0, T]$, (1) is analytically solvable, but it is not possible to transform to SCF and E, F give a counterexample of minimum possible dimension n . Clearly E is singular and $n > 1$. By Lemma 1, E is always singular on $[0, T]$. Let r be such that $\text{rank } E \leq r < n$ on $[0, T]$. Let P be such that

$$(9) \quad PE = \begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix}$$

where E_1 is $r \times r$ and P is analytic on $[0, T]$. Such a P exists by Theorem 1. Multiplying (1) by P gives the still analytically solvable system

$$(10) \quad \begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} + \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

But $[F_{21} \ F_{22}]$ has full rank $n - r$ on $[0, T]$, since if it did not, there would exist a t_0 , $\psi \neq 0$, such that $\psi^T [F_{21}(t) \ F_{22}(t)] = 0$. But then

$$[0, \psi^T] \begin{bmatrix} f_1(t_0) \\ f_2(t_0) \end{bmatrix} = \psi^T f_2(t_0) = 0,$$

which contradicts the fact that f can be an arbitrary function. Now there exists an invertible analytic Q on $[0, T]$ so that $[F_{21} \ F_{22}]Q = [0 \ G_{22}]$ where G_{22} is invertible. Note this follows as (9) by using

$$Q^T \begin{bmatrix} F_{21}^T & 0 \\ F_{22}^T & 0 \end{bmatrix} = \begin{bmatrix} G_{22}^T & 0 \\ 0 & 0 \end{bmatrix}.$$

Making the change of variable $x = Qy$ turns (10) into

$$(11) \quad \begin{bmatrix} \hat{E}_1 & \hat{E}_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y'_1 \\ y'_2 \end{bmatrix} + \begin{bmatrix} \hat{F}_{11} & \hat{F}_{12} \\ 0 & G_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

where $[\hat{E}_1 \ \hat{E}_2] = [E_1 \ E_2]Q$, $[\hat{F}_{11} \ \hat{F}_{12}] = [F_{11} \ F_{12}]Q + [E_1 \ E_2]Q'$. But (11) is equivalent to solving

$$(12) \quad \hat{E}_1 y'_1 + \hat{F}_{11} y_1 = f_1 - \hat{E}_2 (G_{22}^{-1} f_2)' - \hat{F}_{12} G_{22}^{-1} f_2,$$

or

$$(13) \quad \hat{E}_1 y'_1 + \hat{F}_{11} y_1 = \hat{f}$$

for arbitrary smooth \hat{f} . Thus (13) is also an analytically solvable system. Since it has lower dimension than n , by assumption, there exists analytic $R_1(t), R_2(t)$ that puts

(13) into (upper triangular) SCF. Letting

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} R_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

and multiplying by

$$\begin{bmatrix} R_1 & 0 \\ 0 & I \end{bmatrix}$$

changes (11) to

$$(14) \quad \left[\begin{array}{cc|c} I & 0 & \tilde{E}_{13} \\ 0 & N(t) & \tilde{E}_{23} \\ \hline 0 & 0 & 0 \end{array} \right] \begin{bmatrix} z'_{11} \\ z'_{12} \\ z'_2 \end{bmatrix} + \left[\begin{array}{cc|c} C(t) & 0 & \tilde{F}_{13} \\ 0 & I & \tilde{F}_{23} \\ \hline 0 & 0 & G_{22} \end{array} \right] \begin{bmatrix} z_{11} \\ z_{12} \\ z_2 \end{bmatrix} = \begin{bmatrix} f_{11} \\ f_{12} \\ \tilde{f}_2 \end{bmatrix}.$$

Now multiply this equation by

$$\begin{bmatrix} I & 0 & 0 \\ 0 & [I \ \tilde{F}_{23}]^{-1} \\ 0 & [0 \ Q_{22}] \end{bmatrix}$$

to yield

$$(15) \quad \left[\begin{array}{cc|c} I & 0 & \tilde{E}_{13} \\ 0 & N(t) & \tilde{E}_{23} \\ \hline 0 & 0 & 0 \end{array} \right] \begin{bmatrix} z'_{11} \\ z'_{12} \\ z'_2 \end{bmatrix} + \left[\begin{array}{cc|c} C(t) & 0 & \tilde{F}_{13} \\ 0 & I & 0 \\ \hline 0 & 0 & I \end{array} \right] \begin{bmatrix} z_{11} \\ z_{12} \\ z_2 \end{bmatrix} = \begin{bmatrix} \bar{f}_{11} \\ \bar{f}_{12} \\ \bar{f}_2 \end{bmatrix}.$$

Now let

$$z = \begin{bmatrix} I & 0 & -\tilde{E}_{13} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} w$$

and multiply by

$$\left[\begin{array}{ccc} I & 0 & \tilde{E}'_{13} + C\tilde{E}_{13} - \tilde{F}_{13} \\ 0 & I & 0 \\ 0 & 0 & I \end{array} \right]$$

to get the SCF

$$(16) \quad \left[\begin{array}{cc|c} I & 0 & 0 \\ 0 & N(t) & E_{23} \\ \hline 0 & 0 & 0 \end{array} \right] \begin{bmatrix} w'_1 \\ w'_2 \\ w'_3 \end{bmatrix} + \left[\begin{array}{cc|c} C(t) & 0 & 0 \\ 0 & I & 0 \\ \hline 0 & 0 & I \end{array} \right] \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix}.$$

which contradicts the assumption that (1) could not be put into SCF. \square

4. Comments. An examination of the proof of Theorem 2 shows that the analyticity of E, F was used only in applying Theorem 1 to get analytic P, Q such that (8), (9) hold. Since (8), (9) hold for many matrix functions met in practice, it seems plausible that the nonexistence of the SCF is an exceptional event.

The approach in this paper differs from those of earlier authors, for example, Silverman [12], in that we do not assume $E(t)$ in (1) has constant rank. In particular,

we include systems which cannot be put in the form

$$\begin{aligned}x_1' &= A_{11}(t)x_1 + A_{12}(t)x_2 + f_1(x), \\ 0 &= A_{21}(t)x_1 + A_{22}(t)x_2 + f_2(t)\end{aligned}$$

by transformations of the form (3).

The proof of Theorem 2 also provides an algorithmic procedure for obtaining the SCF. Starting with $Ex' + Fx = f$, compute P, Q as in (9), (10) to get (11). Now take the subsystem $\hat{E}_1 y_1' + \hat{F}_{11} y_1 = \hat{f}_1$ of (11) and repeat the procedure again to get again a system in the form (11). At each step we work with a smaller subsystem. At some step we arrive at a system in the form (11) with either \hat{E}_1 identically zero or always invertible and the procedure terminates.

REFERENCES

- [1] H. BART, M. A. KAASHOEK AND D. C. LAY, *Relative inverses of meromorphic operator functions and associated holomorphic projection functions*, Math. Ann., 218 (1975), pp. 199–210.
- [2] P. BERNHARD, *On singular implicit linear dynamical systems*, SIAM J. Control Optim., 20 (1982), pp. 612–633.
- [3] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, London, 1980.
- [4] ———, *Singular Systems of Differential Equations II*, Pitman, London, 1982.
- [5] ———, *Index two linear time-varying singular systems of differential equations*, this Journal, 4 (1983), pp. 237–243.
- [6] ———, *Higher index time varying singular systems*, Proc. 1982 IFAC Workshop on Singular Perturbations and Robustness of Control Systems, Ohrid, Yugoslavia.
- [7] ———, *One canonical form for higher index linear time varying singular systems*, Circuits, Systems & Signal Processing, to appear.
- [8] C. W. GEAR AND L. R. PETZOLD, *ODE methods for the solution of differential/algebraic systems*, preprint, 1982.
- [9] ———, *Differential/algebraic systems and matrix pencils*, Proc. Conference on Matrix Pencils, Pitea, Sweden, March, 1982.
- [10] H. GINGOLD, *A method of global block diagonalization for matrix valued functions*, SIAM J. Math. Anal., 9 (1978), pp. 1076–1082.
- [11] L. R. PETZOLD, *Differential/algebraic equations are not ODE's*, SIAM J. Sci. Stat. Comp., 3 (1982), pp. 367–384.
- [12] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Aut. Control, AC-14 (1969), pp. 270–276.
- [13] L. M. SILVERMAN AND R. S. BUCY, *Generalizations of a theorem of Dolezal*, Math. Systems Theory, 4 (1970), pp. 334–339.

ON THE POLYNOMIALS OF GRAPHS*

KAI WANG†

Abstract. For a graph G , let $A(G)$ be its adjacency matrix. Let $\varphi_G(x)$ be the characteristic polynomial of G . Let J be a matrix with all entries equal to 1. Let $\psi_G(x) = \varphi_{A(G)-J}(x) - \varphi_G(x)$. In this paper, we show that the characteristic polynomials of the join $G+H$, the complement \bar{G} and the composition $G[H]$ can be expressed in terms of φ_G , φ_H , ψ_G and ψ_H .

AMS 1980 mathematics subject classification. 05C50

1. Introduction. A graph G is a pair $(V(G), E(G))$, where $V(G)$ is a finite nonempty set of elements called vertices, and $E(G)$ is a finite set of distinct unordered pairs of distinct elements of $V(G)$ called edges. Two vertices $u, v \in V(G)$ are said to be adjacent, if $(u, v) \in E(G)$. For each $v \in V(G)$, the number of vertices adjacent to v is called the degree of v . If all the vertices of G have the same degree d , then G is said to be a regular graph of degree d . For a graph G with $V(G) = \{v_1, \dots, v_n\}$, the adjacency matrix $A(G)$ is defined by

$$[A(G)]_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E(G), \\ 0 & \text{otherwise.} \end{cases}$$

The characteristic polynomial $\varphi_{A(G)}(x)$ is called the characteristic polynomial of G and is also denoted by $\varphi_G(x)$. The eigenvalues of $A(G)$ are called the eigenvalues of G which comprises the spectrum of G . As usual, two graphs are said to be cospectral if they have the same spectra. The characteristic polynomials and spectra of graphs have been studied by many authors. We refer to the book [1] of Cvetkovic, Doob and Sachs for the results in this field.

In this paper, we will study the characteristic polynomials of the join $G+H$, the complement \bar{G} and the composition $G[H]$ of graphs G and H . Recall that the join $G+H$ is defined by $V(G+H) = V(G) \cup V(H)$, $E(G+H) = E(G) \cup E(H) \cup \{(u, v) | u \in V(G), v \in V(H)\}$; the complement \bar{G} is defined by $V(\bar{G}) = V(G)$, $E(\bar{G}) = \{(u, v) | (u, v) \notin E(G)\}$; the composition $G[H]$ is defined by $V(G[H]) = V(G) \times V(H)$, $E(G[H]) = \{(u_1, v_1), (u_2, v_2) | (u_1, u_2) \in E(G)\} \cup \{(u, v_1), (u, v_2) | u \in V(G), (v_1, v_2) \in E(H)\}$.

For regular graphs G and H , it is known [1] that φ_{G+H} and $\varphi_{\bar{G}}$ can be expressed in terms of φ_G and φ_H . However, this is not true for arbitrary graphs because the joins and the complements of cospectral graphs are not necessarily cospectral. It turns out that a new polynomial ψ_G has to be introduced so that φ_{G+H} and $\varphi_{\bar{G}}$ can be expressed in terms of φ_G , φ_H , ψ_G and ψ_H . Here ψ_G can be defined by the equation

$$\psi_G(x) = \varphi_{A(G)-J}(x) - \varphi_G(x)$$

where J is a square matrix of order $|V(G)|$ with all entries equal to 1. The main result of this paper is the following:

THEOREM 1.1. *Let G and H be two arbitrary graphs with $m = |V(G)|$. Then*

- (i) $\varphi_{G+H}(x) = \varphi_G(x)\varphi_H(x) - \psi_G(x)\psi_H(x)$,
- (ii) $\varphi_{\bar{G}}(x) = (-1)^m(\varphi_G(-x-1) + \psi_G(-x-1))$,
- (iii) $\varphi_{G[H]}(x) = (\psi_H(x))^m \varphi_G(\varphi_H(x)/\psi_H(x))$.

* Received by the editors October 12, 1981, and in revised form December 9, 1982.

† Department of Mathematics, Wayne State University, Detroit, Michigan 48202.

This paper is organized as follows: In § 2, we will prove our main result. In §§ 3 and 4, we will prove some corollaries. In § 5, we will study the functional properties of $\psi_G(x)$. In §§ 6–8, we will study the generating function for numbers of walks, generalized characteristic polynomials and parametrized characteristic polynomials.

2. A proof of Theorem 1.1. For convenience, we use I (0, respectively) to denote an identity (zero, respectively) matrix of appropriate size and use J to denote a matrix of appropriate size with all entries equal to 1.

For an arbitrary square matrix X , we use X_1, X_2, X_3 to denote the matrices obtained from X by setting

$$[X_1]_{1j} = X_{1j}, \quad [X_2]_{1j} = 0, \quad [X_3]_{1j} = -1$$

and, for $i > 1$,

$$[X_1]_{ij} = [X_2]_{ij} = [X_3]_{ij} = X_{ij} - X_{1j}.$$

It is clear that $|X_1| = |X|$.

We will use the following identity for determinants.

$$(*) \quad \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{i1} + b_{i1} & \cdots & a_{in} + b_{in} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{in} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ b_{i1} & \cdots & b_{in} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}.$$

Let K be an arbitrary graph and let $n = |V(K)|$. For $1 \leq i, j \leq n$, let $u_{ij} = (xI - A(K))_{ij}$. Then

$$\begin{aligned} \varphi_{A(K)-J}(x) &= |xI - A(K) + tJ| \\ &= \begin{vmatrix} u_{11} + t & \cdots & u_{1n} + t \\ \vdots & & \vdots \\ u_{n1} + t & \cdots & u_{nn} + t \end{vmatrix} = \begin{vmatrix} u_{11} + t & \cdots & u_{1n} + t \\ u_{21} - u_{11} & \cdots & u_{2n} - u_{1n} \\ \vdots & & \vdots \\ u_{n1} - u_{11} & \cdots & u_{nn} - u_{1n} \end{vmatrix} \\ &= |(xI - A(K))_1| - t|(xI - A(K))_3| \\ &\qquad\qquad\qquad \text{(by (*) and definitions of } X_1 \text{ and } X_3) \\ &= \varphi_K(x) - t|(xI - A(K))_3|. \end{aligned}$$

It follows from the definition of $\psi_K(x)$ that

$$\psi_K(x) = -|(xI - A(K))_3|$$

and

$$|xI - A(K) + tJ| = \varphi_K(x) + t\psi_K(x).$$

We shall now prove identity *i*. It is not difficult to see that with a suitable labelling of vertices, the adjacency matrix of $G + H$ is given by

$$A(G + H) = \begin{bmatrix} A(G) & J \\ J & A(H) \end{bmatrix}.$$

Then

$$\varphi_{G+H}(x) = |xI - A(G + H)| = \begin{vmatrix} xI - A(G) & -J \\ -J & xI - A(H) \end{vmatrix}.$$

By subtracting the top row of the top (bottom, respectively) block from all other rows in the top (bottom, respectively) block and applying (*) twice, we obtain

$$\begin{aligned}
 & \begin{vmatrix} xI - A(G) & -J \\ -J & xI - A(H) \end{vmatrix} \\
 &= \begin{vmatrix} (xI - A(G))_1 & -1 \cdots -1 \\ -1 \cdots -1 & (xI - A(H))_1 \\ 0 & \end{vmatrix} \\
 &= \begin{vmatrix} (xI - A(G))_1 & -1 \cdots -1 \\ 0 \cdots 0 & (xI - A(H))_1 \\ 0 & \end{vmatrix} + \begin{vmatrix} (xI - A(G))_2 & -1 \cdots -1 \\ -1 \cdots -1 & \end{vmatrix} \\
 & \qquad \qquad \qquad + \begin{vmatrix} (xI - A(G))_1 & 0 \cdots 0 \\ -1 \cdots -1 & (xI - A(H))_2 \\ 0 & \end{vmatrix}.
 \end{aligned}$$

The last determinant vanishes because all entries in the top row of $(xI - A(H))_2$ are zero. The first determinant is equal to

$$|(xI - A(G))_1| |xI - A(H))_1| = |xI - A(G)| |xI - A(H)| = \varphi_G(x)\varphi_H(x).$$

The second determinant can be evaluated as follows. By switching the first row of the top block and the first row of the bottom block, we obtain

$$\begin{aligned}
 & \begin{vmatrix} (xI - A(G))_2 & -1 \cdots -1 \\ -1 \cdots -1 & (xI - A(H))_2 \\ 0 & \end{vmatrix} = - \begin{vmatrix} (xI - A(G))_3 & 0 \\ 0 & (xI - A(H))_3 \end{vmatrix} \\
 &= -|(xI - A(G))_3| |(xI - A(H))_3| \\
 &= -\psi_G(x)\psi_H(x).
 \end{aligned}$$

Now we can put it altogether to obtain

$$\varphi_{G+H}(x) = \varphi_G(x)\varphi_H(x) - \psi_G(x)\psi_H(x).$$

The identity (ii) follows almost directly from the definition of $\psi_G(x)$. Note that $A(\tilde{G}) = J - I - A(G)$. Then

$$\begin{aligned}
 \varphi_{\tilde{G}}(x) &= |xI - A(\tilde{G})| = |(x + 1)I + A(G) - J| \\
 &= (-1)^m |(-x - 1)I - (A(G) - J)| \\
 &= (-1)^m \varphi_{A(G)-J}(-x - 1) \\
 &= (-1)^m (\varphi_G(-x - 1) + \psi_G(-x - 1)).
 \end{aligned}$$

We now prove identity (iii) as follows.

It is not difficult to show that if the vertices of $G[H]$ are labeled lexicographically, then the adjacency matrix $A(G[H])$ is given by

$$A(G[H]) = A(G) \otimes J + I \otimes A(H).$$

Since $A(G)$ is a symmetric matrix, there is a nonsingular matrix M such that $M^{-1}A(G)M = \text{diag}(\lambda_1, \dots, \lambda_m)$ where $\{\lambda_1, \dots, \lambda_m\}$ is the spectrum of G . Then

$$\varphi_G(x) = \prod_{i=1}^m (x - \lambda_i).$$

Then we compute $\varphi_{G[H]}(x)$ as follows.

$$\begin{aligned} \varphi_{G[H]}(x) &= |xI - A(G[H])| = |xI - I \otimes A(H) - A(G) \otimes J| \\ &= |I \otimes (xI - A(H)) - A(G) \otimes J| \\ &= |(M \otimes I)^{-1} (I \otimes (xI - A(H)) - A(G) \otimes J) (M \otimes I)| \\ &= |I \otimes (xI - A(H)) - (M^{-1}A(G)M) \otimes J| \\ &= |I \otimes (xI - A(H)) - \text{diag}(\lambda_1, \dots, \lambda_m) \otimes J| \\ &= |\text{diag}(xI - A(H) - \lambda_1 J, \dots, xI - A(H) - \lambda_m J)| \\ &= \prod_{i=1}^m |xI - A(H) - \lambda_i J| \\ &= \prod_{i=1}^m (\varphi_H(x) - \lambda_i \psi_H(x)) \\ &= (\psi_H(x))^m \prod_{i=1}^m \left(\frac{\varphi_H(x)}{\psi_H(x)} - \lambda_i \right) \\ &= (\psi_H(x))^m \varphi_G \left(\frac{\varphi_H(x)}{\psi_H(x)} \right). \end{aligned}$$

This completes the proof of Theorem 1.1.

