SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0001 $01.25/0

AGGREGATION THEOREMS FOR ALLOCATION PROBLEMS*

J. ACZÉL†, C. T. NG† AND C. WAGNER‡

Abstract. Suppose that n individuals assign values to a sequence of m numerical decision variables subject to the constraints that the m values assigned by each individual must be nonnegative and sum to some fixed positive σ. Suppose that we wish to aggregate their individual assignments to produce consensual values of these variables satisfying the aforementioned constraints. Aczél and Wagner have shown that if $m \geqq 3$, then a method of aggregation is based on weighted arithmetic averaging iff (a) the consensual value assigned to each variable depends only on the values assigned by individuals to that variable and (b) the consensual value is zero if all individuals assign that variable the value zero. In the present paper we extend this result in various ways, dropping the unanimity condition (b) and allowing individual and consensual values to be restricted to some subinterval of $[0, \sigma]$.

1. Suppose that a group of n individuals wish to assign values to a sequence of m numerical decision variables. We call such a problem an *allocation problem* if the values assigned must be nonnegative and sum to some fixed positive number σ. Examples of allocation problems abound, including, for example, the assignment of probabilities to a sequence of pairwise disjoint, exhaustive events, and the distribution of a fixed sum of money or other resource σ among m projects.

In general we may expect that individuals will differ in the values that they assign to the variables, and hence be faced with the problem of aggregating their individual assignments to produce consensual values of these variables. Let us denote by the n-dimensional vector z_j the sequence of values assigned by the individuals to the jth variable. (In what follows, lower case Latin letters, other than subscripts and integers describing dimensions, denote vectors, while Greek letters denote real numbers. We abbreviate the vector $(\alpha, \alpha, \cdots, \alpha)$, with equal components, by $\boldsymbol{\alpha}$.) If $m \geqq 3$ a method of aggregation assigns consensual values to decision variables in such a way that (a) the consensual value assigned to the jth variable is $\Phi_j(z_j)$, where $\Phi_j: [0, \sigma]^n \to [0, \sigma]$ and (b) the consensual value is zero if all individuals assign that variable the value zero iff the method is based on weighted arithmetic averaging, with weights invariant over the m decision variables ([7, Thm. 6.4]; see also [2], [3], [4], [9]). In many decisionmaking situations, individual and consensual values of the variables may be constrained to lie in a proper subinterval of $[0, \sigma]$, as, for example, when each of a number of budgetary units must receive not less than a minimal allocation $\mu > 0$, nor more than a maximal allocation $\nu < \sigma$. In such cases aggregation is more appropriately modeled by functions $\Phi_j: I^n \to I$, where $I \subset [0, \sigma]$. Furthermore, the assigned values may be deviations from some preferred value, and thus some may be negative (cf. [2]). So we may wish also to drop the condition $\mu > 0$ (or $\mu \geqq 0$) and let I be any (finite) real interval. In addition, conditions like (b) are only plausible if aggregation is carried out "internally" among the n individuals. If these individuals functioned as advisors to some external decisionmaker and he was responsible for aggregation, he might very well decide to ignore their unanimity. We are thus motivated in the present

* Received by the editors March 19, 1982, and in final revised form February 10, 1983.

† Centre for Information Theory, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. These authors' research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

‡ Mathematics Department, University of Tennessee, Knoxville, Tennessee 37996-1300. This author's work was supported in part by a Summer Fellowship at the Center for Advanced Study in the Behavioral Sciences.

paper to generalize, as described below, the model of aggregation specified by (a) and (b), and to characterize the methods of aggregation which accord with this more general model.

2. With the above observations in mind, we first model aggregation as follows: Let σ be a fixed constant, and I be an interval with end points $\mu < \nu$ compatible with σ in the sense that

$$(1)\qquad \begin{array}{lll} & I=[\mu,\nu] & \text{and}\quad (m-1)\mu+\nu \leqq \sigma \leqq \mu+(m-1)\nu, \\ \text{or} & I=[\mu,\nu[& \text{and}\quad (m-1)\mu+\nu \leqq \sigma < \mu+(m-1)\nu, \\ \text{or} & I=]\mu,\nu] & \text{and}\quad (m-1)\mu+\nu < \sigma \leqq \mu+(m-1)\nu, \\ \text{or} & I=]\mu,\nu[& \text{and}\quad (m-1)\mu+\nu \leqq \sigma \leqq \mu+(m-1)\nu. \end{array}$$

We suppose that there exist bounded functions

$$(2)\qquad \Phi_j : I^n \to \mathbb{R} \qquad (j=1,2,\cdots,m)$$

(in fact, it will suffice to assume merely that at least one Φ_{j_0} is bounded below on some proper rectangle $[\gamma_1,\delta_1]\times[\gamma_2,\delta_2]\times\cdots\times[\gamma_n,\delta_n]\subseteq I^n$) such that

$$(3)\qquad \left(z_j \in I^n \text{ and } \sum_{j=1}^{m} z_j = \boldsymbol{\sigma}\right) \Rightarrow \sum_{j=1}^{m} \Phi_j(z_j) = \sigma.$$