3. Corollaries. It is clear that Theorem 1.1 implies the following formula which is due to Cvetkovic [1, p. 57].

COROLLARY 3.1. *Let G and H be two arbitrary graphs with $m = |V(G)|$ and $n = |V(H)|$. Then*

$$\begin{aligned} \varphi_{G+H}(x) &= (-1)^m \varphi_G(x) \varphi_{\bar{H}}(-x-1) + (-1)^n \varphi_{\bar{G}}(-x-1) \varphi_H(x) \\ &\quad + (-1)^{m+n+1} \varphi_{\bar{G}}(-x-1) \varphi_{\bar{H}}(-x-1). \end{aligned}$$

COROLLARY 3.2. *For an arbitrary graph G , let G^+ be the join of G with the one-point graph. Then*

$$\psi_G(x) = x\varphi_G(x) - \varphi_{G^+}(x).$$

Proof. Let K be the one-point graph. Then $A(K) = 0$ and $\varphi_K(x) = x$. This implies that

$$\psi_K(x) = \varphi_{A(K)-J}(x) - \varphi_K(x) = (x+1) - x = 1.$$

It follows from Theorem 1.1(i) that

$$\varphi_{G^+}(x) = \varphi_G(x) \varphi_K(x) - \psi_G(x) \psi_K(x) = x\varphi_G(x) - \psi_G(x).$$

COROLLARY 3.3. *Let G and H be cospectral graphs. Then the following statements are equivalent.*

- (i) $\psi_G(x) = \psi_H(x)$;
- (ii) \bar{G} and \bar{H} are cospectral;
- (iii) G^+ and H^+ are cospectral;
- (iv) $G + K$ and $H + K$ are cospectral for an arbitrary graph K .

4. Regular graphs. In this section, we will apply Theorem 1.1 to regular graphs. We first prove the following lemma.

LEMMA 4.1. *Let A be an $m \times m$ matrix. If A has a constant row sum y , then*

$$|A + tJ| = (y + mt)|A|/y.$$

Proof. For convenience, let $A = [A_1, \dots, A_m]$ where A_i is the i th column vector of A .

$$\begin{aligned} |A + tJ| &= |[A_1 + tJ, \dots, A_m + tJ]| \\ &= \left| \left[\sum_{i=1}^m A_i + mtJ, A_2 + tJ, \dots, A_m + tJ \right] \right| \\ &= (y + mt)|[J, A_2 + tJ, \dots, A_m + tJ]| \\ &= (y + mt)|[J, A_2, \dots, A_m]| \end{aligned}$$

since $\sum_{i=1}^m A_i = yJ$. In particular,

$$|A| = y|[J, A_2, \dots, A_m]|.$$

It follows that

$$|A + tJ| = \frac{y + mt}{y}|A|.$$

PROPOSITION 4.2. *Let K be a regular graph of degree d and let $n = |V(K)|$. Then*

$$\psi_K(x) = n\varphi_K(x)/(x - d).$$

Proof.

$$\psi_K(x) = \varphi_{A(K)-J}(x) - \varphi_K(x) = |xI - A(K) + J| - \varphi_K(x).$$

Since K is regular of degree d , $xI - A(K)$ has a constant row sum $x - d$. It follows from Lemma 4.2 that

$$\begin{aligned} \psi_K(x) &= (x - d + n)\varphi_K(x)/(x - d) - \varphi_K(x) \\ &= n\varphi_K(x)/(x - d). \end{aligned}$$

Now the following results become corollaries to Theorem 1.1.

COROLLARY 4.3 (Sachs [1, p. 56]). *If G is a regular graph of degree d and $|V(G)| = m$, then*

$$\varphi_{\bar{G}}(x) = (-1)^m \frac{x - m + d + 1}{x + d + 1} \varphi_G(-x - 1).$$

COROLLARY 4.4 (Finck and Grohmann [1, p. 57]). *If G and H are regular graphs of degree d and r , respectively, then*

$$\varphi_{G+H}(x) = \left(1 - \frac{mn}{(x - d)(x - r)} \right) \varphi_G(x) \varphi_H(x)$$

where $m = |V(G)|$ and $n = |V(H)|$.

COROLLARY 4.5 (Schwenk [3]). *If H is a regular graph of degree d with $|V(H)| = n$, then*

$$\varphi_{G[H]}(x) = \left(\frac{n\varphi_H(x)}{x-d}\right)^m \varphi_G\left(\frac{x-d}{n}\right).$$

In view of Proposition 4.2, we propose the following conjecture.

CONJECTURE. If $\psi_G(x)$ is a divisor of $\varphi_G(x)$, then G is regular.

5. Further properties of $\psi_G(x)$. In this section, we will state further properties of $\psi_G(x)$ which may be useful in the future studies. The proofs are left as exercises.

PROPOSITION 5.1. *Let G and H be two arbitrary graphs and let $m = |V(G)|$ and $n = |V(H)|$. Then*

- (i) $\psi_{G \cup H}(x) = \varphi_G(x)\psi_H(x) + \psi_G(x)\varphi_H(x)$,
- (ii) $\psi_{\bar{G}}(x) = (-1)^m \psi_G(-x-1)$,
- (iii) $\psi_{G+H}(x) = \varphi_G(x)\psi_H(x) + \psi_G(x)\varphi_H(x) + 2\psi_G(x)\psi_H(x)$.

COROLLARY 5.2. *Define $\tau_G(x) = \varphi_G(x) + \psi_G(x)$. Then,*

$$\tau_{G+H}(x) = \tau_G(x)\tau_H(x).$$

6. Generalized characteristic polynomials. For an arbitrary graph G , let $D(G)$ be a diagonal matrix such that $[D(G)]_{ii}$ is the order of vertex v_i . In [1], it is proposed to study the following generalized characteristic polynomial:

$$\Phi_G(x, y) = |xI + yD(G) - A(G)|.$$

This polynomial includes $\varphi_G(x) = \Phi_G(x, 0)$ and some other polynomials as special cases. In this section, we only record the formula for $\Phi_{\bar{G}}(x, y)$, $\Phi_{G+H}(x, y)$ and leave the derivations to the readers.

THEOREM 6.1. *For an arbitrary graph G , let*

$$\psi_G(x, y) = (x + (m-1)y)\Phi_G(x, y) - \Phi_{G^+}(x-y, y)$$

where $m = |V(G)|$. Then

- (i) $\psi_G(x, 1) = (m/x)\Phi_G(x, 1)$,
- (ii) $\Phi_{\bar{G}}(x, y) = (-1)^m(\Phi_G(-x-1-(m-1)y, y) + \psi_G(-x-1-(m-1)y, y))$,
- (iii) $\Phi_{G+H}(x, y) = \Phi_G(x+ny, y)\Phi_H(x+my, y) - \psi_G(x+ny, y)\psi_H(x+my, y)$ where $n = |V(H)|$.

COROLLARY 6.2 ([1, p. 58]). *Let $C_G(x) = |xI - D(G) + A(G)|$. Then*

- (i) $C_{\bar{G}}(x) = (-1)^m(x/(x-m))C_G(m-x)$,
- (ii) $C_{G+H}(x) = (1 - (mn/(x-m)(x-n)))C_G(x-n)C_G(x-m)$.

We remark that $C_G(x)$ is related to the complexity of G .

7. Generating function for number of walks. For an arbitrary graph G , let $W_G(t) = \sum_{k=0}^{\infty} N_k t^k$ be the generating function for the numbers N_k of walks of length k in G . The function $W_G(t)$ has been studied extensively in [1]. Theorem 1.3 in the Introduction was originally proved by Cvetkovic via the following formula.

THEOREM 7.1 ([1, p. 44]). *Let $W_G(t)$ be the generating function for the numbers of walks in G . Then*

$$W_G(t) = \frac{1}{t} \left[(-1)^m \frac{\varphi_{\bar{G}}(-(t+1)/t)}{\varphi_G(1/t)} - 1 \right].$$

We refer to [1] for the proof of Theorem 7.1 and further results. The following formula for $W_G(t)$ is more compact which follows easily from Theorem 7.1 and Theorem 1.4.

THEOREM 7.2. *Let $W_G(t)$ be the generating functions for the numbers of walks in G . Then*

$$W_G(t) = \frac{1}{t} \frac{\psi_G(1/t)}{\varphi_G(1/t)}.$$

We remark that above formula and the results in § 5 can be used to deduce formulas for $W_{G+H}(t)$ and $W_{\bar{G}}(t)$ [1, p. 209].

8. Parametrized characteristic polynomials. In [4], Johnson and Newman studied the following parametrized characteristic polynomial:

$$P_G(t, x) = |xI - A_t|$$

where $A_t = (t-1)A(G) + J$.

Among other things, they proved the following theorem.

THEOREM 8.1. *If two graphs G and H are cospectral, then $p_G(t, x) = p_H(t, x)$ if and only if \bar{G} and \bar{H} are cospectral.*

This also follows from the following formula which can be proved by the method of this paper.

PROPOSITION 8.2. *With the above notation,*

$$p_G(t, x) = (t-1)^m \left(\varphi_G \left(\frac{x}{t-1} \right) - \frac{1}{t-1} \psi_G \left(\frac{x}{t-1} \right) \right).$$

Independent of [4], we have studied the following parametrized characteristic polynomial:

$$q_G(t, x) = |xI - (1-t)A(G) - tA(\bar{G})|.$$

Note that $q_G(0, x) = \varphi_G(x)$ and $q_G(1, x) = \varphi_{\bar{G}}(x)$.

The proof of following proposition is routine.

PROPOSITION 8.3. *Let G be an arbitrary graph and let $m = |V(G)|$. Then*

$$q_G(t, x) = (1-2t)^m \left(\varphi_G \left(\frac{x+t}{1-2t} \right) - \frac{t}{1-2t} \psi_G \left(\frac{x+t}{1-2t} \right) \right).$$

COROLLARY 8.4. *Let G and H be two graphs. Then $q_G(t, x) = q_H(t, x)$ if and only if G is cospectral to H and \bar{G} is cospectral to \bar{H} .*

Acknowledgments. The author of this paper would like to thank Charles Johnson for the suggestions and encouragement during the preparation of this paper and Tom Leighton for suggesting the direct proof of Theorem 1.1(i) which replaces a derivative-based proof.

REFERENCES

- [1] D. M. CVETKOVIC, M. DOOB AND H. SACHS, *Spectra of Graphs, Theory and Applications*, Academic Press, New York, 1980.
- [2] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.
- [3] A. J. SCHWENK, *Computing the characteristic polynomial of a graph*, in *Graphs and Combinatorics*, Lecture Notes in Mathematics 406, R. Bari and F. Harary, eds., Springer-Verlag, New York, 1974.
- [4] C. JOHNSON AND M. NEWMAN, *A note on cospectral graphs*, *J. Combinatorial Theory*, 28 (1980), pp. 96–100.

GENERALIZED CONTROLLABILITY, (A, B) -INVARIANT SUBSPACES AND PARAMETER INVARIANT CONTROL*

S. P. BHATTACHARYYA†

Abstract. In this paper a linear state space model of a control system with control and disturbance inputs and subject to a class of structured parameter variations is considered. For this problem a geometric condition is derived which guarantees the existence of a state feedback control law which zeros the disturbance transfer function and maintains it zero for the class of parameter variations given. The result involves the notion of generalized controllability and a generalization of the concept of (A, B) -invariant subspaces due to Wonham.

1. Introduction. The geometric theory of linear time invariant state space systems has been extensively developed by Wonham [1] with the invention of the key concept of (A, B) -invariant subspaces. A basic problem solved via this approach is the problem of zeroing the disturbance transfer function via state feedback. This problem characterizes various other control problems, as can be seen from the synthesis problems dealt with in [1]. Our objective in the present paper is to extend these results to the case where the system model is subject to perturbations. It is well known that arbitrary perturbations of the state space model causes any solution to break-down. However, arbitrary perturbations are also frequently unrealistic, and therefore we consider a class of structured parameter variations and give a sufficient condition for solvability of this problem. This sufficient condition involves a generalization of the concept of (A, B) -invariant subspaces and reduces to the solvability condition for the perturbation free case in the absence of perturbations.

2. Problem formulation. Consider the linear system

$$(1a) \quad \dot{x}(t) = A(\alpha)x(t) + B(\beta)u(t) + D(\delta)\xi(t),$$

$$(1b) \quad y(t) = C(\gamma)x(t),$$

denoted by $S(\alpha, \beta, \gamma, \delta)$, with state x , input u , disturbance ξ , output y and

$$(2a) \quad A(\alpha) = A_0 + \alpha_1 A_1 + \cdots + \alpha_p A_p := A_0 + \Delta A(\alpha),$$

$$(2b) \quad B(\beta) = B_0 + \beta_1 B_1 + \cdots + \beta_q B_q := B_0 + \Delta B(\beta),$$

$$(2c) \quad C(\gamma) = C_0 + \gamma_1 C_1 + \cdots + \gamma_r C_r := C_0 + \Delta C(\gamma),$$

$$(2d) \quad D(\delta) = D_0 + \delta_1 D_1 + \cdots + \delta_s D_s := D_0 + \Delta D(\delta),$$

where

$$A_i \in \mathbb{R}^{n \times n}, \quad i = 0, 1, \dots, p, \quad B_i \in \mathbb{R}^{n \times l}, \quad i = 0, 1, \dots, q,$$

$$C_i \in \mathbb{R}^{m \times n}, \quad i = 0, 1, \dots, r, \quad D_i \in \mathbb{R}^{n \times k}, \quad i = 0, 1, \dots, s,$$

$$\alpha := (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^p, \quad \beta := (\beta_1, \dots, \beta_q) \in \mathbb{R}^q,$$

$$\gamma := (\gamma_1, \dots, \gamma_r) \in \mathbb{R}^r, \quad \delta := (\delta_1, \dots, \delta_s) \in \mathbb{R}^s.$$

In (1) (A_0, B_0, C_0, D_0) represents the nominal system model, A_1, \dots, D_s specify the structure of the perturbations and a specific choice of $\alpha, \beta, \gamma, \delta$ determines a

* Received by the editors September 10, 1982, and in revised form December 2, 1982.

† Department of Electrical Engineering, Texas A&M University, College Station, Texas 77843. This research was partially supported by the National Science Foundation under grant ECS 8200852. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26-29, 1982.

specific perturbation $(\Delta A(\alpha), \Delta B(\beta), \Delta C(\gamma), \Delta D(\delta))$. We introduce the class of linear systems generated by the perturbations

$$(3) \quad \mathbf{S} := \bigcup_{\substack{(\alpha, \beta, \gamma, \delta) \\ \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r \times \mathbb{R}^s}} S(\alpha, \beta, \gamma, \delta)$$

and formulate the following problem.

Determine conditions under which there exists a matrix $F \in \mathbb{R}^{l \times n}$ such that the resulting disturbance transfer function

$$(4) \quad C(\gamma)(sI - A(\alpha) - B(\beta)F)^{-1}D(\delta) \equiv 0$$

for every system $S(\alpha, \beta, \gamma, \delta) \in \mathbf{S}$.

The problem is not altered if \mathbf{S} is defined by confining $(\alpha, \beta, \gamma, \delta)$ to a neighborhood of the origin instead of the arbitrary variations in (3); therefore for simplicity arbitrary variations are considered.

In the following sections the notion of generalized controllability due to Carlson and Hill [2] and a generalization of (A, B) -invariant subspaces are developed to provide a sufficient condition for solving the above problem.

3. Generalized controllability and observability. Define the sets

$$\begin{aligned} \mathbf{A} &:= \{A_0, A_1, \dots, A_p\}, & \mathbf{B} &:= \{B_0, B_1, \dots, B_q\}, \\ \mathbf{C} &:= \{C_0, C_1, \dots, C_r\}, & \mathbf{D} &:= \{D_0, D_1, \dots, D_s\}, \end{aligned}$$

and let $\mathcal{T} := \text{Im } T$ and $\text{Ker } T$ denote the range and null space of a matrix T . Following Carlson and Hill [2] we introduce the following definition.

DEFINITION 1. The *generalized controllability subspace* generated by (\mathbf{A}, \mathbf{B}) , denoted $\mathcal{C}(\mathbf{A}, \mathbf{B})$, is the minimal subspace containing $\sum_{j=0}^q \text{Im } B_j$ which is A_i -invariant, $i = 0, 1, \dots, p$. The *generalized unobservable subspace* generated by (\mathbf{C}, \mathbf{A}) , denoted $\theta(\mathbf{C}, \mathbf{A})$, is the maximal subspace of $\bigcap_{j=0}^r \text{Ker } C_j$ which is A_i -invariant, $i = 0, 1, \dots, p$.

In Carlson and Hill [2] the orthogonal complement of $\mathcal{C}(\mathbf{A}, \mathbf{B})$ is taken as the controllability subspace. Our definition corresponds to the more usual notion of controllability employed in the control field [1].

LEMMA 1. *The subspace $\mathcal{C}(\mathbf{A}, \mathbf{B})$ is $A(\alpha)$ -invariant and contains $\text{Im } B(\beta)$ for every $(\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^q$ and is the smallest subspace with this property. The subspace $\theta(\mathbf{C}, \mathbf{A})$ is $A(\alpha)$ -invariant and is contained in $\text{Ker } C(\gamma)$ for every $(\gamma, \alpha) \in \mathbb{R}^r \times \mathbb{R}^p$ and is the largest subspace with this property.*

The simple proof is omitted. Let the matrices

$$\hat{B} := [B_0, B_1, \dots, B_q], \quad \hat{C} := \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_r \end{bmatrix}.$$

Then $\mathcal{C}(\mathbf{A}, \mathbf{B})$ is also the smallest subspace containing $\text{Im } \hat{B}$ which is $A(\alpha)$ -invariant for every $\alpha \in \mathbb{R}^p$ and $\theta(\mathbf{C}, \mathbf{A})$ is the largest subspace of $\text{Ker } \hat{C}$ which is $A(\alpha)$ -invariant for every $\alpha \in \mathbb{R}^p$; these subspaces may also be defined as follows:

$$(5a) \quad \mathcal{C}(\mathbf{A}, \mathbf{B}) = \sum_{t=0,1,\dots} \sum_{0 \leq K_t < n-1} \sum_{0 \leq j_t \leq p} \text{Im } A_{j_1}^{K_1} A_{j_2}^{K_2} \dots A_{j_t}^{K_t} \hat{B}$$

$$(5b) \quad \theta(\mathbf{C}, \mathbf{A}) = \bigcap_{t=0,1,\dots} \bigcap_{0 \leq K_t \leq n-1} \bigcap_{0 \leq j_t \leq p} \text{Ker } \hat{C} A_{j_1}^{K_1} \dots A_{j_t}^{K_t}$$

$$(A_{j_1}^{K_1} \dots A_{j_t}^{K_t} := I \text{ for } t = 0).$$

4. Generalized (A, B)-invariant subspaces. In [1], (A, B) -invariant subspaces are defined. We generalize this as follows.

DEFINITION 2. A subspace $\mathcal{V} \subset \mathbb{R}^n$ is a *generalized (A, B)-invariant subspace* iff there exists real F such that

$$(6) \quad (A(\alpha) + B(\beta)F)\mathcal{V} \subset \mathcal{V} \quad \text{for all } (\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^q.$$

The motivation for introducing this generalization is that if $\mathbf{F}(A(\alpha), B(\beta), \mathcal{V})$ denotes the family of real matrices F satisfying $(A(\alpha) + B(\beta)F)\mathcal{V} \subset \mathcal{V}$ for all $(\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^q$, then \mathcal{V} is a generalized (A, B) invariant subspace iff

$$\bigcap_{(\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^q} \mathbf{F}(A(\alpha), B(\beta), \mathcal{V}) \neq \emptyset.$$

Now let

$$(7) \quad \mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C}) := \{\mathcal{V} \mid \mathcal{V} \subset \mathbb{R}^n, F \in \mathbb{R}^{l \times n} (A(\alpha) + B(\beta)F)\mathcal{V} \subset \mathcal{V} \subset \text{Ker } C(\gamma)\}$$

for all $(\alpha, \beta, \gamma) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r$.

THEOREM 1. For every $\mathbf{A}, \mathbf{B}, \mathbf{C}$, the family $\mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ has a unique maximal element.

Proof. We are given $A_i, i = 0, 1, \dots, p, B_i, i = 0, 1, \dots, q, C_i, i = 0, 1, \dots, r$. For each subspace \mathcal{V} of \mathbb{R}^n , let

$$B_i^{-1}\mathcal{V} := \{v \in \mathbb{R}^l \mid B_i v \in \mathcal{V}\}, \quad i = 1, \dots, q,$$

let $\mathcal{R} := \bigcap_{i=1}^q B_i^{-1}\mathcal{V}$, and let $Q \in \mathbb{R}^{l \times l}$ project \mathbb{R}^l onto \mathcal{R} along \mathcal{R}^\perp . There exists an $F \in \mathbb{R}^{l \times n}$ such that

$$(8a) \quad (A_0 + B_0 F)\mathcal{V} \subset \mathcal{V}, \quad B_i F \mathcal{V} \subset \mathcal{V}, \quad i = 1, \dots, q,$$

iff

$$(8b) \quad A_0 \mathcal{V} \subset \text{Im } B_0 Q + \mathcal{V}.$$

To see this, if (8a) holds, then $F\mathcal{V} \subset \mathcal{R} = \bigcap_{i=1}^q B_i^{-1}\mathcal{V}$, so that for $v \in \mathcal{V}$, $QFv = Fv$, and

$$(A_0 + B_0 F)v = (A_0 + B_0 QF)v \in \mathcal{V},$$

and

$$A_0 v \in \text{Im } B_0 Q + \mathcal{V}.$$

Conversely, if (8b) holds, let v_1, \dots, v_k be a basis of \mathcal{V} and extend to a basis $v_1, \dots, v_k, \dots, v_n$ of \mathbb{R}^n . By assumption, there exist $r_i \in \mathcal{R}, w_i \in \mathcal{V}, i = 1, \dots, k$ so that

$$A_0 v_i = B_0 r_i + w_i, \quad i = 1, \dots, k.$$

Define $F \in \mathbb{R}^{l \times n}$ so that

$$Fv_i = \begin{cases} -r_i, & i = 1, \dots, k, \\ 0, & i = k + 1, \dots, n. \end{cases}$$

Then

$$(A_0 + B_0 F)v_i = A_0 v_i + B_0 Fv_i = B_0 r_i + w_i - B_0 r_i, \quad i = 1, \dots, k$$

and $F\mathcal{V} \subset \mathcal{R}$, i.e., $B_i F \mathcal{V} \subset \mathcal{V}, i = 1, \dots, q$.

Now subspace \mathcal{V} of \mathbb{R}^n is in $\mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ iff, for some $F \in \mathbb{R}^{l \times n}$ such that

$$(9a) \quad (\mathbf{A}_0 + \mathbf{B}_0 F) \mathcal{V} \subset \mathcal{V} \subset \text{Ker } \hat{\mathbf{C}},$$

$$(9b) \quad \mathbf{B}_i F \mathcal{V} \subset \mathcal{V}, \quad i = 1, \dots, q,$$

$$(9c) \quad \mathbf{A}_i \mathcal{V} \subset \mathcal{V}, \quad i = 1, \dots, q,$$

and by what we have just shown, iff

$$(10) \quad \mathbf{A}_0 \mathcal{V} \subset \text{Im } \mathbf{B}_0 \mathbf{Q} + \mathcal{V},$$

$$(11) \quad \mathcal{V} \subset \text{Ker } \hat{\mathbf{C}},$$

$$(12) \quad \mathbf{A}_i \mathcal{V} \subset \mathcal{V}, \quad i = 1, \dots, p.$$

It is now easy to see that $\mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is closed under subspace addition (although $\mathcal{R} = \mathcal{R}_{\mathcal{V}}$ and $\mathbf{Q} = \mathbf{Q}_{\mathcal{V}}$ depend on \mathcal{V} we have $\mathcal{R}_{\mathcal{V}} + \mathcal{R}_{\mathcal{W}} \subset \mathcal{R}_{\mathcal{V} + \mathcal{W}}$, so that $\text{Im } \mathbf{B}_0 \mathbf{Q}_{\mathcal{V}} + \text{Im } \mathbf{B}_0 \mathbf{Q}_{\mathcal{W}} \subset \text{Im } \mathbf{B}_0 \mathbf{Q}_{\mathcal{V} + \mathcal{W}}$) and by [1, Lemma 4.4] $\mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ has a unique maximal element. \square

Let the maximal element of $\mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ be denoted by $\mathcal{V}^*(\mathbf{A}, \mathbf{B}, \mathbf{C})$. This subspace may be calculated from the following algorithm which extends the procedure given in [1] for calculating maximal (\mathbf{A}, \mathbf{B}) -invariant subspaces.

1. $\mathcal{V}_0 := \text{Ker } \hat{\mathbf{C}}$
2. $\mathcal{R}_K := \bigcap_{i=1}^q \mathbf{B}_i^{-1} \mathcal{V}_K$
3. $\mathbf{Q}_K :=$ projection of \mathbb{R}^l on \mathcal{R}_K along \mathcal{R}_K
4. $\mathbf{B}_{0K}^* := \mathbf{B}_0 \mathbf{Q}_K$
5. $\mathcal{V}_{K+1} = \mathcal{V}_K \cap \mathbf{A}_0^{-1} (\text{Im } \mathbf{B}_{0K}^* + \mathcal{V}_K) \cap \mathbf{A}_1^{-1} \mathcal{V}_K \cdots \cap \mathbf{A}_p^{-1} \mathcal{V}_K$
6. $\mathcal{V}^*(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathcal{V}_n$.

The proof of convergence of this algorithm follows from the fact that the \mathcal{V}_K are nonincreasing and $\mathcal{V}_{K+1} = \mathcal{V}_K$ implies that $\mathcal{V}_K \in \mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$. That the subspace resulting from the algorithm is $\mathcal{V}^*(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is easily seen.

5. Parameter invariant disturbance rejection. The following result on the problem formulated in § 2 can now be stated. For this let

$$\hat{\mathbf{D}} := [\mathbf{D}_0, \mathbf{D}_1, \dots, \mathbf{D}_s].$$

THEOREM 2. *There exists $F \in \mathbb{R}^{l \times n}$ for which*

$$(13) \quad \mathbf{C}(\boldsymbol{\gamma})(s\mathbf{I} - \mathbf{A}(\boldsymbol{\alpha}) - \mathbf{B}(\boldsymbol{\beta})F)^{-1} \mathbf{D}(\boldsymbol{\delta}) = 0 \quad \text{for all } \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathbf{S}$$

if

$$(14) \quad \text{Im } \hat{\mathbf{D}} \subset \mathcal{V}^*(\mathbf{A}, \mathbf{B}, \mathbf{C}).$$

The proof depends on the following lemma.

LEMMA 2.

$$(15) \quad \mathbf{C}(\boldsymbol{\gamma})(s\mathbf{I} - \mathbf{A}(\boldsymbol{\alpha}))^{-1} \mathbf{D}(\boldsymbol{\delta}) = 0 \quad \text{for all } (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathbb{R}^p \times \mathbb{R}^l \times \mathbb{R}^s$$

if

$$(16a) \quad \text{Im } \hat{\mathbf{D}} \subset \theta(\mathbf{C}, \mathbf{A})$$

or equivalently

$$(16b) \quad \mathbf{C}(\mathbf{A}, \mathbf{D}) \subset \text{Ker } \hat{\mathbf{C}}.$$

Proof. The condition (15) can be rewritten

$$C(\boldsymbol{\gamma})A^i(\boldsymbol{\alpha})D(\boldsymbol{\delta})=0 \quad \text{for all } (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathbb{R}^p \times \mathbb{R}^r \times \mathbb{R}^s$$

and this is equivalent to

$$(17) \quad \hat{C}A^i(\boldsymbol{\alpha})\hat{D}=0 \quad \text{for all } \boldsymbol{\alpha} \in \mathbb{R}^p.$$

Clearly (16a) or (16b) implies (17).

Proof of Theorem 2. By the definition of $\mathcal{V}(\mathbf{A}, \mathbf{B}, \mathbf{C})$, there exists an $F \in \mathbb{R}^{l \times n}$ for which (9) holds for $\mathcal{V} = \mathcal{V}^*(\mathbf{A}, \mathbf{B}, \mathbf{C})$. Let

$$\mathbf{A}_F := \{A_0 + B_0F, A_1, \dots, A_p, B_1F, \dots, B_qF\}.$$

By our definition of generalized unobservable subspace, we have

$$\mathcal{V}^*(\mathbf{A}, \mathbf{B}, \mathbf{C}) \subset \theta(\mathbf{C}, \mathbf{A}_F).$$

Suppose (14) holds. Then for the F given above,

$$(18) \quad \text{Im } \hat{D} \subset \theta(\mathbf{C}, \mathbf{A}_F).$$

By Lemma 2, with \mathbf{A} replaced by \mathbf{A}_F , we see that (18) implies (13). \square

6. Concluding remarks. Generalized controllability and generalized (\mathbf{A}, \mathbf{B}) -invariant subspaces are closely related to the problem of zeroing the transfer function under parameter variations. However we note that Theorem 2 does not give a necessary condition for this problem. The reason for this may be seen as follows. Let

$$\mathcal{V}(\boldsymbol{\alpha}) := \bigcap_{i=0}^{n-1} \text{Ker } \hat{C}A^i(\boldsymbol{\alpha}).$$

Then (17) and hence (15) is equivalent to

$$\text{Im } \hat{D} \subset \bigcap_{\boldsymbol{\alpha} \in \mathbb{R}^p} \mathcal{V}(\boldsymbol{\alpha}) := \mathcal{W},$$

but it is not true in general that

$$A(\boldsymbol{\alpha})\mathcal{W} \subset \mathcal{W} \quad \text{for all } \boldsymbol{\alpha} \in \mathbb{R}^p.$$

The problem of obtaining necessary and sufficient conditions for this problem therefore remains open, and it is hoped that the results given here will prove helpful in that effort.

7. Acknowledgment. The author gratefully acknowledges many improvements in the paper resulting from discussions with Professor David Carlson.

REFERENCES

- [1] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.
- [2] D. CARLSON AND R. HILL, *Generalized controllability and inertia theory*, *Linear Alg. and Applic.*, 15 (1976), pp. 177–187.
- [3] S. P. BHATTACHARYYA, *Parameter invariant observers*, *Internat. J. Control*, 32 (1980), pp. 1127–1132.

TENSORS AND GRAPHS*

RUSSELL MERRIS† AND WILLIAM WATKINS‡

Dedicated to Emilie Haynsworth

Abstract. In § 1, we discuss symmetry classes of tensors and their dimensions in the context of the representation theory of the general linear group. The main result is a formula for the Marcus–Chollet index. In § 2, we observe that the free vector space generated by the nonisomorphic graphs on p vertices is a symmetry class of tensors. Thus, we are able to make use of the dimension formulas in § 1 to enumerate the nonisomorphic graphs. For example: let $m = p(p - 1)/2$. Denote by ξ_a the irreducible character of S_m corresponding to the partition $(m - q, q)$, $0 \leq q \leq m/2$. Then the number of nonisomorphic, unlabelled graphs on p vertices is $\sum_{q=0}^{\lfloor m/2 \rfloor} (m - 2q + 1)(1, \xi_a)_p$, where $(1, \xi_a)_p$ is the number of occurrences of the principal character in the restriction of ξ_a to the pair group $S_p^{(2)}$ (i.e., the line group of the complete graph on p vertices). In addition, we use the corresponding representation of the general linear group to catalog distance inventories between graphs. Section 3 contains extensions to multigraphs, and § 4 some combinatorial lemmas and proofs.

AMS (1979) subject classification. Primary, 15A69; secondary, 05C25, 05C30, 20G05.

1. Symmetry classes of tensors. Let V be a vector space of dimension n over a field F of characteristic zero. For m an integer at least 2, denote by $V^{m \otimes}$ the m th tensor power of V and write $v_1 \otimes v_2 \otimes \cdots \otimes v_m$ for the tensor product of the indicated vectors. Then to each permutation $\sigma \in S_m$, there corresponds a (unique) linear operator $P(\sigma)$ on $V^{m \otimes}$ such that $P(\sigma^{-1})(v_1 \otimes v_2 \otimes \cdots \otimes v_m) = v_{\sigma(1)} \otimes v_{\sigma(2)} \otimes \cdots \otimes v_{\sigma(m)}$, for all $v_1, v_2, \dots, v_m \in V$. If G is a subgroup of S_m , and χ is an absolutely irreducible F -valued character of G , define

$$\theta(G, \chi) = \frac{\chi(\text{id})}{o(G)} \sum_{\sigma \in G} \chi(\sigma) P(\sigma),$$

where $o(G)$ is the cardinality of G and $\chi(\text{id})$ is the degree of χ . Then $\theta(G, \chi)$ is a projection onto its range $V_\chi(G)$. The subspace $V_\chi(G)$ is one variety of what has come to be known as a symmetry class of tensors.

Let $L(V)$ be the set of linear operators on V . Then each $T \in L(V)$ induces a natural linear operator $T^{m \otimes} \in L(V^{m \otimes})$ such that

$$(1) \quad T^{m \otimes}(v_1 \otimes v_2 \otimes \cdots \otimes v_m) = Tv_1 \otimes Tv_2 \otimes \cdots \otimes Tv_m,$$

for all $v_1, v_2, \dots, v_m \in V$. Since $T^{m \otimes}$ evidently commutes with $\theta(G, \chi)$, it follows that $V_\chi(G)$ is an invariant subspace of $T^{m \otimes}$, for all $T \in L(V)$. Let $K_\chi^G(T)$ denote the restriction of $T^{m \otimes}$ to $V_\chi(G)$. When $G = S_m$, write $K_\chi^m(T)$. If, for example, $G = S_m$ and $\chi = \epsilon$, the alternating character, then $V_\epsilon(S_m) = \bigwedge^m V$, the m th exterior power of V , and $K_\epsilon^m(T)$ is the m th compound. If $G = \{\text{id}\}$ and $\chi = 1$, then $V_\chi(G) = V^{m \otimes}$ and $K_\chi^G(T) = T^{m \otimes}$. In general, it follows from (1) that

$$(2) \quad K_\chi^G(T_1)K_\chi^G(T_2) = K_\chi^G(T_1T_2),$$

for all $T_1, T_2 \in L(V)$. In particular, K_χ^G is a representation of the full linear group $GL(n, F) = \{T \in L(V): T \text{ is invertible}\}$, provided $V_\chi(G) \neq \{0\}$. M. Marcus and J. Chollet

* Received by the editors July 9, 1982, and in revised form January 31, 1983.

† Department of Mathematics, California State University, Hayward, California 94542. The research of this author was supported in part by the National Science Foundation under grant MCS 77-28437.

‡ California State University, Northridge, California 91330.

[20] defined the *index* of $V_\chi(G)$ to be the largest value of $n = \dim V$ such that $V_\chi(G) = \{0\}$. In Theorem 1 (below), we give a means of determining the index.

Consider, for the moment, the special case $G = S_m$. (Every irreducible F -representation of S_m is absolutely irreducible.) Let $\sigma \rightarrow Q(\sigma) = (q_{ij}(\sigma))$, $\sigma \in S_m$, be an irreducible F -representation of S_m which affords χ . Define

$$(3) \quad \theta_i(S_m, Q) = \frac{\chi(\text{id})}{m!} \sum_{\sigma \in S_m} q_{ii}(\sigma) P(\sigma).$$

Then, by the Schur relations for the coordinate functions q_{ij} ([14] or [32, p. 16]), $\{\theta_i(S_m, Q) : 1 \leq i \leq \chi(\text{id})\}$ is a set of pairwise annihilating idempotents which sum to $\theta(S_m, \chi)$. Since $\theta_i(S_m, Q)$ commutes with $T^{m \otimes}$, $T \in L(V)$, it follows that $V^i_Q(S_m)$, the range of $\theta_i(S_m, Q)$, is an invariant subspace of $K_\chi^m(T)$. In particular, the representation $T \rightarrow K_\chi^m(T)$, $T \in GL(n, F)$, reduces into $\chi(\text{id})$ pieces. It is proved in [29, Lemma 1] that each of these pieces is equivalent and in [21, Chap. VII] that each of them is irreducible. (The Young symmetrizers are of the form (3).) Denote by $J_{Q,i}^m(T)$ the restriction of $K_\chi^m(T)$ (i.e., of $T^{m \otimes}$) to $V^i_Q(S_m)$. Then, in summary, the representation $T \rightarrow T^{m \otimes}$, $T \in GL(n, F)$, is equivalent to the direct sum of the representations K_χ^m , as χ ranges over the irreducible characters of S_m . Moreover, each of the representations K_χ^m is equivalent to the direct sum of $J_{Q,i}^m$, $1 \leq i \leq \chi(\text{id})$. Finally, $J_{Q,i}^m$ is an irreducible representation of $GL(n, F)$ and, for a fixed χ , all of the representations $J_{Q,i}^m$, $1 \leq i \leq \chi(\text{id})$, are equivalent. (In particular, $\dim V^i_Q(S_m) = \dim V^j_Q(S_m)$, for $i, j = 1, 2, \dots, \chi(\text{id})$.) In an effort to keep the notation under control, we will use J_χ^m to denote any one of the equivalent representations $J_{Q,i}^m$. Observe that the degree of the representation J_χ^m is the dimension of $V^i_Q(S_m)$ for any/every $i = 1, 2, \dots, \chi(\text{id})$ and for any irreducible representation Q of S_m which affords χ . We will use the notation $\dim(n, m, \chi)$ to denote the degree of J_χ^m . Thus, in particular,

$$\dim V_\chi(S_m) = \chi(\text{id}) \dim(n, m, \chi),$$

where $n = \dim V$.

We now return to the general case. Let G be a subgroup of S_m . Suppose χ is an absolutely irreducible F -valued character of G . The same analysis shows (possibly over some extension of F) that $T \rightarrow K_\chi^G(T)$ is equivalent to a direct sum of $\chi(\text{id})$ equivalent pieces, $T \rightarrow J_\chi^G(T)$. In general, however, J_χ^G is further reducible. Indeed (see [21] and [29]), J_χ^G is equivalent to a direct sum of components, each of the form J_λ^m for some irreducible character λ of S_m with $\dim(n, m, \lambda) > 0$. Moreover, the multiplicity of such a J_λ^m in J_χ^G is equal to the multiplicity of χ in the restriction of λ to G , i.e., to $(\chi, \lambda)_G$. Since K_χ^G is equivalent to a direct sum of $\chi(\text{id})$ copies of J_χ^G , we obtain the (for our purposes) fundamental formula

$$(4) \quad \dim V_\chi(G) = \sum_{\lambda \in \mathcal{S}_m} \chi(\text{id})(\chi, \lambda)_G \dim(n, m, \lambda),$$

where \mathcal{S}_m is the set of irreducible characters of S_m . (See [22].)

In order to discuss $\dim(n, m, \lambda)$, $\lambda \in \mathcal{S}_m$, we need the notion of a (proper) partition of m . Each irreducible character of S_m corresponds to a partition $(\lambda) = (\lambda_1, \lambda_2, \dots)$, a sequence of nonnegative integers such that $\lambda_1 \geq \lambda_2 \geq \dots$ and $\lambda_1 + \lambda_2 + \dots = m$. The nonzero λ_i are called the parts of (λ) ; the number of parts is the *length* of (λ) , denoted $r = r(\lambda)$. We will sometimes write $(\lambda) = (\lambda_1, \lambda_2, \dots, \lambda_r)$. It is useful to draw pictures of partitions. These pictures are called Young diagrams or Ferrers–Sylvester graphs. The picture for (λ) consists of $r(\lambda)$ left justified rows of boxes. The number of boxes

in the i th row is λ_i . Thus, for example, the diagram corresponding to $(\lambda) = (3, 2^2, 1)$, an abbreviation for $(3, 2, 2, 1)$, is shown in Fig. 1.

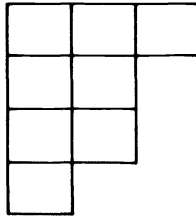


FIG. 1

Let λ be an irreducible character of S_m corresponding to the partition $(\lambda) = (\lambda_1, \lambda_2, \dots, \lambda_r)$. Suppose $\{A(\sigma) = (a_{ij}(\sigma)): \sigma \in S_m\}$ is an F -representation of S_m which affords λ . Since $\theta_i(S_m, A)$ is idempotent,

$$\begin{aligned} \dim(n, m, \lambda) &= \dim V_A^i(S_m) = \text{trace } \theta_i(S_m, A) \\ &= \frac{\lambda(\text{id})}{m!} \sum_{\sigma \in S_m} a_{ii}(\sigma) \text{trace } P(\sigma) \\ &= \frac{\lambda(\text{id})}{m!} \sum_{\sigma \in S_m} a_{ii}(\sigma) \rho(\sigma), \end{aligned}$$

where ρ is the character afforded by the representation $\{P(\sigma): \sigma \in S_m\}$. We may use the Schur relations to write

$$(5) \quad \dim(n, m, \lambda) = \frac{1}{m!} \sum_{\sigma \in S_m} \lambda(\sigma) \rho(\sigma) = (\lambda, \rho)_{S_m},$$

the number of occurrences of λ in the reduction of ρ . (In S_m , σ and σ^{-1} are conjugate.)