To motivate the compatibility conditions in (1) for the study of (3), we observe that, in general the set $S=\{z_1 \in I^n \mid \exists z_2,\cdots,z_m \in I^n \text{ such that } \sum_{j=1}^{m} z_j = \boldsymbol{\sigma}\}$ is a subinterval (n-dimensional) of I^n and (3) provides information about Φ_j only on S. Thus it is natural to assume that the domain I^n of Φ_j is equal to S, which is equivalent to (1). For instance, if $I=]\mu,\nu[$, then $z_1=\boldsymbol{\sigma}-\sum_{j=2}^{m} z_j < \boldsymbol{\sigma}-(m-1)\boldsymbol{\mu}$ (inequality meant componentwise). But z_1 can be anywhere in $]\boldsymbol{\mu},\boldsymbol{\nu}[$, so it is natural to assume $\nu \leqq \sigma-(m-1)\mu$. The argument is similar for $\mu \geqq \sigma-(m-1)\nu$ and for the other cases in (1). Note that (1) implies

$$(4)\qquad \mu < \frac{1}{m}\sigma < \nu.$$

The following theorem characterizes the aggregation methods which satisfy (2) and (3) when there are at least three decision variables ($m \geqq 3$):

THEOREM 1. *For fixed $m \geqq 3$, an aggregation method satisfies* (2) *and* (3) *if, and only if, there exist real numbers $\omega_1,\omega_2,\cdots,\omega_n$ and $\beta_1,\beta_2,\cdots,\beta_m$, with*

$$(5)\qquad \sum_{j=1}^{m} \beta_j = \left(1-\sum_{i=1}^{n} \omega_i\right)\sigma,$$

such that, for all $z=(\zeta_1,\cdots,\zeta_n)\in I^n$,

$$(6)\qquad \Phi_j(z)=\sum_{i=1}^{n} \omega_i\zeta_i+\beta_j \qquad (j=1,\cdots,m).$$

The ω_i $(i=1,2,\cdots,n)$ will be called *weights*.

Proof. Sufficiency is clear. To prove necessity we bring in new intervals, variables and functions. On first thought, one would want to move $\boldsymbol{\mu}$ into $\mathbf{0}$ and thus go back to the problem previously considered (as described in § 1). However, the intervals *open at* 0 ($]\mathbf{0}, \boldsymbol{\nu}-\boldsymbol{\mu}[$ or $]\mathbf{0}, \boldsymbol{\nu}-\boldsymbol{\mu}]$) would lead to complications and also the necessary extension (cf. Appendix) is easier if $\mathbf{0}$ is in the *interior* of the new domain. So we define

$$\tilde{I} = I - \frac{1}{m}\sigma, \qquad \tilde{z}_j = z_j - \frac{1}{m}\boldsymbol{\sigma}, \tag{7}$$
$$\tilde{\Phi}_j(\tilde{z}_j) = \Phi_j\left(\tilde{z}_j + \frac{1}{m}\boldsymbol{\sigma}\right) = \Phi_j(z_j) \qquad (j = 1, \cdots, m)$$

in order to transform (3) into

$$\left(\tilde{z}_j \in \tilde{I}^n \text{ and } \sum_{j=1}^m \tilde{z}_j = \mathbf{0}\right) \Rightarrow \sum_{j=1}^m \tilde{\Phi}_j(\tilde{z}_j) = \sigma. \tag{8}$$

Note that $\tilde{I}$ contains 0 as an interior point because of (4). Putting $\tilde{z}_j = \mathbf{0}$ $(j = 1, \cdots, m)$ we obtain $\sum_{j=1}^m \tilde{\Phi}_j(\mathbf{0}) = \sigma$ and so the functions defined by

$$\psi_j(\tilde{z}) = \tilde{\Phi}_j(\tilde{z}) - \tilde{\Phi}_j(\mathbf{0}) \tag{9}$$

satisfy

$$\left(\tilde{z}_j \in \tilde{I}^n \text{ and } \sum_{j=1}^m \tilde{z}_j = \mathbf{0}\right) \Rightarrow \sum_{j=1}^m \psi_j(\tilde{z}_j) = 0, \qquad \psi_j(\mathbf{0}) = 0 \qquad (j = 1, \cdots, m). \tag{10}$$

We first consider (10) within a fixed symmetric subinterval $[-\varepsilon, \varepsilon] \subseteq \tilde{I}$ (with $\varepsilon > 0$). In particular we get

$$\left(\tilde{z}_j \in [-\varepsilon, \varepsilon]^n \text{ and } \sum_{j=1}^m \tilde{z}_j = \mathbf{0}\right) \Rightarrow \sum_{j=1}^m \psi_j(\tilde{z}_j) = 0, \qquad \psi_j(\mathbf{0}) = 0. \tag{11}$$