Since $P(\sigma) \in L(V^{m \otimes})$, ρ depends on $n = \dim V$. Indeed, it is well known (see, e.g., [19, p. 75]) that $\rho(\sigma) = n^{c(\sigma)}$, where $c(\sigma)$ is the number of cycles, including cycles of length 1, in the disjoint cycle factorization of σ . Thus (5) may be rewritten

$$(6) \quad \dim(n, m, \lambda) = \frac{1}{m!} \sum_{\sigma \in S_m} \lambda(\sigma) n^{c(\sigma)}.$$

Taking advantage of the corresponding partition $(\lambda) = (\lambda_1, \lambda_2, \dots, \lambda_r)$, one may give more explicit versions of (6). The following ‘‘Frame–Robinson–Thrall type’’ formula is given in [16, p. 326], [17], [18, p. 189], and [31]:

$$(7) \quad \dim(n, m, \lambda) = \frac{\prod_{t=1}^r \prod_{j=1}^{\lambda_t} (n - t + j)}{\prod_{t=1}^r \prod_{j=1}^{\lambda_t} h_{tj}},$$

where h_{tj} is the ‘‘hook-length’’ of the (t, j) box in the Young frame for (λ) . That is, $h_{tj} = 1 + d + e$ where $d = \lambda_t - j$ is the number of boxes in row t to the right of the (t, j) th box, and e is the number of boxes in column j below the (t, j) th box. For example, if $(\lambda) = (3, 2^2, 1)$, the number appearing in each box of Fig. 2 is the hook-length of that box.

It is possible to give a similar interpretation to the numerator of (7). If $(\lambda) = (3, 2^2, 1)$, notice that the numerator is the product of the integers in Fig. 3.

6	4	1
4	2	
3	1	
1		

FIG. 2

n	$n+1$	$n+2$
$n-1$	n	
$n-2$	$n-1$	
$n-3$		

FIG. 3

The Frame–Robinson–Thrall hook-length formula for the degree of λ is

$$(8) \quad \lambda(\text{id}) = \frac{m!}{\prod_{i=1}^r \prod_{j=1}^{\lambda_i} h_{ij}}$$

If we define

$$(9) \quad p_\lambda(x) = \prod_{i=1}^r \prod_{j=1}^{\lambda_i} (x - i + j),$$

then (7) becomes

$$(10) \quad \dim(n, m, \lambda) = \frac{\lambda(\text{id})p_\lambda(n)}{m!}.$$

One easy but important consequence of (7)/(10) is that

$$(11) \quad \dim(n, m, \lambda) = 0 \quad \text{if and only if} \quad n < r(\lambda).$$

When $n \geq r(\lambda)$, (7) may be modified to give the following ‘‘Frobenius type’’ formula (see [1, p. 201], [12, p. 387] and [30, p. 129]):

$$(12) \quad \dim(n, m, \lambda) = \frac{\prod_{t=2}^n \prod_{j=1}^{t-1} (\lambda_j - \lambda_t + t - j)}{\prod_{t=2}^n (t-1)!},$$

where $\lambda_t = 0$ for $t > r(\lambda)$.

We are now ready to state and prove the main result of this section. Returning to (2), we recall that K_x^G is a representation of $GL(n, F)$ only if $\dim V_x(G) \neq 0$. As we remarked, Marcus and Chollet have defined the *index* of $V_x(G)$ to be the largest value of $n = \dim V$ such that $V_x(G) = \{0\}$.

THEOREM 1. *Let G be a subgroup of S_m . Let χ be an absolutely irreducible F -valued character of G . Then the index of $V_x(G)$ is $\min \{r(\lambda) - 1 : \lambda \in \mathcal{I}_m \text{ and } (\chi, \lambda)_G \neq 0\}$.*

Proof. Immediate from (4) and (11).

The problem of explicitly listing all groups, characters, and values of n corresponding to the degeneracy $V_x(G) = \{0\}$ has received extensive attention in the recent literature. (See [3]–[6], [9] and [35].) Currently it is known exactly which groups and characters lead to this kind of degeneracy in all cases for which $m \leq 4n$.

Example 1. Let $m = 4$. Suppose G is the subgroup of S_4 generated by (12) and (13)(24). (Then G is the dihedral group D_4 .) Let χ_1 through χ_4 be the linear (i.e., degree 1) characters of G defined by Table 1, and let χ_5 be the irreducible character of G of degree 2.

TABLE 1

	(12)	(13)(24)
χ_1	1	1
χ_2	1	-1
χ_3	-1	-1
χ_4	-1	1

Then

$$(4)|_G = \chi_1, \quad (3, 1)|_G = \chi_2 + \chi_5, \quad (2^2)|_G = \chi_1 + \chi_4,$$

$$(2, 1^2)|_G = \chi_3 + \chi_5, \quad (1^4)|_G = \chi_4,$$

where, e.g., $(2, 1^2)|_G$ is to be read as the restriction to G of the character of S_4 corresponding to the partition $(2, 1, 1)$. By Theorem 1, the index of $V_{\chi_1}(G)$ is $\min \{r(\lambda) - 1 : (\lambda) = (4) \text{ or } (\lambda) = (2^2)\}$. Thus, the index of $V_{\chi_1}(G)$ is zero. Similarly, $\text{index } V_{\chi_2}(G) = \text{index } V_{\chi_4}(G) = \text{index } V_{\chi_5}(G) = 1$, and $\text{index } V_{\chi_3}(G) = 2$.

Explicit formulas for $\dim V_{\chi}(G)$ may be obtained by substituting any one of (5), (6), (7), (10) or (12) into (4). The best known of these arises by plugging (6) into (4) and using the orthogonality relations of the second kind. (One may achieve the same result more directly simply by taking the trace of $\theta(G, \chi)$.) In any case,

$$(13) \quad \dim V_{\chi}(G) = \frac{\chi(\text{id})}{o(G)} \sum_{\sigma \in G} \chi(\sigma) n^{c(\sigma)}.$$

It was probably S. G. Williamson ([39] and [40]) who first explicitly noticed the connection between dimensions of symmetry classes of tensors and the Pólya–Redfield theorem of combinatorial enumeration. (Also see [26].) As pointed out in [23], one may use (10) in place of (6) to obtain an equivalent theorem. Section 2 consists of a detailed exploration of this idea for a particular example. Part of the motivation for writing § 2 came from a recent *Monthly* article by H. S. Wilf [38].

2. The symmetry class of graphs. The purpose of this section is to apply some of the preceding material to certain aspects of graph theory. Although our observations will be made for “1-graphs”, it should be noted that they apply to hypergraphs.

Let \mathcal{G} be a labelled graph on p vertices. We may describe \mathcal{G} by means of a coloring of the $m = \binom{p}{2}$ edges of the complete (labelled) graph K_p . An edge of K_p is colored 1 if it is an edge of \mathcal{G} and 0 if it is not. Thus, there is a one-to-one correspondence between labelled graphs and $\Gamma_{m,2}$, the set of all functions $\gamma: \{1, 2, \dots, m\} \rightarrow \{0, 1\}$.

Suppose W is a vector space over F of dimension 2. Let $\{e_0, e_1\}$ be a basis of W . Then $\{e_{\gamma}^{\otimes} = e_{\gamma(1)} \otimes e_{\gamma(2)} \otimes \dots \otimes e_{\gamma(m)} : \gamma \in \Gamma_{m,2}\}$ is a basis of $W^{m \otimes}$. Thus, we may view $W^{m \otimes}$ as the (free) vector space spanned by the labelled graphs on p vertices.

Expressed as permutations of the vertices, the group of automorphisms of K_p is S_p . Expressed as a group of permutations of the edges, it is a subgroup of S_m , $m = p(p - 1)/2$, called the *pair* group of S_p (or the *line* group of K_p). This group is commonly denoted $S_p^{(2)}$. Consider the symmetry class of $W^{m \otimes}$ corresponding to $G = S_p^{(2)}$ and the principal (identically 1) character. For vectors $w_1, w_2, \dots, w_m \in W$, denote $\theta(S_p^{(2)}, 1)w_1 \otimes w_2 \otimes \dots \otimes w_m$ by $w_1 * w_2 * \dots * w_m$ and $e_{\gamma(1)} * e_{\gamma(2)} * \dots * e_{\gamma(m)}$ by e_{γ}^* . Thus $W_1(S_p^{(2)})$ is spanned by $\{e_{\gamma}^* : \gamma \in \Gamma_{m,2}\}$.

If $\alpha, \beta \in \Gamma_{m,2}$, we say that β is equivalent to α modulo $S_p^{(2)}$ and write $\alpha \equiv \beta$, if there is a permutation $\sigma \in S_p^{(2)}$ such that $\alpha\sigma = \beta$. Let $\Delta_{m,2}$ be a system of distinct representatives for the equivalence classes modulo $S_p^{(2)}$. Multilinear algebraists will recognize $\{e_\gamma^* : \gamma \in \Delta_{m,2}\}$ as a basis for $W_1(S_p^{(2)})$ [19, p. 97]. Indeed, $e_\alpha^* = e_\beta^*$ if and only if $\alpha \equiv \beta$. Graph theorists, on the other hand, may recognize $\Delta_{m,2}$ as elements of $\Gamma_{m,2}$ corresponding to a complete set of nonisomorphic labelled graphs on p vertices [13, p. 83]. The equivalence classes represented by $\Delta_{m,2}$ are the nonisomorphic *unlabelled* graphs (we will simply say *graphs*) on p vertices. That is, the labelled graph corresponding to α is isomorphic to the labelled graph corresponding to β if and only if $\alpha \equiv \beta$. In particular, we may view $W_1(S_p^{(2)})$ as the (free) vector space spanned by the graphs on p vertices.

The well-known Cauchy–Frobenius–Burnside–Pólya–Redfield formula for the number of nonisomorphic (unlabelled) graphs on p vertices follows from these remarks by taking $n = 2$ in (13), i.e.,

$$(14) \quad \dim W_1(S_p^{(2)}) = \frac{1}{p!} \sum 2^{c(\sigma)},$$

where the summation is over $\sigma \in S_p^{(2)}$ [33], [10, Chap. 7], [13], [16, p. 170], [37]. In Example 2 (below), $S_4^{(2)}$ is explicitly listed. Note, however, that (14) depends only on the cycle structures of the permutations in $S_p^{(2)}$.

As Wilf [38] observes, (14) is not a polynomial-time formula. It is not inconceivable that such a formula could be based on some alternate computation of the dimension. As a possible step in this direction we present the following:

THEOREM 2. *Let p be a positive integer. Let $m = p(p-1)/2$. Denote by ξ_q the irreducible character of S_m corresponding to the partition $(m-q, q)$, $0 \leq q \leq m/2$. Then the number of nonisomorphic, unlabelled graphs on p vertices is given by the formula*

$$(15) \quad \dim W_1(S_p^{(2)}) = \sum_{q=0}^{\lfloor m/2 \rfloor} (m-2q+1)(1, \xi_q)_p,$$

where $(1, \xi_q)_p$ is the number of occurrences of the principal character in the restriction of ξ_q to $S_p^{(2)}$.

Proof. Let $m = p(p-1)/2$, $G = S_p^{(2)}$ and $\chi = 1$ in (4). Taking $n = 2$, we see from (11) that

$$\dim W_1(S_p^{(2)}) = \sum_{q=0}^{\lfloor m/2 \rfloor} (1, \xi_q)_p \dim(2, m, \xi_q).$$

It follows from (12) that $\dim(2, m, \xi_q) = (m-2q+1)$.

Of course, the missing ingredient in Theorem 2 is an analogue of “Young’s rule” ([15, p. 51], [1, Chap. 6], [12, Chap. 7] or [18, Chap. 5]) for the computation of the “Kostka-like” coefficient $(1, \xi_q)_p$. On the other hand, the characters ξ_q are among the easiest characters of S_m to compute. For these characters, the Frobenius formulas are particularly simple. (See, e.g., [1, p. 213], [11], [12, p. 206], [16, p. 236] [18, § 8.1], [30, p. 143], or [34].)

Example 2. Consider the case $p = 4$. Number the vertices of K_4 so that the vertex set may be identified with $\{1, 2, 3, 4\}$. The edge set is then $\{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\}$. We number these 6 edges **1, 2, 3, 4, 5, 6**, respectively. Thus, for example, edge **3** in $\{1, 4\}$. As a group of permutations of the vertices, the automorphism group of K_4 is S_4 . We compute $S_4^{(2)}$, the automorphism group as a group of permutations of the edges. Let, for example, $\sigma = (1234) \in S_4$. We denote by $\hat{\sigma}$ the corresponding

permutation of the edges. Then

$$\hat{\sigma}\{1, 2\} = \{\sigma(1), \sigma(2)\} = \{2, 3\}.$$

Thus, $\hat{\sigma}(1) = 4$, i.e., $\hat{\sigma}$ sends the first edge to the fourth. Similarly,

$$\hat{\sigma}\{2, 3\} = \{\sigma(2), \sigma(3)\} = \{3, 4\},$$

$$\hat{\sigma}\{3, 4\} = \{\sigma(3), \sigma(4)\} = \{4, 1\} = \{1, 4\},$$

$$\hat{\sigma}\{1, 4\} = \{\sigma(1), \sigma(4)\} = \{2, 1\} = \{1, 2\}.$$

It follows that **(1463)** is a cycle in the disjoint cycle factorization of $\hat{\sigma}$. Since $\hat{\sigma}\{1, 3\} = \{2, 4\}$ and $\hat{\sigma}\{2, 4\} = \{1, 3\}$, we have $\hat{\sigma} = \mathbf{(1463)(25)}$. The remaining elements of $S_4^{(2)}$ may be computed in the same way, and we list them in Table 2. (The cycle structures of the permutations in $S_p^{(2)}$ may be found, e.g., in [10] and [13].)

TABLE 2

σ	$\hat{\sigma}$	σ	$\hat{\sigma}$	σ	$\hat{\sigma}$
id	id	(124)	(153)(246)	(13)(24)	(16)(34)
(12)	(24)(35)	(132)	(124)(365)	(14)(23)	(16)(25)
(13)	(14)(36)	(134)	(145)(263)	(1234)	(1463)(25)
(14)	(15)(26)	(142)	(135)(264)	(1243)	(1562)(34)
(23)	(12)(56)	(143)	(154)(236)	(1324)	(2453)(16)
(24)	(13)(46)	(234)	(123)(465)	(1342)	(1265)(34)
(34)	(23)(45)	(243)	(132)(456)	(1423)	(2354)(16)
(123)	(142)(356)	(12)(34)	(25)(34)	(1432)	(1364)(25)

In the absence of a ‘‘Young’s rule’’ type formula for computing $(1, \xi_q)_p$, we proceed directly:

$$(16) \quad (1, \xi_q)_4 = \frac{1}{4!} \sum_{\sigma \in S_4} \xi_q(\hat{\sigma}).$$

From [30, p. 144],

$$(17) \quad \xi_1(\hat{\sigma}) = a_1(\hat{\sigma}) - 1,$$

$$(18) \quad \xi_2(\hat{\sigma}) = \frac{1}{2}a_1(\hat{\sigma})(a_1(\hat{\sigma}) - 3) + a_2(\hat{\sigma}),$$

$$(19) \quad \xi_3(\hat{\sigma}) = \frac{1}{6}a_1(\hat{\sigma})(a_1(\hat{\sigma}) - 1)(a_1(\hat{\sigma}) - 5) + (a_1(\hat{\sigma}) - 1)a_2(\hat{\sigma}) + a_3(\hat{\sigma})$$

where a_r is the number of cycles of length r in the disjoint cycle factorization (of $\hat{\sigma}$). Of course, $\xi_0 = 1$. From the tabulated description of $S_4^{(2)}$, the computations from (16) are

$$(1, \xi_0)_4 = 1,$$

$$(1, \xi_1)_4 = \frac{1}{24}[5 + 6 \times 1 + 8 \times (-1) + 3 \times 1 + 6 \times (-1)] = 0,$$

$$(1, \xi_2)_4 = \frac{1}{24}[9 + 6 \times 1 + 8 \times 0 + 3 \times 1 + 6 \times 1] = 1,$$

$$(1, \xi_3)_4 = \frac{1}{24}[5 + 6 \times 1 + 8 \times 2 + 3 \times 1 + 6 \times (-1)] = 1.$$

Putting these values into (15) yields 11, i.e., there are 11 (nonisomorphic, unlabelled) graphs on 4 vertices.

Example 3. Since $\xi_0 = 1$, $(1, \xi_0)_p$ will always be 1 (for $p \geq 2$). From the Cauchy-Frobenius-Burnside lemma,

$$(20) \quad \frac{1}{p!} \sum_{\sigma \in S_p} a_1(\hat{\sigma}) = 1$$

because $S_p^{(2)}$ is transitive. It follows from (17) that $(1, \xi_1)_p = 0$ (for $p \geq 3$). Using a TRS-80 microcomputer, one of the authors produced Table 3. The number in row p , column q is $(1, \xi_q)_p$, the number of occurrences of the principal character in the restriction of ξ_q to $S_p^{(2)}$. From (15) and the 5th row of the table ($m = 5(4)/2$) we see there are $(11)(1) + (9)(0) + (7)(1) + (5)(2) + (3)(2) + (1)(0) = 34$ nonisomorphic graphs on 5 vertices. Similarly, from row 6 we confirm that there are 156 nonisomorphic graphs on 6 vertices. Note that the table is incomplete from row 7 down.

TABLE 3

	0	1	2	3	4	5	6	7	8	...
3	1	0								
4	1	0	1	1						
5	1	0	1	2	2	0				
6	1	0	1	3	4	6	6	3		
7	1	0	1	3	5	11	20	24	32	
8	1	0	1	3	6	13	32	59	106	
9	1	0	1	3	6	14	38	85	197	...
10	1	0	1	3	6	15	40	99	263	
11	1	0	1	3	6	15	41	105	295	
12	1	0	1	3	6	15	42	107	310	

THEOREM 3. Let p be an integer at least 3. Let $m = p(p - 1)/2$. Denote by ξ_q the irreducible character of S_m corresponding to the partition $(m - q, q)$, $0 \leq q \leq m/2$. If $(1, \xi_q)_p$ is the number of occurrences of the principal character in the restriction of ξ_q to $S_p^{(2)}$, then $(1, \xi_q)_p$ is constant for $p \geq 2q$, i.e., if Table 3 were continued, column q would be constant from row $2q$ on down.

We will prove Theorem 3 in § 4. It follows from this result and Table 3 that $(1, \xi_0)_p = 1, p \geq 3$; $(1, \xi_1)_p = 0, p \geq 3$; $(1, \xi_2)_p = 1, p \geq 4$; $(1, \xi_3)_p = 3, p \geq 6$; $(1, \xi_4)_p = 6, p \geq 8$; $(1, \xi_5)_p = 15, p \geq 10$; and $(1, \xi_6)_p = 42, p \geq 12$. The sequence 1, 0, 1, 3, 6, 15, 42, ... is unfamiliar to us.

Example 4. Let $P(s, t)$ be the smallest value of p such that any 2-coloring of the edges of K_p , using the colors 0 and 1, contains either a subgraph isomorphic to K_s , each edge of which is colored 0, or a subgraph isomorphic to K_t , each edge of which is colored 1. The numbers $P(s, t)$ are called *Ramsey numbers* [10]. In our context, $P(s, t)$ is the smallest value of p such that the system of distinct representatives can be chosen so that every $\alpha \in \Delta_{2,m}$ either begins with $\binom{s}{2}$ zeros or with $\binom{t}{2}$ ones.

There are other advantages to looking at graph theory through a multilinear lens. One motivation for studying $V_\chi(G)$ is that it is a representation module for $\{K_\chi^G(T): T \in GL(n, F)\}$. In fact, these same operators have a graph-theoretic interpretation. Before this can be demonstrated, we are obliged to introduce some additional notation. First recall that $K_\chi^G = J_\chi^G$ when $\chi(\text{id}) = 1$. We will use the notation J_p to denote K_χ^G when $G = S_p^{(2)}$ and $\chi = 1$.

As above, let $E = \{e_0, e_1\}$ be a basis of W . Then $E^* = \{e_\gamma^*: \gamma \in \Delta_{m,2}\}$ is a basis of $W_1(S_p^{(2)})$, where $m = p(p - 1)/2$. Suppose $T \in L(W)$. Let $A = (a_{ij})$ be the $(2 \text{ by } 2)$

matrix representation of T with respect to E . Denote by $J_p(A)$ the matrix representation of $J_p(T)$ with respect to the induced base E^* (with some fixed but arbitrary order). The entries of $J_p(A)$ are given by the following well-known result [19, p. 122]:

For $\alpha, \beta \in \Delta_{m,2}$, the (α, β) entry of $J_p(A)$ is

$$(21) \quad \frac{1}{\nu(\alpha)} \sum_{\sigma \in S_p} \prod_{t=1}^m a_{\alpha\hat{\sigma}(t),\beta(t)},$$

where $\nu(\alpha)$ is the cardinality of the stabilizer subgroup of α in $S_p^{(2)}$ (i.e., $\nu(\alpha) = o\{\hat{\sigma} \in S_p^{(2)} : \alpha\hat{\sigma} = \alpha\}$).

THEOREM 4. Let $A = \begin{pmatrix} x_0 & 0 \\ 0 & x_1 \end{pmatrix}$. Then $J_p(A)$ is a diagonal matrix. For $\alpha \in \Delta_{m,2}$, the (α, α) entry of $J_p(A)$ is $x_0^{m-s}x_1^s$, where s is the number of edges in the graph on p vertices to which α corresponds. In particular, the trace of $J_p(A)$ is the Pólya inventory ([13, p. 84]), a polynomial of degree m in x_0 and x_1 . The coefficient of $x_0^{m-s}x_1^s$ in this polynomial is the number of nonisomorphic, unlabelled graphs on p vertices with s edges.

Proof. In (21) the contribution corresponding to σ is nonzero if and only if $\alpha\hat{\sigma}(t) = \beta(t)$, $t = 1, 2, \dots, m$, i.e., if and only if $\alpha\hat{\sigma} = \beta$. But, $\Delta_{m,2}$ is a system of distinct representatives for the equivalence classes modulo $S_p^{(2)}$. That is to say, for $\alpha, \beta \in \Delta_{m,2}$, $\alpha \equiv \beta$ if and only if $\alpha = \beta$. Thus, $J_p(A)$ is diagonal.

The same considerations show that the (α, α) entry of $J_p(A)$ is

$$\frac{1}{\nu(\alpha)} \sum \prod_{t=1}^m a_{\alpha\hat{\sigma}(t),\alpha(t)} = \prod_{t=1}^m a_{\alpha(t),\alpha(t)},$$

where the summation on the left is over the stabilizer subgroup of α . The right-hand side of this equation is $x_0^{m-s}x_1^s$, where $s = o\{i : \alpha(i) = 1\}$, i.e., the number of edges in the graph corresponding to α .

We see that J_p can be viewed as a bookkeeping device. This role is even more apparent in Theorem 5, below.

DEFINITION 1. Let $\alpha, \beta \in \Gamma_{m,2}$. The distance between α and β is $h(\alpha, \beta) = o\{i : \alpha(i) \neq \beta(i)\}$.

Of course, one may identify an element $\gamma \in \Gamma_{m,2}$ with a sequence of length m of zeros and ones, i.e., a codeword. The distance $h(\alpha, \beta)$ is the usual Hamming distance from coding theory. We are interested in an inventory of the distances between unlabelled graphs (i.e., between equivalence classes of $\Gamma_{m,2}$ modulo $S_p^{(2)}$).

DEFINITION 2. Let $\alpha, \beta \in \Gamma_{m,2}$. For $i = 0, 1, \dots, m$, let

$$z_i(\alpha, \beta) = o\{\sigma \in S_p : h(\alpha\hat{\sigma}, \beta) = i\}.$$

Observe that $z_0(\alpha, \beta) \neq 0$ if and only if $\alpha \equiv \beta$, and that $z_0(\alpha, \alpha) = \nu(\alpha)$, the cardinality of the stabilizer subgroup of α . Indeed, $z_i(\alpha, \beta) = z_i(\beta, \alpha)$ is a multiple of both $\nu(\alpha)$ and $\nu(\beta)$ for all $i = 0, 1, \dots, m$. Finally, we note that if $\alpha \equiv \alpha'$ and $\beta \equiv \beta'$, then $z_i(\alpha, \beta) = z_i(\alpha', \beta')$. Define $z(\alpha, \beta) = \sum_{i=0}^m z_i(\alpha, \beta)x^i$.

Example 5. Consider the graphs in Figs. 4 and 5. With respect to a clockwise numbering of the vertices, starting with the top left corner, they correspond to $\alpha = (1, 1, 1, 0, 0, 1)$ and $\beta = (0, 0, 1, 1, 0, 1)$, respectively. (Here $\alpha(i)$ is the i th entry in the sequence α .) We may compute

$$(22) \quad z(\alpha, \beta) = 4x + 16x^3 + 4x^5.$$

From the polynomial, we see that α and β are not isomorphic, but are at a (minimum) distance of 1, i.e., α is isomorphic to a graph α' which can be transformed into β by a single edge-change (the addition or deletion of a single edge). The probability is

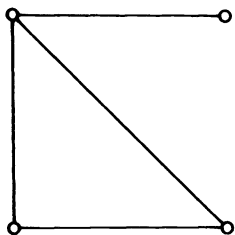


FIG. 4

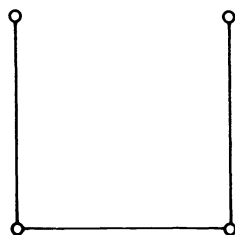


FIG. 5

4/24 that a randomly chosen isomorphic image, α'' , of α is a distance 1 from β . Finally, the expected (minimum) number of edge-changes needed to transform α'' to β is $[4 + (16 \times 3) + (4 \times 5)]/24 = 3$. (Of course, the last remark follows more easily, in this case, because the polynomial happens to be symmetric.) Note that the stabilizer subgroup of α is $\{\text{id}, (23)(45)\}$. Thus, (22) suffers some redundancy.

DEFINITION 3. For $\alpha, \beta \in \Gamma_{m,2}$, let $w(\alpha, \beta) = z(\alpha, \beta)/v(\alpha)$.

THEOREM 5. Let $A = \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix}$. Then, for $\alpha, \beta \in \Delta_{m,2}$, the (α, β) entry of $J_p(A)$ is $w(\alpha, \beta)$.

Proof. In (21),

$$\prod_{t=1}^m a_{\alpha\hat{\sigma}(t),\beta(t)} = x^{o\{i: \alpha\hat{\sigma}(i) \neq \beta(i)\}}.$$

The result follows from the definitions.

A similar concept of distance between general Pólya patterns was discussed in [27].

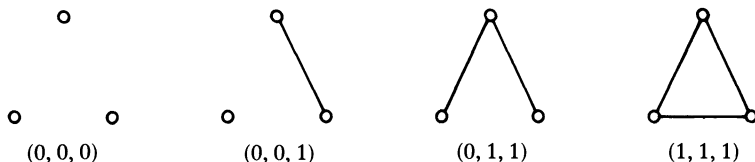


FIG. 6

Example 6. The four nonisomorphic, unlabelled graphs on 3 vertices correspond to $\Delta_{3,2} = \{(0, 0, 0), (0, 0, 1), (0, 1, 1), (1, 1, 1)\}$. With respect to the appropriate numbering, these graphs are shown in Fig. 6. If we order $\Delta_{3,2}$ as shown, then

$$(23) \quad J_3 \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 & x^3 \\ 3x & 1+2x^2 & 2x+x^3 & 3x^2 \\ 3x^2 & 2x+x^3 & 1+2x^2 & 3x \\ x^3 & x^2 & x & 1 \end{pmatrix}.$$

Thus, for example, $w((0, 0, 1), (0, 1, 1))$ is the polynomial in the (2, 3) position, namely $2x + x^3$. Of course, we may write (23) as

$$J_3 \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix} = J_3 \left(I_2 + x \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right) = I_4 + xD_1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + x^2D_2 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + x^3J_3 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

where D_1 and D_2 are derivation operators [19, § 3.2].

3. Extension to edge colored graphs. The results in § 2 were achieved by identifying a certain symmetry class of tensors with the free vector space generated by the nonisomorphic graphs on p vertices. This method may be extended to any situation in which permutation groups act on colored sets. The generalities of such extensions are straightforward. Only the details vary from case to case. In this section we indicate how the details work out for edge colored graphs.

In Theorem 2 we counted the nonisomorphic graphs on p vertices. Our approach was to count the inequivalent spanning subgraphs of the complete graph K_p . To count *inequivalent* spanning subgraphs of an arbitrary graph \mathcal{G} on p vertices one may proceed analogously, i.e., color the edges of \mathcal{G} using the two colors 0 and 1. (Counting nonisomorphic subgraphs is more difficult. When $\mathcal{G} \neq K_p$, inequivalent subgraphs may be isomorphic.) Suppose we modify this approach by coloring the edges of \mathcal{G} using the $n + 1$ colors, 0 through n . This is equivalent to choosing a subgraph of \mathcal{G} (consisting of the p vertices together with the edges not colored 0) and then coloring its edges using the n colors, 1 through n . It is also equivalent to considering multigraphs. If r edges connect vertices i and j , then “edge” $\{i, j\}$ receives “color” r . How many inequivalent, n -colored, spanning subgraphs does \mathcal{G} have? We may label the vertices of \mathcal{G} , say $1, 2, \dots, p$. If we mimic the approach of § 2, we are interested in the subgroup of $S_p^{(2)}$ which stabilizes \mathcal{G} . The disadvantage of such mimicry involves viewing the automorphism group of \mathcal{G} as a subgroup of S_m , $m = p(p - 1)/2$. If \mathcal{G} has k edges, it may be that k is much smaller than m . In this section, we let H be the automorphism group of (the edges of) \mathcal{G} as a subgroup of S_k .

Suppose U is a vector space over F of dimension $n + 1$. Let $\{e_0, e_1, \dots, e_n\}$ be a basis of U . Then $\{e_\gamma^\otimes : \gamma \in \Gamma_{k,n+1}\}$ is a basis of $U^{k \otimes}$, where $\Gamma_{k,n+1}$ is the set of all functions $\gamma : \{1, 2, \dots, k\} \rightarrow \{0, 1, \dots, n\}$. With respect to a fixed labelling of the k edges of \mathcal{G} , each $\alpha \in \Gamma_{k,n+1}$ corresponds to a labelled, edge-colored subgraph of \mathcal{G} . Two such colored subgraphs, α and β , are equivalent if there is a $\sigma \in H$ such that $\alpha\sigma = \beta$. Thus, $U_1(H)$, the symmetry class of tensors in $U^{k \otimes}$ corresponding to H and the principal character, is the free vector space corresponding to the inequivalent, unlabelled, n -colored subgraphs of \mathcal{G} . It follows that we may use (4) with $V = U$, $G = H$ and $\chi \equiv 1$ together with any of our formulas for $\dim(n + 1, k, \lambda)$ to answer our question. To be specific, we state one such result.

THEOREM 6. *Let \mathcal{G} be a graph with p vertices and k edges. Let H be the automorphism group of (the edges of) \mathcal{G} as a subgroup of S_k . The number of inequivalent ways to choose a spanning subgraph of \mathcal{G} and color its edges using n colors is*

$$(24) \quad \dim U_1(H) = \sum_{\lambda \in \mathcal{S}_k} \frac{\lambda(\text{id})(1, \lambda)_{HP_\lambda}(n + 1)}{k!}$$

where $p_\lambda(x)$ is the polynomial defined in (9).

Results analogous to Theorems 4 and 5 are available in this context as well. If $\mathcal{G} = K_p$, we may replace “inequivalent” with “nonisomorphic”.

4. Combinatorial lemmas. In this section we prove Theorem 3 from § 2. To do so we must examine how the cycle structure of a permutation σ in S_p on the p vertices of the complete graph determines the cycle structure of the corresponding permutation $\hat{\sigma}$ on the graph’s $m = \binom{p}{2}$ edges. For σ in S_p and r a positive integer, let $b_r(\sigma)$ denote the number of r cycles in the cycle decomposition of σ as a permutation in S_p and let $a_r(\hat{\sigma})$ be the number of r cycles in $\hat{\sigma}$ as a permutation in $S_p^{(2)} \subset S_m$.

LEMMA 1. *For each positive integer r , there are positive integers $c_{s,i}, d, e, f$ such that*

$$a_r(\hat{\sigma}) = d \binom{b_r(\sigma)}{2} + eb_r(\sigma) + fb_{2r}(\sigma) + \sum c_{s,i} b_s(\sigma) b_i(\sigma)$$

for all p and all σ in S_p . The sum is taken over all pairs (s, t) such that $s \neq t$ and $\text{LCM}\{s, t\} = r$.

Example 7. If $r = 1$ or 2 then

$$(25) \quad a_1(\hat{\sigma}) = \binom{b_1(\sigma)}{2} + b_2(\sigma), \quad a_2(\hat{\sigma}) = 2\binom{b_2(\sigma)}{2} + b_4(\sigma) + b_1(\sigma)b_2(\sigma).$$

A proof of Lemma 1 along with a discussion of how to compute the coefficients $c_{s,t}$, d , e , f is contained in [10, pp. 283–290].

Next we define the *degree* of a “monomial” of the form $b_{r_1}(\sigma) \cdots b_{r_u}(\sigma)$, or $a_{r_1}(\hat{\sigma}) \cdots a_{r_u}(\hat{\sigma})$, to be $r_1 + \cdots + r_u$. To proceed all we need from Lemma 1 is the fact that $a_r(\hat{\sigma})$ is a linear combination of monomials in $b_t(\sigma)$ of degree $\leq 2r$.

LEMMA 2. *Every monomial $a_{r_1}(\hat{\sigma}) \cdots a_{r_u}(\hat{\sigma})$ of degree r is a linear combination of monomials $b_{t_1}(\sigma) \cdots b_{t_u}(\sigma)$ of degree $\leq 2r$.*

This follows from Lemma 1 since if $\text{LCM}\{s, t\} = r$, then $s + t \leq 2r$.

Example 8. From (25) in Example 7, we have the monomial

$$(26) \quad a_1(\hat{\sigma})a_2(\hat{\sigma}) = \left[\binom{b_1(\sigma)}{2} + b_2(\sigma) \right] \left[b_1(\sigma)b_2(\sigma) + 2\binom{b_2(\sigma)}{2} + b_4(\sigma) \right].$$

This formula holds for all p and all σ in S_p . The left side of (26) is a monomial of degree 3 and the right side is a linear combination of monomials of degrees $\leq 2 \cdot 3 = 6$.

We have not yet discussed the main ingredient of Theorem 3, namely the characters ξ_q corresponding to the partitions $(m - q, q)$ of $m = \binom{p}{2}$. The next lemma describes the relation between ξ_q and the monomials $b_{r_1}(\sigma) \cdots b_{r_u}(\sigma)$ and $a_{r_1}(\hat{\sigma}) \cdots a_{r_u}(\hat{\sigma})$.

LEMMA 3. *For each positive integer $q \leq m/2$, $\xi_q(\hat{\sigma})$ is a fixed linear combination of monomials $a_{r_1}(\hat{\sigma}) \cdots a_{r_u}(\hat{\sigma})$ of degrees $\leq q$. Hence $\xi_q(\hat{\sigma})$ is a fixed linear combination of monomials $b_{r_1}(\sigma) \cdots b_{r_u}(\sigma)$ of degrees $\leq 2q$.*

The first statement in Lemma 3 means that there are formulas for ξ_q analogous to (17), (18) and (19). This fact is a result of formula (5.9) in [30, p. 144]. The second statement in Lemma 3 is a result of the first and of Lemma 2.

It follows from Lemma 3 that $(1, \xi_q)_p$ is a linear combination of averages of the form

$$(27) \quad \frac{1}{p!} \sum_{\sigma \in S_p} b_{r_1}(\sigma) \cdots b_{r_u}(\sigma)$$

with $r_1 + \cdots + r_u \leq 2q$, for any p satisfying $q \leq p(p - 1)/4$.

Example 9. From equations (17) and (25) we get

$$(28) \quad \begin{aligned} (1, \xi_1)_p &= \frac{1}{p!} \sum (a_1(\hat{\sigma}) - 1) \\ &= \frac{1}{p!} \sum \left(\frac{b_1(\sigma)(b_1(\sigma) - 1)}{2} + b_2(\sigma) - 1 \right) \\ &= \frac{1}{2} \left(\frac{1}{p!} \sum b_1^2(\sigma) \right) - \frac{1}{2} \left(\frac{1}{p!} \sum b_1(\sigma) \right) + \left(\frac{1}{p!} \sum b_2(\sigma) \right) - 1, \end{aligned}$$

for all p . (All sums are taken over σ in S_p .)

The next lemma will be used to show that averages of the form (27) are constant for $p \geq r_1 + r_2 + \cdots + r_u$ and thus that $(1, \xi_q)_p$ is constant for $p \geq 2q$. For b and k positive integers, define $(b)_k = b(b - 1) \cdots (b - k + 1)$, the “falling factorial”.

LEMMA 4. Let k_1, k_2, \dots, k_t be nonnegative integers. Then

$$(29) \quad \frac{1}{p!} \sum_{\sigma \in S_p} (b_1(\sigma))_{k_1} \cdots (b_t(\sigma))_{k_t} = \begin{cases} \frac{1}{1^{k_1} 2^{k_2} \cdots t^{k_t}}, & \text{if } p \cong k_1 + 2k_2 + \cdots + tk_t, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 4 and its proof appear in [16, p. 229].

Next, we prove Theorem 3. It follows from Lemma 4 that sums of the form

$$\frac{1}{p!} \sum_{\sigma \in S_p} (b_1(\sigma))_{k_1} \cdots (b_t(\sigma))_{k_t}$$

are constant for all $p \cong k = k_1 + 2k_2 + \cdots + tk_t$.

Let $b_{r_1}(\sigma) \cdots b_{r_u}(\sigma)$ be a monomial of degree $r = r_1 + \cdots + r_u$. A simple inductive argument shows that this monomial is a sum of terms of the form $(b_1(\sigma))_{k_1} \cdots (b_t(\sigma))_{k_t}$, where $k_1 + 2k_2 + \cdots + tk_t \cong r$. It follows that averages of the form (27) are constant for $p \cong r_1 + \cdots + r_u$. This fact combined with the comments before Lemma 4 proves Theorem 3.

Example 10. From (28) and Lemma 4 we have

$$(1, \xi_1)_p = \frac{1}{2} \left(\frac{1}{p!} \sum (b_1(\sigma))_2 \right) + \left(\frac{1}{p!} \sum (b_2(\sigma))_1 \right) - 1 = \frac{1}{2} \left(\frac{1}{1^2} \right) + \left(\frac{1}{2^1} \right) - 1 = 0,$$

for all $p \cong 2$.