Putting $\tilde{z}_1 = y$, $\tilde{z}_2 = -y$, $\tilde{z}_3 = \cdots = \tilde{z}_m = \mathbf{0}$ in (11) we get $\psi_1(y) + \psi_2(-y) = 0$ on $[-\varepsilon, \varepsilon]^n$. Similarly $\psi_j(y) + \psi_k(-y) = 0$ on $[-\varepsilon, \varepsilon]^n$ for all $j \neq k$. This implies (as $m \geqq 3$)

$$\psi_1 = \psi_2 = \cdots = \psi_m = \text{an odd function } \psi \text{ on } [-\varepsilon, \varepsilon]^n. \tag{12}$$

In view of (12), we get from (11), with $\tilde{z}_1 = x$, $\tilde{z}_2 = y$, $\tilde{z}_3 = -x - y$, $\tilde{z}_4 = \cdots = \tilde{z}_m = \mathbf{0}$, the Cauchy equation

$$\psi(x) + \psi(y) = \psi(x + y), \qquad x, y, x + y \in [-\varepsilon, \varepsilon]^n. \tag{13}$$

This ψ can be extended uniquely to a function $\bar{\psi}$: $\mathbb{R}^n \to \mathbb{R}$ satisfying

$$\bar{\psi}(x) + \bar{\psi}(y) = \bar{\psi}(x + y) \quad \text{all } x, y \in \mathbb{R}^n. \tag{14}$$

This extension theorem is due to Daróczy and Losonczi [5]. (For completeness we include in the appendix a shorter proof, cf. [3], [6], [8]).

We now claim that $\psi_1 = \psi_2 = \cdots = \psi_m = \bar{\psi}$ on $\tilde{I}^n$. We first consider

$$\phi_j(\tilde{z}) = \psi_j(\tilde{z}) - \bar{\psi}(\tilde{z}), \qquad z \in \tilde{I}^n \qquad (j = 1, \cdots, m),$$

and we need to show that $\phi_j = 0$ on $\tilde{I}^n$ for all j. Since $\bar{\psi}$ extends ψ, we already have from (12)

$$\phi_j(x) = 0 \quad \text{for all } x \in [-\varepsilon, \varepsilon]^n. \tag{15}$$

From (10) we get

$$\left(\tilde{z}_j \in \tilde{I}^n \text{ and } \sum_{j=1}^{m} \tilde{z}_j = \mathbf{0}\right) \Rightarrow \sum_{j=1}^{m} \phi_j(\tilde{z}_j) = 0. \tag{16}$$

Let $z \in \tilde{I}^n$ be arbitrarily given and we will show that $\phi_j(z) = 0$ for all j. From (1) and the discussion following (3), there exist $z_2, \cdots, z_m \in \tilde{I}^n$ such that $z + \sum_{j=2}^{m} z_j = \mathbf{0}$, i.e. $\sum_{j=2}^{m} z_j = -z$. Therefore the mean of $z_2, \cdots, z_m$, which equals $-z/(m-1)$ is also in $\tilde{I}^n$. By (16) we have

$$\phi_1(z) + \phi_2\left(\frac{-1}{m-1}z\right) + \phi_3\left(\frac{-1}{m-1}z\right) + \cdots + \phi_m\left(\frac{-1}{m-1}z\right) = 0. \tag{17}$$

Repeating this, using $(-1/(m-1))^l z$ in place of z, we get a sequence of equations

$$\begin{aligned} &\phi_1\left(\left(\frac{-1}{m-1}\right)^{l} z\right) + \phi_2\left(\left(\frac{-1}{m-1}\right)^{l+1} z\right) \\ &\quad + \phi_3\left(\left(\frac{-1}{m-1}\right)^{l+1} z\right) + \cdots + \phi_m\left(\left(\frac{-1}{m-1}\right)^{l+1} z\right) = 0 \end{aligned} \tag{18}$$

for $l = 0, 1, 2, \cdots$. Since (16) is symmetric in the ϕ_j's, (18) remains valid under any permutation of ϕ_j's. For large enough l, $(-1/(m-1))^{l+1}z$ will be in the interval $[-\varepsilon, \varepsilon]^n$ where the ϕ_j's are zero, and so by (18) $\phi_1((-1/(m-1))^l z) = 0$. By symmetry $\phi_j((-1/(m-1))^l z) = 0$ for all $j = 1, \cdots, m$. Using (18) recursively we get $\phi_j((-1/(m-1))^0 z) = \phi_j(z) = 0$ as claimed.