REFERENCES

- [1] H. BOERNER, *Representations of Groups*, North-Holland/Elsevier, Amsterdam/London/New York, 1970.
- [2] G. H. CHAN, *A note on symmetrizers of rank 1*, Nanta Math., 11 (1978), pp. 130–133.
- [3] ———, *On the triviality of a symmetry class of tensors*, Linear and Multilinear Algebra, 6 (1978), pp. 78–82.
- [4] ———, *(k)-characters and the triviality of symmetry classes*, Linear Algebra Appl., 25 (1979), pp. 139–149.
- [5] ———, *(k)-characters and the triviality of symmetry classes, II*, Nanta Math., 12 (1979), pp. 7–15.
- [6] ———, *Minimal (k)-groups of degree n, 3k < n ≦ 4k*, Linear and Multilinear Algebra, 7 (1979), pp. 155–166.
- [7] G. H. CHAN AND M. H. LIM, *Nonzero symmetry classes of smallest dimension*, Canad. J. Math., 32 (1980), pp. 957–968.
- [8] ———, *Nonzero symmetry classes of smallest dimension II*, manuscript.
- [9] S. C. CHANG, *On the vanishing of a (G, σ) space*, Chinese J. Math., 4 (1976), pp. 1–7.
- [10] D. I. A. COHEN, *Basic Techniques of Combinatorial Theory*, John Wiley, New York, 1978.
- [11] A. GAMBA, *Sui caratteri delle rappresentazioni del gruppo simmetrico*, Atti Accad. Naz. Lincei, 8 (1952), pp. 167–169.
- [12] M. HAMERMESH, *Group Theory*, Addison-Wesley, Reading, MA, 1962.
- [13] F. HARARY AND E. M. PALMER, *Graphical Enumeration*, Academic Press, New York, 1973.
- [14] I. M. ISAACS, *Character Theory of Finite Groups*, Academic Press, New York, 1976.
- [15] G. D. JAMES, *The Representation Theory of the Symmetric Groups*, Lecture Notes in Mathematics 682, Springer-Verlag, New York, 1978.
- [16] G. D. JAMES AND A. KERBER, *The Representation Theory of the Symmetric Group*, Addison-Wesley, Reading, MA, 1981.
- [17] R. C. KING, *The dimensions of irreducible representations of linear groups*, Canad. J. Math., 22 (1970), pp. 436–448.
- [18] D. E. LITTLEWOOD, *The Theory of Group Characters*, Oxford Univ. Press, London, 1958.
- [19] M. MARCUS, *Finite Dimensional Multilinear Algebra*, Part 1, Marcel Dekker, New York, 1973.
- [20] M. MARCUS AND J. CHOLLET, *The index of a symmetry class of tensors*, Linear and Multilinear Algebra, 11 (1982), pp. 277–281.

- [21] R. MERRIS, *Multilinear Algebra*, Monograph Series, Institute for the Interdisciplinary Appl. of Algebra and Combinatorics, Univ. California, Santa Barbara, CA, 1975.
- [22] ———, *The dimensions of certain symmetry classes of tensors II*, *Linear and Multilinear Algebra*, 4 (1976), pp. 205–207.
- [23] ———, *Pattern enumeration and Young diagrams*, Abstract 78T-A163, *Notices Amer. Math. Soc.*, 25 (1978), p. A-571.
- [24] ———, *Recent advances in symmetry classes of tensors*, *Linear and Multilinear Algebra*, 7 (1979), pp. 317–328.
- [25] ———, *Manifestations of Pólya's counting theorem*, *Linear Algebra Appl.*, 32 (1980), pp. 209–234.
- [26] ———, *Pólya's counting theorem via tensors*, *Amer. Math. Monthly*, 88 (1981), pp. 179–185.
- [27] ———, *Generalized matrix functions and pattern inventory*, *Linear and Multilinear Algebra*, 12 (1983), pp. 315–327.
- [28] R. MERRIS AND M. A. RASHID, *The dimensions of certain symmetry classes of tensors*, *Linear and Multilinear Algebra*, 2 (1974), pp. 245–248.
- [29] R. MERRIS AND W. WATKINS, *Elementary divisors of induced transformations on symmetry classes of tensors*, *Linear Algebra Appl.*, 38 (1981), pp. 17–26.
- [30] F. D. MURNAGHAN, *The Theory of Group Representations*, Dover, New York, 1963.
- [31] G. MURTAZA AND M. A. RASHID, *Duality of a Young diagram describing a representation and dimensionality formulas*, *J. Math. Phys.*, 14 (1973), pp. 1196–1198.
- [32] M. NEWMAN, *Matrix Representations of Groups*, National Bureau of Standards Applied Math. Series 60, Superintendent of Documents, Washington, D.C. 1968.
- [33] G. PÓLYA, *Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen*, *Acta Math.*, 68 (1937), pp. 145–254.
- [34] T. TSUZUKU, *On multiple transitivity of permutation groups*, *Nagoya Math. J.*, 18 (1961), pp. 93–109.
- [35] R. WESTWICK, *A note on symmetry classes of tensors*, *J. Algebra*, 15 (1970), pp. 309–311.
- [36] D. E. WHITE, *Multilinear techniques in Pólya enumeration theory*, *Linear and Multilinear Algebra*, 7 (1979), pp. 299–315.
- [37] ———, *Monotonicity and unimodality of the pattern inventory*, *Adv. in Math.*, 38 (1980), pp. 101–108.
- [38] H. S. WILF, *What is an answer?*, *Amer. Math. Monthly*, 89 (1982), pp. 289–292.
- [39] S. G. WILLIAMSON, *Operator theoretic invariants and the enumeration theory of Pólya and de Bruijn*, *J. Combin. Theory*, 8 (1970), pp. 162–169.
- [40] ———, *Pólya's counting theorem and a class of tensor identities*, *J. London Math. Soc. (2)*, 3 (1971), pp. 411–421.
- [41] ———, *Symmetry operators of Krantz products*, *J. Combin. Theory*, 11 (1971), pp. 122–138.

PERFECT STORAGE REPRESENTATIONS FOR FAMILIES OF DATA STRUCTURES*

F. R. K. CHUNG†, A. L. ROSENBERG‡ AND LAWRENCE SNYDER§

Abstract. In this paper we investigate the problem of finding efficient universal storage representations for certain families of data structures, such as the family T_n of n -node binary trees, where the constituent parts of family members are labelled according to a uniform naming scheme. For example, each node of a tree in T_n can be labelled by a binary string describing the sequence of left and right edges taken to reach that node from the root. If one preassigns a distinct memory location to each possible distinct name, then any member of T_n can be stored by storing the contents of each node in the location assigned to the label of that node. However, this would require $2^n - 1$ memory locations and is wasteful of space, since certain labels can never occur together in a tree in T_n and hence could share a single memory location. We consider the problem of minimizing the number of memory locations needed, viewed in the following general form:

Consider a collection Γ of labelled finite graphs, where each graph has distinctly labelled vertices but different graphs in Γ may share certain vertex labels. A graph U is universal for Γ if U contains every graph $G \in \Gamma$ as a subgraph; U is perfect-universal for Γ if it is universal and there exists a perfect hash function h that maps the labels of graphs in Γ to vertices of U such that h is one-to-one on the vertex labels of each $G \in \Gamma$.

We will show that the smallest perfect-universal graph for T_n has size roughly the square root of the size of the name space and the vertex degree need be no more than 9. We also consider several other families of graphs motivated by data structures: the family $T_n^{(k)}$ of n node k -ary trees, the family C_n of $(\leq n)$ -position two-dimensional chaotic arrays, the family R_n of $(\leq n)$ -position two-dimensional ragged arrays and the family of A_n of $(\leq n)$ -position rectangular arrays. Sharp bounds for the perfect universal graphs for $T_n^{(k)}, C_n, R_n, A_n$ are established and perfect hash functions are explicitly constructed.

Introduction. In this paper we study three questions related to the problem of finding flexible storage representations for data structures. The formal framework evolved from the following considerations.

Many families of graphs and of data structures admit consistent families of naming schemes for their constituent parts (cf. [11]). For example, the atomic entries of d -dimensional arrays are referenced via d -tuples of integers; entries of ragged arrays are also often so referenced; and the "name" $\langle i, j \rangle$ refers to the same entry of a two-dimensional array no matter what size or shape the array has. Similarly, the leftmost grandchild of the root of an ordered binary tree is often referred to as "left, left" or as "LL" or, in the case of LISP S -expressions, as "car(car(root))"; and all of these naming schemes assign the same name to this leftmost grandchild no matter what shape or size the tree has. Thus one can consistently regard these as being a single familial naming scheme for binary trees or for d -dimensional arrays.

One consequence of the existence of these familial naming schemes is that, in place of dynamically allocating storage for an n -element member of a family of data structures that admits such a naming scheme, one could statically allocate storage for

*Received by the editors November 24, 1980, and in revised form January 24, 1983. This paper was typeset at Bell Laboratories, Murray Hill, New Jersey, using the **troff** program running under the UNIX™ operating system. Final copy was produced on April 6, 1983.

†Bell Laboratories, Murray Hill, New Jersey 07974.

‡Department of Computer Science, Duke University, Durham, North Carolina 27706. A portion of the research of this author was supported by the National Science Foundation under grant MCS 8116522, and a portion was done while this author was with the IBM Research Center, Yorktown Heights, New York.

§Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907. The research of this author was supported in part by the National Science Foundation under grant MCS 78-04749, while this author was at the Department of Computer Science, Yale University, New Haven, Connecticut.

the entire subfamily of n -element structures by using the familial naming scheme: One could set aside storage space sufficient to accommodate all names that could ever occur together in some n -element member of the family. (For instance, the names $\langle 1,4 \rangle$ and $\langle 4,1 \rangle$ cannot coexist in the same 8-element rectangular array and, so, when allocating space for such arrays, one would permit these elements to share the same space.) One would then store (and retrieve) any particular n -element structure by hashing (with a guarantee of no collisions) to the locations corresponding to the names of the particular elements used. Rosenberg and Stockmeyer [15] use this strategy to allocate storage for rectangular two-dimensional arrays. The most straightforward realization of this strategy for ordered binary trees is based on the fact that every n -node binary tree is a subtree of the depth- n complete binary tree; or, equivalently, the name space for n -node binary trees is a transliteration of the set of binary strings of length less than n . Thus, by laying out the complete depth- n binary tree in contiguous memory locations without pointers (e.g., the root is assigned to relative location 1, $\text{location}(\text{left}(x)) = 2\text{location}(x)$, and $\text{location}(\text{right}(x)) = 2\text{location}(x)+1$), one has effectively stored any n -node binary tree without pointers. The obvious flaw in this example is that one has allocated 2^n-1 storage locations to save $n-1$ pointers. Similar scenarios can be described using the $n \times n$ array to "store" all n -element two-dimensional rectangular arrays or ragged arrays. The first question we shall address in this paper is: Do more efficient static allocation strategies exist, and if so, how conservative of storage can they be? In the course of answering this question, we shall be extending Sprugnoli's [16] work on collision-free hashing schemes to data structures rather than unstructured sets; Rosenberg and Stockmeyer's [15] work on storage schemes for rectangular arrays of unspecified sizes to data structures other than rectangular arrays; and Lipton, Rosenberg, and Yao's [9] work on hashing schemes for extendible data structures to the case where the hashing schemes must be *perfect* in Sprugnoli's sense (i.e., collision-free). (The use of "perfect" in our title derives from Sprugnoli's use of this term.)

A second (but closely related) use one can make of familial naming schemes is the following. There are a variety of situations in which one wishes to view either data structures [10], [13], [14] or circuits [17] as graphs. In such circumstances one often wishes to deal with families of graphs but soon finds it onerous to have to deal with each graph in the family individually: one would like to have a single "universal" graph that contains as a subgraph each of the graphs in the family in question. In the context of [17], for instance, one would be able, in the presence of a universal graph for trees, to design a single circuit that can be specialized to any individual tree circuit, rather than having to design a special circuit for each individual tree. Indeed, F. R. K. Chung and R. L. Graham have (with coauthors) [2]-[6] studied in detail the problem of finding universal graphs for the family of n -node trees, as well as the special version of the problem where the universal graph must itself be a tree. Aleliunas and Rosenberg [1] study the analogous problem for rectangular arrays. What is not addressed in the cited works is the problem of how hard it is to find the placement of a given graph in the universal graph. One approach to this problem is to have the placement of each individual graph be given by a perfect hash function from the sets of vertices of the individual graphs into the set of vertices of the universal graph. One would not be surprised to learn that universal graphs that are *perfect* in this sense must often be bigger than universal graphs that are constructed without an eye to the layout problem, but how much bigger need they be? The second task of this paper is to quantify the difference in the sizes of universal graphs and *perfect-universal* graphs for the families of graphs that we study.

In [10], [13] it is proposed to study the problem of finding linked storage representations for data structures via a type of graph embedding. One problem with the approach advocated in these papers and their successors is that they overlook the question of programmability. Specifically, they describe the following scenario: One has a structurally complicated graph that represents one's logical data structure. For reasons related to the particulars of one's computing environment, it would prove onerous to store this graph structure directly. Instead, one "encodes" this graph in a structurally simpler one, replacing edges in the complicated graph by paths in the simpler one, thereby trading traversal time for simplicity of storage management. The problem not addressed in these studies is: When one is traversing one's data structure in its encoding form, how does one find the paths corresponding to the edges one wishes to traverse. An attempt is made in [14] to resolve this problem by imposing a notion of "uniformity" on data encodings; but the message of that paper is that uniformity leads to insufferable time- and space-inefficiency in encodings. We make another attempt at the programmability problem here. Say that one has a universal graph for the family of target graphs in a data encoding situation. Then any ensemble of encodings of the source graphs into this target graph induces, in a natural way, a universal graph for the source family: vertices of the universal-source are the images (under the encodings) of the vertices of the source graphs; and edges are induced in the obvious way by the source edges (so vertices v and v' are adjacent in the universal-source whenever they are the images of adjacent vertices in one of the individual source graphs). The programmability problem is alleviated somewhat by this use of universal graphs, since there is now only one encoding to worry about, namely the encoding of the universal-source in the universal-target, rather than a whole family of encodings. Of course this latter allegation is true only if the universal graphs in question have bounded vertex-degrees, for otherwise one still has unboundedly many edge-path associations to distinguish among at each step. Indeed, we shall place great stock here on our universal graphs' having bounded vertex-degrees.

Having outlined the problems that motivated our study, let us begin to develop our formal framework.

1. Basic definitions.

A. Perfect hash functions. Let S_1, S_2, \dots, S_n be a collection of finite sets, and let T be a set. The function

$$\phi: S_1 \cup S_2 \cup \dots \cup S_n \rightarrow T$$

is a *perfect hash function from the collection of S_i into T* if ϕ is one-to-one on each set S_i .

B. Perfect universal graphs. Let $\Gamma = \{G_i\}$ be a family of finite labelled undirected graphs. For brevity, we shall often call a graph $G \in \Gamma$ a " Γ -graph." The *size* of the collection Γ , denoted $\text{Size}(\Gamma)$, is defined to be the cardinality of the union of the vertex-sets of all Γ -graphs.

An undirected graph U is *universal for n -vertex Γ -graphs* if U contains each n -vertex Γ -graph as a subgraph.

An undirected graph U is *perfect-universal for n -vertex Γ -graphs* if it is universal for these graphs and if its universality is witnessed by a perfect hash function from the collection of vertex-sets of Γ -graphs into the vertex-set of the graph

U ; that is to say, there is a total *allocation* function

$$\alpha: \bigcup_{G \in \Gamma} \text{Vertices}(G) \rightarrow \text{Vertices}(U)$$

satisfying the following. For every Γ -graph G having at most n vertices, if there is an edge in G connecting vertices v and v' , then there is an edge in U connecting vertices $\alpha(v)$ and $\alpha(v')$. If the graphs in Γ have mutually disjoint vertex-sets, then the notions of perfect and ordinary universal graph coincide; however, for the cases we shall be studying, the graphs in each Γ will share many names, and so perfect universality will be a much stronger property than unrestricted universality.

C. The perfection number of a family of graphs. The *perfection number* of the family of graphs Γ is denoted $\text{Perf}(\Gamma)$ and is defined to be the size (= number of vertices) of the smallest perfect-universal graph for Γ . It is an immediate consequence of our framework that $\text{Perf}(\Gamma)$ is also the size of the smallest set T into which the vertex-sets of the graphs in Γ can be hashed perfectly.

D. The families of graphs of interest. We now present the families of graphs we shall be studying, by formal definition and by picture (see Fig. 1). In preparation, we present the following notational conventions.

For each nonnegative integer n , we denote

- by $[n]$ the set $\{0, 1, \dots, n-1\}$;
- by $\{0, 1\}^n$ the set of all the 2^n length- n *binary* strings;
- by $\{0, 1\}^*$ the set of all finite-length binary strings;
- and by $[k]^*$ the set of all finite-length k -ary strings.

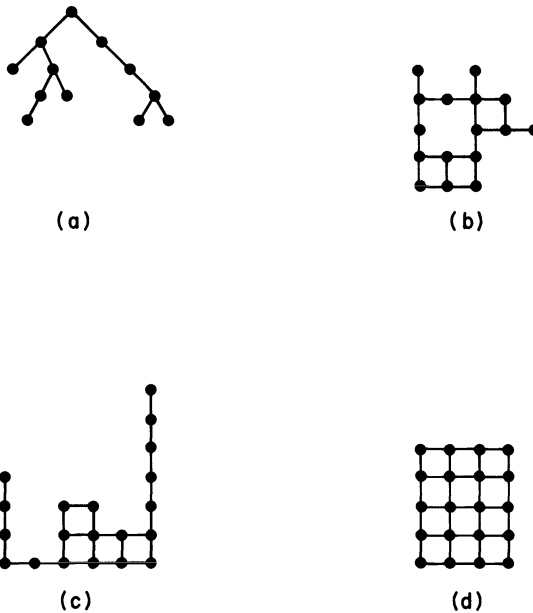


Fig. 1. Instances of the four graph families: (a) binary trees, (b) chaotic arrays, (c) ragged arrays, (d) rectangular arrays.

Trees. An n -node (rooted, ordered) binary tree is a graph whose vertex-set is an n -element prefix-closed subset of $\{0,1\}^*$ (i.e., the string x is in the set whenever either $x0$ or $x1$ is), and whose edge-set comprises all 2-element subsets of the vertex-set of the form $(x, x\sigma)$ where $x \in \{0,1\}^*$ and $\sigma \in \{0,1\}$. We denote by T_n the family of binary trees having n or fewer nodes and by $T_n^{(k)}$ the family of k -ary trees whose vertex sets are $(\leq n)$ -element prefix-closed subsets of $[k]^*$.

Chaotic arrays. An n -position (two-dimensional) chaotic array is a graph whose vertex-set is an order-closed n -element subset of $N \times N$ (i.e., each vertex is a pair of nonnegative integers, and for each pair $\langle r,s \rangle \neq \langle 0,0 \rangle$ in the vertex-set, at least one of $\langle r-1,s \rangle$ and $\langle r,s-1 \rangle$ is also in the set), and whose edge-set comprises all 2-element subsets of the vertex-set of the form $(p, p+\delta)$ where $p \in N \times N$ and $\delta \in \{\langle 0,1 \rangle, \langle 1,0 \rangle\}$. We denote by C_n the family of chaotic arrays having n or fewer vertices.

Ragged arrays. An n -position (two-dimensional) ragged array is a chaotic array whose vertex-set satisfies the following two conditions.

- If $\langle r,s \rangle \in N \times N$ is a vertex of the array, then so also is every element of the set $\{r\} \times [s]$;
- if $\langle r,0 \rangle$ is a vertex of the array, then so also is every element of $[r] \times \{0\}$.

We denote by R_n the family of ragged arrays having n or fewer vertices.

Rectangular arrays. An n -position rectangular array is an n -position ragged array whose vertex-set is of the form

$$[a] \times [b]$$

for some nonnegative integers a and b (perforce, $ab = n$). We denote by A_n the family of rectangular arrays having n or fewer vertices.

2. Perfect universal graphs for binary trees and k -ary trees. The main result of this section is that, even though $\text{Size}(T_n) = 2^n - 1$, there are perfect-universal graphs for T_n whose size is roughly only the square root of this quantity. This savings of a square root appears to be very positive, but it contrasts unfavorably with the result by Chung and Graham [5] to the effect that there are (unrestricted) universal graphs for T_n (or $T_n^{(k)}$) of n vertices and $O(n \log n)$ edges, and with the result by Chung, Coppersmith, and Graham [3] to the effect that there are (unrestricted) universal trees for T_n (or $T_n^{(k)}$) of size roughly $n^{O(\log n)}$. For the case of k -ary trees, $k \geq 3$, the perfect-universal graph of $T_n^{(k)}$ is again of size roughly the square root of $\text{Size}(T_n^{(k)}) = (k^n - 1)/(k - 1)$.