The boundedness (2) implies that $\bar{\psi}$ is bounded below on some rectangle and so (14) yields $\psi_j(z) = \bar{\psi}(z) = \sum_{i=1}^{n} \omega_i \zeta_i$ with appropriate constants (weights) $\omega_1, \cdots, \omega_n$ on $\tilde{I}^n$ (see e.g. [1, pp. 214–216]). This, (9), (7), and (14) give

$$\begin{aligned} \Phi_j(z) - \Phi_j(\mathbf{0}) &= \tilde{\Phi}_j\left(z - \frac{1}{m}\boldsymbol{\sigma}\right) - \tilde{\Phi}_j\left(-\frac{1}{m}\boldsymbol{\sigma}\right) = \bar{\psi}\left(z - \frac{1}{m}\boldsymbol{\sigma}\right) - \bar{\psi}\left(-\frac{1}{m}\boldsymbol{\sigma}\right) \\ &= \bar{\psi}(z) = \sum_{i=1}^{n} \omega_i \zeta_i \end{aligned}$$

that is, (6) holds with $\beta_j = \Phi_j(0)$. The functions Φ_j given by (6) satisfy (3) if and only if the constants satisfy (5). This proves the theorem. □

We note that the "weights" ω_i may be negative. It is easy to show that these weights are nonnegative if and only if $\Phi_j(z) \geqq \Phi_j(x)$ for all $z = (\zeta_1, \cdots, \zeta_n)$ and $x = (\xi_1, \cdots, \xi_n)$ with $\zeta_i \geqq \xi_i$, $i = 1, \cdots, n$. We remark also that *the above theorem continues to hold, without essential change in the proof, if the bounds μ and ν are replaced by possibly different bounds μ_i and ν_i, $1 \leqq i \leqq n$, on each individual's assignments.*

3. We next investigate aggregation methods which supplement the hypothesis of Theorem 1 by (a) conditions requiring that aggregation respect unanimity among the individuals and (b) a narrowing of the range of the functions Φ_j to I, the same interval where individuals assign values to the variables in question.

It turns out that very weak unanimity conditions have rather substantial consequences:

THEOREM 2. *Let $m \geqq 3$ and suppose that an aggregation method satisfies* (2) *and* (3). *If $\Phi_j(\boldsymbol{\sigma}/m) = \sigma/m$ $(j = 1, \cdots, m)$ then, and only then, for all $z = (\zeta_1, \cdots, \zeta_n) \in I^n$,*

$$\Phi_j(z) = \sum_{i=1}^{n} \omega_i \zeta_i + \beta \qquad (j = 1, \cdots, m), \tag{19}$$

where

$$\beta=\left(1-\sum_{i=1}^{n}\omega_i\right)\frac{\sigma}{m}; \tag{20}$$

and if, for some $\alpha\in I$ *with* $\alpha\neq\sigma/m$, $\Phi_j(\boldsymbol{\alpha})=\alpha$, $j=1,\cdots,m$, *then, and only then, for all* $z=(\zeta_1,\cdots,\zeta_n)\in I^n$,

$$\Phi_j(z)=\sum_{i=1}^{n}\omega_i\zeta_i \qquad (j=1,\cdots,m), \tag{21}$$

where $\sum_{i=1}^{n}\omega_i=1$.

Proof. Assume

$$\phi_j(\boldsymbol{\alpha})=\alpha \quad \text{for a fixed } \alpha\in I.$$

From Theorem 1, (6)

$$\alpha=\sum_{i=1}^{n}\omega_i\alpha+\beta_j \qquad (j=1,2,\cdots,m).$$

Thus

$$\beta_1=\beta_2=\cdots=\beta_m=\beta=\left(1-\sum_{i=1}^{n}\omega_i\right)\alpha.$$

Comparison with (5) gives

$$\left(1-\sum_{i=1}^{n}\omega_i\right)(m\alpha-\sigma)=0.$$

So there are two cases. Either $\alpha=\sigma/m$ and then we have (19) and (20). Or $\alpha\neq\sigma/m$, in which case

$$\sum_{i=1}^{n}\omega_i=1.$$

Thus $\beta_1=\beta_2=\cdots=\beta_m=0$, so that (6) goes over into (21). □

We consider next the effect of specifying, in place of (2), the stronger (and more natural) condition

$$\Phi_j\colon I^n\to I, \qquad 1\leqq j\leqq m \tag{22}$$

(in fact, it suffices to posit this range restriction for $j\in J$, where J is some nonempty subset of $\{1,\cdots,m\}$).

THEOREM 3. *If* $m\geqq 3$, *an aggregation method satisfies* (22) *and* (3) *if, and only if, the aggregation functions* Φ_j *are of the form*

$$\Phi_j(z)=\sum_{i=1}^{n}\omega_i\zeta_i+\beta_j \qquad (j=1,2,\cdots,m), \tag{23}$$

where

$$\sum_{j=1}^{m}\beta_j=\left(1-\sum_{i=1}^{n}\omega_i\right)\sigma, \tag{24}$$

and for each $j\in J$ [*where* (22) *is supposed to hold*]

$$\mu-\mu\Sigma^{**}-\nu\Sigma^{*}<_1\beta_j<_2\nu-\nu\Sigma^{**}-\mu\Sigma^{*} \tag{25}$$

where $\Sigma^* = \Sigma^*\omega_i$ *denotes the sum of the negative weights and* $\Sigma^{**} = \Sigma^{**}\omega_i$ *denotes the sum of the positive weights, and the two inequality symbols* $<_1$ *and* $<_2$ *are either* $<$ *or* $\leqq$ *according to the type of the interval* I:

(A) *When* $I = [\mu, \nu]$, $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$.

(B) *When* $I = [\mu, \nu[$,

(i) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* \neq 0$ *and* $\Sigma^{**} \neq 0$,
(ii) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* = 0$ *and* $\Sigma^{**} \neq 0$,
(iii) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $<$ *if* $\Sigma^* \neq 0$ *and* $\Sigma^{**} = 0$,
(iv) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $<$ *if* $\Sigma^* = 0$ *and* $\Sigma^{**} = 0$.

(C) *When* $I =]\mu, \nu]$,

(i) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* \neq 0$ *and* $\Sigma^{**} \neq 0$,
(ii) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* = 0$ *and* $\Sigma^{**} \neq 0$,
(iii) $<_1$ *is* $<$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* \neq 0$ *and* $\Sigma^{**} = 0$,
(iv) $<_1$ *is* $<$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* = 0$ *and* $\Sigma^{**} = 0$.

(D) *When* $I =]\mu, \nu[$,

(i) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* \neq 0$ *and* $\Sigma^{**} \neq 0$,
(ii) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* = 0$ *and* $\Sigma^{**} \neq 0$,
(iii) $<_1$ *is* $\leqq$ *and* $<_2$ *is* $\leqq$ *if* $\Sigma^* \neq 0$ *and* $\Sigma^{**} = 0$
(iv) $<_1$ *is* $<$ *and* $<_2$ *is* $<$ *if* $\Sigma^* = 0$ *and* $\Sigma^{**} = 0$.

Proof. The specification (22) implies (2) as $J \neq \varnothing$ and so, by Theorem 1, (22) and (3) imply (23) and (24). What remains to be done is to examine what relations between the constants ω_i's and β_j's in (23) should correspond to the requirement $\Phi_j(I^n) \subseteq I$ in (22). We analyze the case (B) when $I = [\mu, \nu[$ in full and omit the details for the cases (A), (C) and (D).

With $I = [\mu, \nu[$, the range of Φ_j over I^n given by (23) is:

(i) $]\nu\Sigma^* + \mu\Sigma^{**}, \mu\Sigma^* + \nu\Sigma^{**}[+ \beta_j$ if $\Sigma^* \neq 0$ and $\Sigma^{**} \neq 0$,

(ii) $[\mu\Sigma^{**}, \nu\Sigma^{**}[+ \beta_j$ if $\Sigma^* = 0$ and $\Sigma^{**} \neq 0$,

(iii) $]\nu\Sigma^*, \mu\Sigma^*] + \beta_j$ if $\Sigma^* \neq 0$ and $\Sigma^{**} = 0$

and

(iv) $\{0\} + \beta_j$ if $\Sigma^* = 0$ and $\Sigma^{**} = 0$.

For each case, the inclusion of the range of Φ_j by $[\mu, \nu[$ corresponds to the inequalities (25) under (B) (i)–(iv). □

Remark. For each Φ_j given by (23) on I^n, the range is an interval of length $(\nu - \mu)(\Sigma^{**} - \Sigma^*) = (\nu - \mu)\sum_{i=1}^n |\omega_i|$. The inequalities (25) imply in particular that $\mu - \mu\Sigma^{**}\omega_i - \nu\Sigma^*\omega_i \leqq \nu - \nu\Sigma^{**}\omega_i - \mu\Sigma^*\omega_i$ which is equivalent to

$$(\nu - \mu) \sum_{i=1}^{n} |\omega_i| \leqq \nu - \mu$$

and reflects the fact that, if the range of Φ_j is to be in I, its length must not exceed that of I. Since $\nu - \mu > 0$, we get

$$\sum_{i=1}^{n} |\omega_i| \leqq 1, \tag{26}$$

and it implies in particular that $\sum_{i=1}^n \omega_i \leqq 1$.

If $\sum_{i=1}^{n} \omega_i < 1$, we may rewrite (23) as

$$\Phi_j(z) = \sum_{i=1}^{n} \omega_i(\zeta_i - \sigma_j) + \sigma_j, \qquad j = 1, \cdots, m, \tag{27}$$

where $\sigma_j = \beta_j/(1 - \sum_{i=1}^{n} \omega_i)$ and, by (24), $\sum_{j=1}^{m} \sigma_j = \sigma$. The aggregation functions Φ_j might arise in practice in the above form (27), if our group of n individuals are advisors to some external decisionmaker whose preferred allocations, prior to consulting with the group, are given by the numbers $\sigma_1, \cdots, \sigma_m$. (Note that (19) with (20) is the special case $\sigma_j = \sigma/m$ of (27).)