Notation. Let $x = \sigma_1 \dots \sigma_n$ be a binary string (each $\sigma_i \in \{0,1\}$). We define:

$$\text{prefix}(x;k) = \text{if } 1 \leq k \leq n \text{ then } \sigma_1 \dots \sigma_k \text{ else } \lambda;$$

and

$$\text{suffix}(x;k) = \text{if } 1 \leq k \leq n \text{ then } \sigma_{n-k+1} \dots \sigma_n \text{ else } \lambda.$$

λ here and throughout denotes the null string. We say that the string x has length $|x| = n$. When x is viewed as a node in a tree, it is said to reside at level n of the tree.

A. Basic lemmas.

LEMMA 2.1. *If the binary strings x and y satisfy*

$$|x| \leq \lfloor (n-1)/2 \rfloor \text{ and } |y| \leq \lfloor n/2 \rfloor$$

then x and y can coexist in the same n -node binary tree (i.e., there is such a tree whose vertex-set contains both x and y).

Proof. The path from x to the root node λ to y traverses a rooted ordered tree. The number of nodes in this tree is at most the number of nodes encountered on the path, namely,

$$|x| + |y| + 1 \leq \lfloor (n-1)/2 \rfloor + \lfloor n/2 \rfloor + 1 = n. \quad \square$$

LEMMA 2.2. *Let x and y be binary strings of common length $\lfloor n/2 \rfloor + m$, where $0 \leq m \leq \lfloor (n-1)/2 \rfloor$. Then x and y can coexist in the same ($\leq n$)-node binary tree if and only if*

$$\text{prefix}(x; 2m+1-(n \bmod 2)) = \text{prefix}(y; 2m+1-(n \bmod 2)).$$

Proof. We shall force x and y into the same tree and then see what happens if we insist that the trees have at most n nodes.

Consider the tree that has unary nodes at levels 0 through $k-1$; a binary node at level k ; and two length- l chains from this binary node to x and y . See Fig. 2. Now this tree has depth $l+k$, and it has $2l+k+1$ nodes. By assumption on the size of T , then,

$$l+k = \lfloor n/2 \rfloor + m;$$

and by our insistence,

$$2l+k+1 \leq n.$$

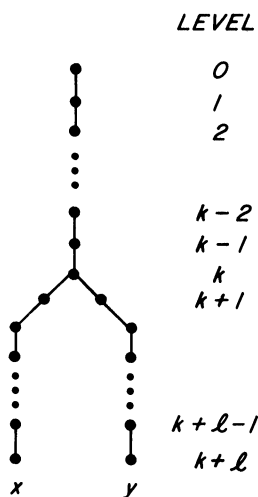


Fig. 2. Illustrating Lemma 2.2: the n -node tree containing both x and y .

Now if n is even (say $n=2a$), the first equation becomes

$$l+k = a+m,$$

while the inequality becomes

$$2l+k+1 \leq 2a,$$

so that

$$k \geq 2m+1.$$

Similar manipulation shows that when n is odd,

$$k \geq 2m.$$

Since k is the length of the prefix that is common to x and y , we have thus established the lemma. \square

LEMMA 2.3. *The binary strings $x = 0x'$ and $y = 1y'$ can coexist in the same ($\leq n$)-node binary tree if and only if*

$$|x|+|y| \leq n-1.$$

Proof. Necessity follows from the fact that the path connecting x and y passes through at least $|x|+|y|+1$ distinct nodes (including x and y). Sufficiency follows from the fact that the shortest path connecting x and y is a $(|x|+|y|+1)$ -node tree. \square

B. The lower bound. Our lower bound on the size of a perfect-universal graph for n -node binary trees is independent of any degree constraints: it holds for unbounded-degree universal graphs as well as for bounded-degree ones.

THEOREM 2.4. *For each integer n ,*

$$\text{Perf}(T_n) \geq (3-(n \bmod 2)) \cdot \exp 2(\lfloor (n-1)/2 \rfloor) - 1.$$

(Throughout this paper, $\exp k(a) = k^a$.)

Proof. By definition of “perfect”, the allocation function α for T_n must be one-to-one on the set of nodes of any n -node binary tree. By Lemma 2.1, therefore, α cannot identify (= map to the same vertex) any two strings that both have length $\leq l = \lfloor (n-1)/2 \rfloor$; this forces the target set of α to have at least $\sum_{0 \leq i \leq l} 2^i = 2^{l+1} - 1$ elements. When n is odd, this number cannot be raised on the basis of Lemma 2.1. However, Lemma 2.1 informs us that any string of length $\lfloor n/2 \rfloor$ can also coexist with any of the already mentioned strings in an n -node binary tree; and when n is even, any string of length $\lfloor n/2 \rfloor = l+1$ can coexist with any of the strings of length $\leq l$. According to Lemma 2.2, two strings of length $l+1$ can coexist in an n -node binary tree only if they start with the same symbol. Therefore, we need assign only 2^l images for these “long” strings. Thus, when n is even, the range of α must have $2^{l+1} - 1 + 2^l = 3 \exp 2(\lfloor (n-1)/2 \rfloor) - 1$ vertices. \square

C. The upper bound. Obviously, the complete binary tree of depth $n-1$, that is, the tree whose vertex-set is the prefix-closure of the set of binary strings of length $n-1$, is perfect-universal for n -node binary trees. However, this tree has $2^n - 1$ nodes, the square of the number that Theorem 2.4 asserts a perfect-universal graph must have. In fact, Theorem 2.4’s necessary number of nodes is correct not only in order of

magnitude, it is *exactly* the right number: the upper bound we derive now coincides exactly with the lower bound of that theorem.

THEOREM 2.5. *For each integer n ,*

$$\text{Perf}(T_n) = (3 - (n \bmod 2)) \cdot \exp 2(\lfloor (n-1)/2 \rfloor) - 1.$$

Moreover, there is a perfect-universal graph for n -node binary trees having just this many vertices and having vertex-degree ≤ 9 .

Proof. We shall describe the universal graph U and the allocation function α in tandem.

Vertices. Letting $m = \lfloor (n-1)/2 \rfloor$, the vertex set of U is the set

$$A \cup B$$

where

$$A = \bigcup_{0 \leq k \leq m} \{0,1\}^k$$

and

$$B = \text{if } n \text{ even then } \{0,1\}^m \text{ else EMPTY.}$$

Edges. We shall not enumerate the edges of U explicitly, choosing instead to specify them implicitly by the rule:

Vertices v and v' of U are connected by an edge precisely when $\alpha^{-1}(v)$ contains a tree-node $x \in \{0,1\}^*$ and $\alpha^{-1}(v')$ contains either $x0$ or $x1$, or when this situation occurs with the roles of v and v' reversed.

Allocation. We describe the allocation function α by cases. Let the string $x \in \{0,1\}^*$ be the argument to α :

(1) if $|x| \leq \lfloor (n-1)/2 \rfloor$, then

$$\alpha(x) = x;$$

(2) else if $|x| = \lfloor n/2 \rfloor$, and n is even, then

$$\alpha(x) = 0 \cdot \text{suffix}(x; \lfloor (n-1)/2 \rfloor);$$

(3) else if x is of the form $\sigma x'$ for $\sigma \in \{0,1\}$, and if $|x| = \lfloor (n-1)/2 \rfloor + m$ for $1 \leq m \leq \lfloor n/2 \rfloor$, then

$$\alpha(x) = (\sim\sigma) \cdot \text{suffix}(x; \lfloor n/2 \rfloor - m).$$

where $\sim\sigma$ denotes the element in $\{0,1\} - \{\sigma\}$. (See Fig. 3.)

Verification. It remains to show that the function α is a perfect hash function, i.e., is one-to-one when restricted to any n -node binary tree. This demonstration follows simply from Lemmas 2.2 and 2.3.

First of all, by Lemma 2.2, strings x and y , both of length $\lfloor n/2 \rfloor + m$, cannot coexist in the same n -node binary tree unless their length- $(2m+1-(n \bmod 2))$ prefixes are identical. The first consequence of this is that, in particular, strings $x = 0x'$ and $y = 1y'$, both of length $\lfloor n/2 \rfloor$, cannot coexist in the same n -node binary tree if n is even. Hence, clause (2) in the definition of α cannot keep α from being one-to-one on n -node trees. The second consequence is that the identifications of "long" strings in clause (3) of the definition of α cannot keep α from being one-to-one on n -node binary trees. Specifically, in that clause, α identifies all length- $(\lfloor n/2 \rfloor + m)$ strings that begin with the same symbol and have the same length- $(\lfloor n/2 \rfloor - m)$ suffix.

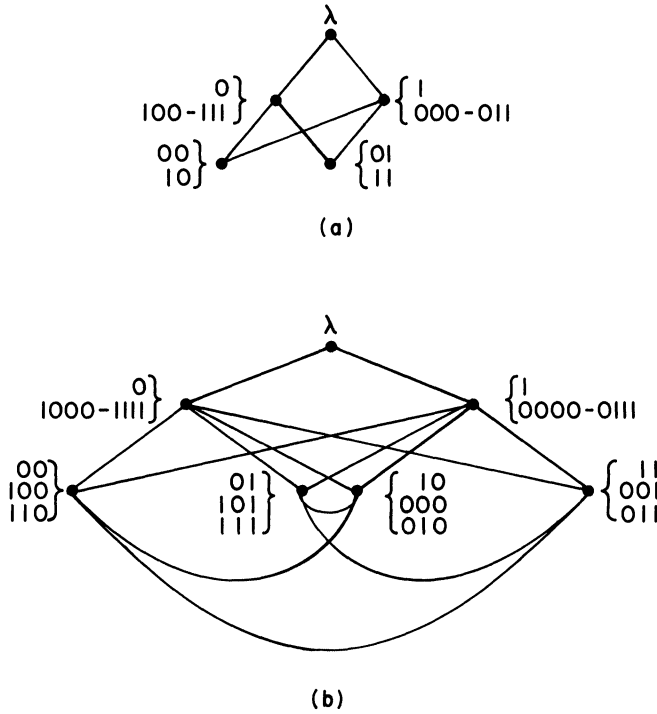


Fig. 3. Sample perfect-universal graphs for n -node binary trees: (a) $n = 4$, (b) $n = 5$.

Since the identical strings share this long suffix, they cannot share the long prefix required by Lemma 2.2 and yet remain distinct. In other words, distinct identified strings of length $\geq \lfloor n/2 \rfloor + m$ cannot coexist in the same n -node tree.

Secondly, by Lemma 2.3, strings $x = 0x'$ and $y = 1y'$ cannot coexist in an n -node binary tree if one of these has length $\lfloor n/2 \rfloor + m$ and the other has length $> \lfloor n/2 \rfloor - m$. Consequently, the (“long” string)- (“short” string) identifications made by α in clause (3) cannot prevent its being one-to-one on n -node trees: α identifies length- $(\lfloor n/2 \rfloor + m)$ strings with length- $(\lfloor n/2 \rfloor - m + 1)$ strings; and it follows directly from Lemma 2.3 that no such long-short pair of strings can coexist in the same n -node binary tree.

The bound on vertex-degrees in U is immediate by calculation: a vertex in U has at most 1 edge “entering it from above”, at most 2 edges “leaving it to below”, at most 4 edges “entering it from below”, and at most 2 edges “leaving it to above”. The edges of U are of course undirected, but the suggestive “entering, leaving, above, below” should be helpful in following the enumeration. \square

The reader can verify easily that, at the cost of (at most) doubling the number of nodes in U , one can make U a perfect ordered universal graph, i.e., a graph for which every ordered n -node binary tree appears as an ordered subgraph.

Our results on binary trees can be easily generalized to the family $T_n^{(k)}$ for $k \geq 3$.

THEOREM 2.6. *For each integer n ,*

$$\text{Perf}(\mathbf{T}_n^{(k)}) = (k+1-(n \bmod 2)) \exp k(\lfloor (n-1)/2 \rfloor) - 1 .$$

Moreover, there is a perfect-universal graph for n -node k -ary trees having just this many vertices and having vertex-degree $\leq 1+2k+k^2$.

Proof. Lemmas 2.1, 2.2, 2.3 hold for k -ary trees. Therefore, for n odd we have $\text{Perf}(\mathbf{T}_n^{(k)}) \geq (k^{l+1}-1)/(k-1)$ where $l = \lfloor (n-1)/2 \rfloor$, since any two strings of length $\leq l$ can coexist in the same $(\leq n)$ -node k -ary tree. For n even, we need an additional k^l vertices in the perfect-universal graph for $\mathbf{T}_n^{(k)}$, since any string of length $l+1$ can coexist with any of the strings of length $\leq l$, and two strings of length $l+1$ can coexist in an n -node k -ary tree only if they start with the same symbol. Thus we have

$$\text{Perf}(\mathbf{T}_n^{(k)}) \geq (k+1-(n \bmod 2))k^l - 1 .$$

To prove the equality we need a perfect hash function α which can be defined the same way as that in Theorem 2.5 except for replacing (3) by (3').

(3') **else if** x is of the form $\sigma x'$ for $\sigma \in [k]$, **and if** $|x| = \lfloor (n-1)/2 \rfloor + m$ for $1 \leq m \leq \lfloor n/2 \rfloor$, **then**

$$\alpha(x) = (\sigma+1 \pmod k) \cdot \text{suffix}(x; \lfloor n/2 \rfloor - m) .$$

It is straightforward to check that α is indeed a well-defined perfect hash function. The bound on vertex-degree in the perfect universal graph U can be calculated as follows: a vertex in U has at most 1 edge "entering it from above", at most k edges "leaving it to below", at most k^2 edges "entering it from below", and at most k edges "leaving it to above". Therefore, U has vertex-degree no more than $1+2k+k^2$. \square

3. Perfect universal graphs for chaotic arrays. The main result of this section is that no material compaction of chaotic arrays is possible as it was with binary trees. In particular, $\text{Size}(\mathbf{C}_n) = n(n+1)/2$; and any perfect-universal graph for \mathbf{C}_n must have at least roughly half this number of vertices. The (unrestricted) universal graph for \mathbf{C}_n has $O(n)$ vertices and $O(n^{3/2})$ edges [2]. (In fact, this is the bound for the universal graph for the family of planar graphs on n vertices.)

Notation. Let $p = \langle p_1, p_2 \rangle$ be an ordered pair of nonnegative integers. The pair p is said to have *size* $\Sigma(p) = p_1 + p_2$. When p is viewed as a position of a chaotic array, it is said to *reside at level* $\Sigma(p)$ of the chaotic array. If $q = \langle q_1, q_2 \rangle$ is another pair of nonnegative integers, then we denote by $M(p, q)$ the third pair

$$M(p, q) = \langle \max(p_1, q_1), \max(p_2, q_2) \rangle .$$

A. The basic lemma.

LEMMA 3.1. *The integer pairs p and q can coexist in the same $(\leq n)$ -position chaotic array if and only if $\Sigma(M(p, q)) < n$.*

Proof. Sufficiency. Consider the graph whose vertex set consists of lattice points encountered in the following walk: start at the origin and proceed as directly as possible to the point $\langle \mu, \nu \rangle$, where

$$\mu = \min(p_1, q_1)$$

and

$$\nu = \min(p_2, q_2) .$$

Now go as directly as possible from $\langle \mu, \nu \rangle$ to p , backtrack to $\langle \mu, \nu \rangle$ using no new vertices, and go from $\langle \mu, \nu \rangle$ as directly as possible to q . A straightforward calculation shows that the number of distinct points encountered along this walk is

$$1 + \mu + \nu + (p_1 - \mu) + (q_1 - \mu) + (p_2 - \nu) + (q_2 - \nu).$$

But this quantity is just $1 + \Sigma(M(p, q))$, which by assumption is at most n . Therefore, if we take this set of lattice points and augment it by a set of edges using the rules determining the edges in chaotic arrays, then we find that we have a $(\leq n)$ -position chaotic array holding both p and q . The sufficiency of the lemma's condition follows.

Necessity. Necessity of the lemma's condition is obvious, for any path including the "origin" $\langle 0, 0 \rangle$ and both p and q must encounter at least

$$1 + \max(p_1, q_1)$$

distinct lattice points while "moving up" and another

$$\max(p_2, q_2)$$

while "moving across" (viewing the points as first-quadrant lattice points). By the order-closure of the set of vertices of a chaotic array, any two points in a chaotic array are connected to the origin via a path encountering only the points in the array. Hence the number of points encountered cannot exceed n in number, since p and q are assumed to reside in the same n -position chaotic array. \square

B. The lower bound.

THEOREM 3.2. *For all integers n ,*

$$\text{Perf}(C_n) \geq \lfloor n/2 \rfloor (\lfloor n/2 \rfloor + 1).$$

Proof. By Lemma 3.1, two points p and q can coexist in the same n -position chaotic array whenever $\Sigma(M(p, q)) < n$. It follows that a perfect hash function α cannot identify any two points p and q for which

$$\max(p_i, q_i) < \lfloor n/2 \rfloor$$

and

$$\max(p_j, q_j) \leq \lfloor n/2 \rfloor$$

where $\{i, j\} = \{1, 2\}$. Therefore, the image space of α must contain enough points to give all these pairs distinct images. \square

C. The upper bound. As was the case with trees, we establish here an upper bound on the perfect number for chaotic arrays that coincides exactly with the lower bound.

THEOREM 3.3. *For each integer n ,*

$$\text{Perf}(C_n) \leq \lfloor n/2 \rfloor (\lfloor n/2 \rfloor + 1).$$

Moreover, there is a perfect-universal graph for n -position chaotic arrays having just this number of vertices and having vertex-degree ≤ 5 .

Proof. The graph U . The universal graph U will have for vertices the set

$$\lfloor \lfloor n/2 \rfloor \rfloor \times \lfloor \lfloor n/2 \rfloor + 1 \rfloor.$$

U 's edges will be induced by the allocation function α as follows: there will be an edge connecting vertices v and v' of U just when $\alpha^{-1}(v)$ contains a point $p \in N \times N$ and $\alpha^{-1}(v')$ contains either $p + \langle 0, 1 \rangle$ or $p + \langle 1, 0 \rangle$. It remains only to describe and validate the function α .

Allocation. The allocation function α is defined by cases:

(1) if $p \in [\lfloor n/2 \rfloor] \times [\lfloor n/2 \rfloor + 1]$, then

$$\alpha(p) = p;$$

(2) if $p_1 \geq \lfloor n/2 \rfloor$, then

$$\alpha(p) = \langle p_2, n - p_1 \rangle;$$

(3) if $p_2 > \lfloor n/2 \rfloor$, then

$$\alpha(p) = \langle n - p_2, p_1 \rangle.$$

Verification. We must show that the function α is both well defined and one-to-one on n -position chaotic arrays. Both tasks are immediate by Lemma 3.1. By hypothesis, each p in the domain of α resides in some n -position chaotic array; hence, $\Sigma(p) < n$. Thus, if p_1 (resp., p_2) is big, in the sense of case (2) (resp., (3)) above, then neither of p_2 (resp., p_1) or $n - p_1$ (resp., $n - p_2$) can be big. It follows that the mapping α is well defined in the sense that it maps positions of n -position chaotic arrays into vertices of U . Now, each vertex v of U is the image of either one or two chaotic array positions. If v receives only one position, then it cannot prevent α from being one-to-one. If v receives two positions, then one has the form $\langle q_1, q_2 \rangle$ where $q_1 < \lfloor n/2 \rfloor$ and $q_2 \leq \lfloor n/2 \rfloor$, and the other has the form $\langle p_1, p_2 \rangle$ where either $p_1 = q_2$ and $p_2 = n - q_1$, or vice-versa. In either case, $\Sigma(M(p, q)) = n$, so p and q cannot coreside in the same n -position chaotic array. Thus α is one-to-one on all such chaotic arrays and so is a witness to U 's being a perfect-universal graph for such arrays, as was claimed. It can be easily verified that most of the vertices in U have degree ≤ 4 . Only those vertices $\langle p_1, p_2 \rangle$ with $p_1 \in \{\lfloor n/2 \rfloor - 1, \lfloor n/2 \rfloor\}$ or $p_2 \in \{\lfloor n/2 \rfloor, \lfloor n/2 \rfloor + 1\}$ are of degree ≤ 5 . \square

4. Perfect universal graphs for ragged arrays. Although ragged arrays seem to be closer to rectangular than to chaotic arrays in terms of the amount of uniformity in their structure, they behave for the purposes of our study much more like chaotic arrays. Specifically, we shall see in the next section that rectangular arrays have a perfection number of n . In contrast, we have seen in the last section that chaotic arrays have a perfection number that is only half of the number of vertices in the most naive possible universal graph for chaotic arrays. We shall see now that ragged arrays' perfection number is roughly $n^2/6$, while $\text{Size}(\mathbf{R}_n) = n(n+1)/2$.