If $\sum_{i=1}^{n} \omega_i = 1$ in (23), we obtain from (26) that $\Sigma^* \omega_i = 0$ and so all weights are nonnegative. Furthermore (25) gives $\beta_j = 0$ for all $j \in J$. In conclusion, Theorem 2 and Theorem 3 can be combined to give the following characterization of aggregation by ordinary weighted arithmetic means:

THEOREM 4. *Let $m \geqq 3$ and $I = [\mu, \nu]$, $[\mu, \nu[$, $]\mu, \nu]$ or $]\mu, \nu[$ be an interval where $\mu < \nu$ are constants satisfying* (1). *Then a sequence of functions $\Phi_j : I^n \to I$ satisfies* (3) *and $\Phi_j(\alpha) = \alpha$ for some $\alpha \in I$ where $\alpha \neq \sigma/m$ if, and only if, there exists a sequence of weights $\omega_1, \cdots, \omega_n$, nonnegative with sum* 1, *such that, for all $z = (\zeta_1, \cdots, \zeta_n) \in I^n$,*

$$\Phi_j(z) = \sum_{i=1}^{n} \omega_i \zeta_i \qquad (j = 1, \cdots, m).$$

Several proofs of the above have previously appeared in the literature for the special case $I = [0, \sigma]$ and $\alpha = 0$. See [3], [4] and [7, Thm. 6.4].

Appendix.

PROPOSITION. *If, for $\psi : [-\varepsilon, \varepsilon]^n \to \mathbb{R}$*

$$\psi(x + y) = \psi(x) + \psi(y) \text{ whenever } x, y, x + y \in [-\varepsilon, \varepsilon]^n, \tag{28}$$

then there exists a $\bar{\psi} : \mathbb{R}^n \to \mathbb{R}$ satisfying

$$\bar{\psi}(x + y) = \bar{\psi}(x) + \bar{\psi}(y) \quad \text{for all } x, y \in \mathbb{R}^n, \tag{29}$$

and

$$\bar{\psi}(x) = \psi(x) \quad \text{for all } x \in [-\varepsilon, \varepsilon]^n. \tag{30}$$

Proof. From (28), $\psi(kx) = k\psi(x)$ for $kx \in [-\varepsilon, \varepsilon]^n$, that is,

$$\psi\left(\frac{1}{k} z\right) = \frac{1}{k} \psi(z) \quad \text{for all } z \in [-\varepsilon, \varepsilon]^n. \tag{31}$$

Let $x \in \mathbb{R}^n$ be arbitrary. There exists a positive integer k such that $u = x/k \in [-\varepsilon, \varepsilon]^n$. Define

$$\bar{\psi}(x) = k\psi(u) = k\psi\left(\frac{1}{k} x\right) \qquad (x \in \mathbb{R}^n, u \in [-\varepsilon, \varepsilon]^n). \tag{32}$$

This definition is unambiguous: If $x = ku = lv$ $(u, v \in [-\varepsilon, \varepsilon]^n)$ then, by (31),

$$l\psi(v) = kl\psi\left(\frac{v}{k}\right) = kl\psi\left(\frac{u}{l}\right) = k\psi(u),$$

as asserted. Note that (32) implies (30) for $k = 1$.

Finally, we prove (32) by choosing, for given $x, y \in \mathbb{R}^n$, an integer k so that x/k, y/k, $(x+y)/k$ are all in $[-\varepsilon, \varepsilon]^n$. By (32) and (28)

$$\bar{\psi}(x+y) = k\psi\left[\frac{1}{k}(x+y)\right] = k\psi\left(\frac{1}{k}x\right) + k\psi\left(\frac{1}{k}y\right) = \bar{\psi}(x) + \bar{\psi}(y)$$

for all $x, y \in \mathbb{R}^n$. □

REFERENCES

[1] J. ACZÉL, *Lectures on Functional Equations and their Applications*, Academic Press, New York–London, 1966.

[2] J. ACZÉL AND C. WAGNER, *A characterization of weighted arithmetic means*, SIAM J. Alg. Discrete Meth., 1 (1980), pp. 259–260.

[3] ———, *Rational group decisionmaking generalized: The case of several unknown functions*, C.R. Math. Rep. Acad. Sci. Canada, 3 (1981), pp. 138–142.

[4] J. ACZÉL, P. L. KANNAPPAN, C. T. NG AND C. WAGNER, *Functional equations and inequalities in rational group decisionmaking*, General Inequalities 3: Proc. Third International Conference on General Inequalities in Oberwolfach, 1981, to appear.

[5] Z. DARÓCZY AND L. LOSONCZI, *Über Erweiterungen der auf einer Punktmenge additiven Funktionen*, Publi. Math. Debrecen, 14 (1967), 239–245.

[6] J. G. DHOMBRES AND R. GER, *Conditional Cauchy Equations*, Glasnik Mat., 13 (33) (1978), pp. 39–62. (In Russian.)

[7] K. LEHRER AND C. WAGNER, *Rational Consensus in Science and Society*, Reidel Publishing, Boston–Dordrecht, 1981.

[8] C. T. NG, *Representation for measures of information with the branching property*, Inform. Control, 25 (1974), pp. 45–56.

[9] C. WAGNER, *Allocation, Lehrer models, and the consensus of probabilities*, Theory and Decision, 14 (1982), pp. 207–220.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0002 $01.25/0

ON THE DIMENSION OF BI-INFINITE SYSTEMS*

HARALD K. WIMMER†

Abstract. The kernel of banded Toeplitz matrices is studied.

The inversion of bi-infinite matrices is important for problems in interpolation theory (see e.g. [2]). In this note we describe the kernel of banded Toeplitz matrices and generalize a result in [3].

Let A_ν, $\nu = 0, 1, \cdots, p$ be complex $m \times n$ matrices. Put

$$A(z) := \sum_{\nu=0}^{p} A_{p-\nu}$$

and define

$$K_A := \left\{ (\cdots, x_{-1}, x_0, x_1, \cdots)^T,\ x_i \in \mathbb{C}^n, \text{ such that } \sum_{\nu=0}^{p} A_\nu x_{i+\nu} = 0 \text{ for all } i \in \mathbb{Z} \right\}.$$

Thus K_A is the right null-space of the bi-infinite block Toeplitz matrix

$$\begin{pmatrix} \cdots 0 & A_0 & A_1 & \cdots & A_p & 0 & \cdots \\ \cdots & 0 & A_0 & A_1 & \cdots & A_p & 0 \cdots \end{pmatrix}.$$

We shall relate the dimension of K_A to the rank and the characteristic polynomial of A.

It is convenient to work with bi-infinite formal power series instead of bi-infinite sequences. Put

$$\mathbb{C}^n\langle z\rangle := \left\{ \sum_{i=-\infty}^{\infty} x_i z^i,\ x_i \in \mathbb{C}^n \right\}.$$

To each $x = (\cdots, x_{-1}, x_0, x_1 \cdots)^T$ we associate an element $\bar{x} \in \mathbb{C}^n\langle z\rangle$ given by

$$\bar{x}(z) = \sum_{i=-\infty}^{\infty} x_i z^i.$$

We define

$$\bar{K}_A := \{\bar{x} \in \mathbb{C}^n\langle z\rangle,\ A\bar{x} = 0\}.$$

Then $x \in K_A$ if and only if $\bar{x} \in \bar{K}_A$ and the vector spaces K_A and $\bar{K}_A$ are isomorphic. In a natural way $\bar{K}_A$ is a $\mathbb{C}[z]$-module if we put

$$z\bar{x} = \sum_{i=-\infty}^{\infty} x_{i-1} z^i.$$

We first consider the case $A = (d)$, $d \in \mathbb{C}[z]$. Let d be factored into $d(z) = cz^k q(z)$ with $q(z) = \prod_{i=1}^{t} (z - a_i)^{s_i}$ where the roots a_i are nonzero and distinct. Then

$$\bar{K}_d = \bigoplus_{i=1}^{t} \bar{K}_{(z-a_i)s_i}.$$

* Received by the editors January 24, 1983.

† Mathematisches Institut, Universität Würzburg, D-8700 Würzburg, West Germany.

By adapting standard results on linear difference equations (see [1]) we establish a basis of the vector space $\bar{K}_{(z-a)s}$ of the form

$$\sum_{i=-\infty}^{\infty} i^{\sigma} a^{-i} z^{i}, \qquad \sigma = 0, 1, \cdots, s-1.$$

Hence

$$\dim \bar{K}_d = \deg q.$$

If $A \in \mathbb{C}^{m \times n}[z]$ has rank r (over $\mathbb{C}[z]$) then we denote the greatest common divisor of all $r \times r$ minors of A by $\chi(a)$ and call it (by abuse of language) the *characteristic polynomial* of A. Let F and G be two unimodular matrices which transform A into Smith form

$$FAG = S = \begin{pmatrix} D & O \\ O & O \end{pmatrix} \tag{1}$$

with $D = \operatorname{diag}(d_1, \cdots, d_r)$, $d_i | d_{i+1}$. Then

$$\chi(A) = \chi(D) = \prod_{\rho=1}^{r} d_\rho. \tag{2}$$

THEOREM. *Let the characteristic polynomial $\chi(A)$ of A be given by*

$$\chi(A) = cz^l h(z), \qquad h(0) \neq 0. \tag{3}$$

(a) *The $\mathbb{C}[z]$-module $\bar{K}_A$ is finitely generated. It is a torsion module if and only if A has full column rank.*

(b) *The vectorspace K_A has finite dimension if and only if A has full column rank. In this case* $\dim K_A = \deg h$ *where h is given by* (3).