A. The basic lemma.

LEMMA 4.1. Let $p = \langle p_1, p_2 \rangle$ and $q = \langle q_1, q_2 \rangle$ be integer pairs with $p_1 \neq q_1$. The pairs p and q can coexist in the same ($\leq n$)-position ragged array if and only if

$$\max(p_1, q_1) + p_2 + q_2 < n.$$

Proof. Sufficiency. The ragged array with vertices

$$([\max(p_1, q_1) + 1] \times \{0\}) \cup (\{p_1\} \times [p_2 + 1]) \cup (\{q_1\} \times [q_2 + 1])$$

contains both $\langle p_1, p_2 \rangle$ and $\langle q_1, q_2 \rangle$ and has precisely $\max(p_1, q_1) + p_2 + q_2 + 1$ vertices.

Necessity. Follows directly from the fact that, by definition, if the point $\langle p_1, p_2 \rangle$ is in the ragged array R , then $([p_1 + 1] \times \{0\}) \cup (\{p_1\} \times [p_2 + 1]) \subseteq R$. \square

B. The lower bound.

THEOREM 4.2. *For all integers n ,*

$$\text{Perf}(\mathbf{R}_n) \geq ([n/3] + 1)(3[2n/3] - n)/2.$$

Proof. We consider the set S of all points (x_1, x_2) satisfying

$$x_1 + x_2 < [2/3 n], 0 \leq x_2 \leq [n/3], 0 \leq x_1.$$

There are $([n/3] + 1)(3[2n/3] - n)/2$ such points. Suppose (p_1, p_2) and (q_1, q_2) are in S and $p_1 < q_1$. Then

$$\max(p_1, q_1) + p_2 + q_2 \leq p_2 + q_1 + q_2 < n.$$

Thus by Lemma 4.1 any two points in S can coexist in the same ($\leq n$)-position ragged array and a perfect hash function α cannot identify any two points in S . Therefore we have

$$\text{Perf}(\mathbf{R}_n) \geq |S| \geq ([n/3] + 1)(3[2n/3] - n)/2. \quad \square$$

C. The upper bound. We will establish here an upper bound on the perfect number for ragged arrays that coincides exactly with the lower bound.

THEOREM 4.3. *For each integer n ,*

$$\text{Perf}(\mathbf{R}_n) = ([n/3] + 1)(3[2n/3] - n)/2.$$

Moreover, there is a perfect-universal graph for n -position ragged arrays having just this number of vertices and having vertex degree ≤ 16 .

Proof. Consider the graph U with vertex set S as defined in Theorem 4.2. The edges of U will be induced by the allocation function α which maps points in $\{(x_1, x_2) : 0 \leq x_1, x_2, 0 \leq x_1 + x_2 < n\}$ to S as defined by cases as follows:

- (1) If $p \in S$, then $\alpha(p) = p$.
- (2) If $x_1 + x_2 \geq [2/3 n]$, $x_1 > [n/3]$, then

$$\alpha(x_1, x_2) = (x_2, n - x_1 - x_2).$$

- (3) If $x_2 > [n/3]$ and $x_1 \leq [n/3]$, then

$$\alpha(x_1, x_2) = (n - x_1 - x_2, x_1).$$

It is straightforward to verify that the function α is well-defined in the sense that it maps positions of n -position ragged arrays into vertices of U . Now, each vertex v of U is the image of at most three ragged array positions. It can be easily checked that α is one-to-one on any n -position ragged array using Lemma 4.1. A vertex $p = (p_1, p_2)$ in U is adjacent to vertices $(p_1 + \epsilon, p_2 + \epsilon')$ for any $\epsilon, \epsilon' \in \{0, 1, -1\}$ and ϵ, ϵ' not both 0 if $(p_1 + \epsilon, p_2 + \epsilon')$ is in $V(U) - \{p\}$. In fact most vertices of U have degree 8 except for a few with degree ≤ 16 . \square

5. Perfect universal graphs for rectangular arrays. This section contains two main results. First, in common with trees, rectangular arrays admit significant compaction: $\text{Size}(A_n)$ is roughly $n \log n$ [12], yet there is a perfect-universal graph for A_n having only n vertices. Second, although the universal graphs just mentioned have vertex-degrees that are not bounded, independent of n , there are perfect-universal graphs for A_n having only $2n$ vertices whose vertex-degrees do not exceed 4.

We shall present the initial results about perfect-universal graphs for A_n in a somewhat cursory manner, since these graphs were studied under a different guise by Rosenberg and Stockmeyer [15].

A. The basic lemma.

LEMMA 5.1. [15] (a) *The point $\langle p_1, p_2 \rangle \in N \times N$ resides in some $(\leq n)$ -position rectangular array if and only if*

$$(p_1+1)(p_2+1) \leq n.$$

(b) *The points $p = \langle p_1, p_2 \rangle$ and $q = \langle q_1, q_2 \rangle$ can coexist in the same $(\leq n)$ -position rectangular array if and only if the point $M(p, q)$ resides in some n -position rectangular array.*

B. Upper and lower bounds.

THEOREM 5.2. [15] *For each integer n ,*

$$\text{Perf}(A_n) = n.$$

Proof. The lower bound on Perf being immediate by the pigeon-hole principle, we turn to the upper bound.

The graph U . Fix on n , let the graph U have for vertices the set $[n]$, and let U 's edges be induced by the allocation function

$$\alpha: \{p \in N \times N \mid (p_1+1)(p_2+1) \leq n\} \rightarrow \text{Vertices}(U)$$

as in the previous sections.

Allocation. Define the function

$$\alpha: \{p \in N \times N \mid (p_1+1)(p_2+1) \leq n\} \rightarrow [n]$$

as follows:

- (1) For each $m \in [n]$, $\alpha(\langle m, 0 \rangle) = m$;
- (2) for each $m \in [n] - \{0\}$, α "assigns" to the points in $[\lfloor n/(m+1) \rfloor] \times \{m\}$ the first $\lfloor n/(m+1) \rfloor$ integers in increasing order in the set $[n] - \alpha([\lfloor n/(m+1) \rfloor] \times \{m\})$.

Verification. The fact that α is indeed an allocation function for a perfect-universal graph for the family A_n of rectangular arrays follows immediately from the proof in [15] that the function α is one-to-one on all rectangular arrays having n or fewer positions. \square

C. A bounded-degree perfect-universal graph for A_n Although the perfect-universal graph constructed in Theorem 5.2 is optimal in size, it is deficient in one major respect: the maximum degree of the vertices of the graph grows with the size of the array-graphs being imbedded. We believe that this growth is inevitable, but we have been unable to verify or refute the following.

CONJECTURE. *Let the family of graphs U_1, U_2, \dots , each $|U_n| = n$, be perfect-universal for the collections A_1, A_2, \dots , respectively. There is no constant c such that*

every graph U_n has vertex-degree $\leq c$.

We do know, however, that one can attain the $(n \log n)$ -to- n compactification of Theorem 5.2 together with bounded degrees if one is willing to suffer a modest increase in the number of vertices in the perfect-universal graph.

THEOREM 5.3. *For each integer n , there is a perfect-universal graph U for A_n having $|U| = 2n$ vertices and vertex-degree 4.*

Proof. We describe the graph U explicitly.

Vertices. The graph U has the vertex-set

$$\text{Vertices}(U) = \{p \in N \times N \mid 0 \leq \text{both}(p) < \lfloor \sqrt{n} \rfloor \text{ or } \lfloor \sqrt{n} \rfloor \leq \text{both}(p) < 2\lfloor \sqrt{n} \rfloor\}$$

where references to $\text{both}(p)$ indicate that the inequalities govern both p_1 and p_2 . Thus, when pictured as a plane set, U looks like two square blocks joined at one corner; see Figure 4(a).

As usual, we shall let the edges of U be induced by our specification of the allocation function

$$\alpha: \{p \in N \times N \mid (p_1+1)(p_2+1) \leq n\} \rightarrow \text{Vertices}(U).$$

Allocation. The function α will be symmetric in the sense that, if $\alpha(\langle p_1, p_2 \rangle) = \langle a_1, a_2 \rangle$, then $\alpha(\langle p_2, p_1 \rangle) = \langle a_2, a_1 \rangle$; hence, in defining $\alpha(p)$, we shall assume with no loss of generality that $p_2 < \lfloor \sqrt{n} \rfloor$ (since by Lemma V.1, at least one of p_1, p_2 must be this small). Let $p \in N \times N$ satisfy the following conditions.

- (1) $(p_1+1)(p_2+1) \leq n$;
- (2) $p_2 < \lfloor \sqrt{n} \rfloor$;
- (3) $\lfloor p_1 / (2\lfloor \sqrt{n} \rfloor) \rfloor = \sum_k \delta_k 2^k$.

Then, letting

$$\pi_1 = p_1 \bmod 2\lfloor \sqrt{n} \rfloor,$$

and

$$\pi_2 = p_2 + \sum_k \delta_k \lfloor \sqrt{n} \rfloor / 2^{k+1}$$

we have

$$\alpha(p) = \text{if } \pi_1 < \lfloor \sqrt{n} \rfloor \text{ then } \langle \pi_1, \pi_2 \rangle \text{ else } \langle \pi_1, \pi_2 + \lfloor \sqrt{n} \rfloor \rangle.$$

Verification. We have three things to verify: that the function α is well-defined, that it is a perfect hash function for A_n , and that the resulting graph U has vertex-degrees ≤ 4 . We treat each issue in turn.

First, let

$$a = \lfloor \log_2 \lfloor p_1 / (2\lfloor \sqrt{n} \rfloor) \rfloor \rfloor.$$

By condition (1), we must have

$$p_2 < n / (2^{a+1} \lfloor \sqrt{n} \rfloor + 1).$$

It follows, therefore, that $\pi_2 < \lfloor \sqrt{n} \rfloor$; moreover, it is immediate by definition that π_1 , which is the first coordinate of $\alpha(p)$, is less than $2\lfloor \sqrt{n} \rfloor$. Hence, when $\pi_1 < \lfloor \sqrt{n} \rfloor$, both coordinates of $\alpha(p)$ are nonnegative but less than $\lfloor \sqrt{n} \rfloor$; and when $\pi_1 \geq \lfloor \sqrt{n} \rfloor$, both coordinates of $\alpha(p)$ are at least $\lfloor \sqrt{n} \rfloor$ but less than $2\lfloor \sqrt{n} \rfloor$. In other words, for every vertex p of a $(\leq n)$ -position rectangular array, $\alpha(p) \in \text{Vertices}(U)$.

Second, assume that there are distinct pairs p and q of the right form such that $\alpha(p) = \alpha(q)$. Trivially, we must have $p_1 \neq q_1$, since p and q are assumed to be

distinct. Assume that $q_1 > p_1$, and let

$$a = \lfloor \log_2 \lfloor q_1 / (2 \lfloor \sqrt{n} \rfloor) \rfloor \rfloor.$$

Since α identifies p and q , it must be that

$$p_2 \geq \lfloor \sqrt{n} \rfloor / 2^{a+1}.$$

But simple calculation now demonstrates that

$$(p_2+1)(q_1+1) > n$$

so that the point $M(p, q)$ cannot reside in any n -position rectangular array (by Lemma 5.1a), so the points p and q cannot coreside in the same n -position such array (by Lemma 5.1b). It follows that the function α is a perfect hash function for A_n , as was claimed.

It is easy to verify that the vertex degrees in U are bounded by 4 since any edge e of U is in one of the following types (see Fig. 4):

- (1) $e = \{ \langle p_1, p_2 \rangle, \langle p_1, p_2+1 \rangle \}$;
- (2) $e = \{ \langle p_1, \lfloor \sqrt{n} \rfloor - 1 \rangle, \langle p_1 + \lfloor \sqrt{n} \rfloor, \lfloor \sqrt{n} \rfloor \rangle \}$ where p_1 satisfies $a \leq p_1 < \lfloor \sqrt{n} \rfloor$;
- (3) $e = \{ \langle p_1, 0 \rangle, \langle p_1, 2 \lfloor \sqrt{n} \rfloor - 1 \rangle \}$ where

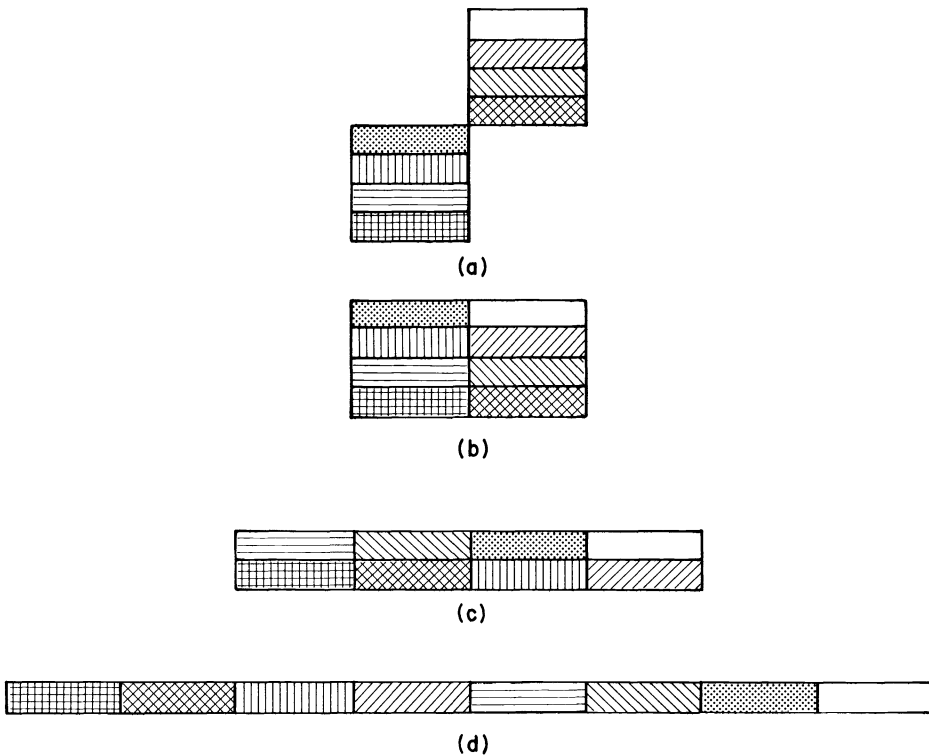


Fig. 4. Illustrating how the $2n$ -vertex degree-4 perfect-universal graph for rectangular arrays "covers" short wide arrays: (a) a schematic view of the graph U ; (b) U covering arrays with $\leq 2\lfloor \sqrt{n} \rfloor$ columns and $\lfloor \sqrt{n} \rfloor$ rows; (c) U covering arrays with $\leq 4\lfloor \sqrt{n} \rfloor$ columns and $0.5\lfloor \sqrt{n} \rfloor$ rows; (d) U covering arrays with $\leq 8\lfloor \sqrt{n} \rfloor$ columns and $0.25\lfloor \sqrt{n} \rfloor$ rows.

$$p_1 = \sum_k \delta_k \lfloor \sqrt{n} \rfloor / 2^{k+1},$$

$$p_2 = \sum_k \delta'_k \lfloor \sqrt{n} \rfloor / 2^{k+1},$$

$$\sum \delta_k 2^k + 1 = \sum \delta'_k 2^k.$$

(4) $e = \{ \langle p_1, p_2 \rangle, \langle q_1, q_2 \rangle \}$ where

$e' = \{ \langle p_2, p_1 \rangle, \langle q_2, q_1 \rangle \}$ is of type 1, 2, or 3.

6. Summary. We summarize our results in Table 1.

TABLE 1

	Size	The order of the universal graph	The order of the perfect universal graph
T_n	$2^n - 1$	$O(n \log n)$	$(3 - (n \bmod 2))2^{\lfloor (n-1)/2 \rfloor} - 1$
$T_n^{(k)}$	$\frac{k^n - 1}{k - 1}$	$O(n \log n)$	$(k + 1 - (n \bmod 2))k^{\lfloor (n-1)/2 \rfloor} - 1$
C_n	$n(n+1)/2$	$O(n^{3/2})$	$\lfloor n/2 \rfloor (\lfloor n/2 \rfloor + 1)$
R_n	$n(n+1)/2$	$O(n^{3/2})$	$(\lfloor n/3 \rfloor + 1)(3 \lfloor 2n/3 \rfloor - n)/2$
A_n	$n \log n$	n	n

where

T_n : The family of $(\leq n)$ -position binary trees,

$T_n^{(k)}$: The family of $(\leq n)$ -position k -ary trees,

C_n : The family of $(\leq n)$ -position chaotic arrays,

R_n : The family of $(\leq n)$ -position ragged arrays,

A_n : The family of $(\leq n)$ -position rectangular arrays.

REFERENCES

[1] R. ALELIUNAS and A. L. ROSENBERG, *On embedding rectangular grids in square grids*, IEEE Trans. Computers, C-31 (1982), pp. 907-913.
 [2] L. BABAI, F. R. K. CHUNG, P. ERDOS, R. L. GRAHAM and J. H. SPENCER, *On graphs which contain all sparse graphs*, Annals of Discrete Math., 12 (1982), pp. 21-26.
 [3] F. R. K. CHUNG, D. COPPERSMITH and R. L. GRAHAM, *On trees containing all small trees*, in The Theory and Applications of Graphs, G. Chartrand, ed., John Wiley, New York, 1981, pp. 265-272.

- [4] F. R. K. CHUNG and R. L. GRAHAM, *On graphs which contain all small trees*, J. Comb. Th.(B) 24 (1978), pp. 14-23.
- [5] ———, *On universal graphs for spanning trees*. Typescript, 1979.
- [6] F. R. K. CHUNG, R. L. GRAHAM, and N. J. PIPPENGER, *On graphs which contain all small trees, II.*, Proc. 1976 Hungarian Colloq. on Combinatorics, North-Holland, Amsterdam, 1978, pp. 213-223.
- [7] R. A. DeMILLO, S. C. EISENSTAT and R. J. LIPTON, *On small universal data structures and related combinatorial problems*, Proc. John Hopkins Conference on Information Sciences and System, 1978, pp. 408-411.
- [8] J.-W. HONG, K. MEHLHORN and A. L. ROSENBERG, *Cost tradeoffs in graph embeddings, with applications*, J. ACM, to appear.
- [9] R. J. LIPTON, A. L. ROSENBERG, A. C. YAO, *External hashing schemes for collections of data structures*, J. ACM, 27 (1980), pp. 81-95.
- [10] R. J. LIPTON, S. C. EISENSTAT and R. A. DeMILLO, *Storage hierarchies for classes of control structures and data structures*, J. ACM, 23, (1976), pp. 720-732.
- [11] A. L. ROSENBERG, *Data graphs and addressing schemes*, J. Comp. Syst. Sci., 5 (1971), pp. 193-238.
- [12] ———, *Managing storage for extendible arrays*, this Journal, (1975), pp. 287-306.
- [13] ———, *Data encodings and their costs*, Acta Inform, 9 (1978), pp. 273-292.
- [14] A. L. ROSENBERG, L. SNYDER and L. J. STOCKMEYER, *Uniform data encodings*, Theor. Comp. Sci., 11 (1980), pp. 145-165.
- [15] A. L. ROSENBERG and L. J. STOCKMEYER, *Storage schemes for boundedly extendible arrays*, Acta Inform., 7 (1977), pp. 289-303.
- [16] R. SPRUGNOLI, *Perfect hashing functions: a single probe retrieving method for static sets*, C. ACM, 20 (1977), pp. 841-850.
- [17] L. VALIANT, *Universality considerations in VLSI circuits*, IEEE Trans. Computers, C-30 (1981), 135-140.