Proof. Let the Smith form of A be given as in (1) and put $\tilde{S} := (D, O_{r,n-r})$. Then $\bar{x} = G\bar{y}$ is a $\mathbb{C}[z]$-module isomorphism between $\bar{K}_A$ and $\bar{K}_{\tilde{S}}$. Furthermore

$$\bar{K}_{\tilde{S}} = \bar{K}_D \oplus \underbrace{\mathbb{C}\langle z \rangle \oplus \cdots \oplus \mathbb{C}\langle z \rangle}_{(n-r)\text{-times}}. \tag{4}$$

If the column rank of A is not maximal i.e. $n > r$, then a block $O_{r,n-r}$ appears in the Smith form (1) and a summand $\mathbb{C}\langle z \rangle$ in (4). Obviously $\mathbb{C}\langle z \rangle$ is torsion free and its dimension as a $\mathbb{C}$-vectorspace is infinite. For $\bar{y} \in \bar{K}_D$ we have $d_r \bar{y} = 0$. Hence $\bar{K}_D$ is a torsion module over $\mathbb{C}[z]$. Because of $\bar{K}_D = \bigoplus_{\rho=1}^{r} \bar{K}_{d_\rho}$ we are back at the scalar case and (2) yields the dimension of $\bar{K}_D$. □

Remark. In the special case $A(z) = Bz + C$, part (b) of the theorem has been proved in [3]. The approach in [3] is based on the Kronecker normal form of the pencil $Bz + C$ and on deflating subspaces.

REFERENCES

[1] L. BRAND, *Differential and Difference Equations*, John Wiley, New York, 1966.

[2] A. S. CAVARETTA JR., W. A. DAHMEN, C. A. MICCHELLI AND P. W. SMITH, *On the solvability of certain systems of linear difference equations*, SIAM J. Math. Anal., 12 (1981), pp. 833–841.

[3] P. W. SMITH AND H. WOLKOWICZ, *Dimensionality of bi-infinite systems*, submitted to Lin. Alg. Appl.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0003 $01.25/0

PREEMPTIVE SCHEDULING, LINEAR PROGRAMMING AND NETWORK FLOWS*

D. DE WERRA†

Abstract. A refinement of the Birkhoff–von Neumann theorem on bistochastic matrices is derived by a simple argument based on network flow theory; this result is then used for solving a problem of preemptive scheduling on unrelated processors with constraints involving subsets of processors and subsets of jobs. A simple construction procedure is given; it generalizes a two-stage method (linear programming + network flow) developed by E. L. Lawler [J. Assoc. Comput. Mach., 25 (1978), pp. 612–619] and extended by R. Slowinski [Przeglad Statyst., 24 (1977), pp. 409–415, RAIRO Inform., 15 (1981), pp. 155–166].

Key words. preemptive scheduling, network flows, bistochastic matrices

1. Introduction.

The purpose of this paper is to give a general formulation of the preemptive scheduling problem for which one can use a method in 2 stages developed by E. L. Lawler et al. [3] and R. Slowinski et al. [5], and refined by R. Slowinski [4]. This procedure consists in solving first a linear programming problem for obtaining the partial processing times of the jobs on the various processors; the second phase is the construction of a schedule based on the previous processing times; it is performed by solving a finite sequence of compatible flow problems.

In the proposed formulation, more general constraints than in the previous models can be introduced; one may take into account the fact that some subsets of processors or some subsets of jobs use resources which are available in limited amounts.

In the next section, the problem of preemptive scheduling on unrelated processors will be discussed. The main result will be established in terms of subpermutation matrices in § 3, while in § 4 an example will be discussed.

For all terms related to flows, the reader is referred to Ford and Fulkerson [2].

2. Preemptive scheduling on unrelated processors.

We shall first describe the model which is an extension of the model used by Slowinski [4]. We are given a set $\tau = \{T_1, \cdots, T_n\}$ of jobs, a set $\mathcal{P} = \{P_1, \cdots, P_m\}$ of unrelated processors and a set $\mathcal{R} = R^P \cup R^T$ of renewable resources; each job T_q can be processed during some time on a processor, then one may interrupt its processing at any time and continue it later on some other processor (or on the same); π_q is the set of processors which can be used for job T_q while τ_p is the set of jobs which can use processor P_p.

For each job T_q we are given the times t_{pq} which would be needed to process T_q completely on processor P_p. We are also given a family $\mathcal{A} = (A_i | i \in I)$ of subsets of $\mathcal{P}$: A_i is the subset of processors which use the same resource $r_i^P \in R^P$; we are also given for each A_i a positive integer α_i: it is the number of units of r_i^P which are available at any time, so that no more than α_i processors in A_i (i.e. using r_i^P) can be working at the same time. We may assume that for each processor P_p there is a singleton $A_i = \{P_p\}$ in $\mathcal{A}$ with $\alpha_i = 1$. The reason will be made clear later.

Furthermore, we are given a family $\mathcal{B} = (B_j | j \in J)$ of subsets of τ: B_j is the subset of jobs which use the same resource $r_j^T \in R^T$; β_j units of this resource are available at any time, so that at most β_j jobs of B_j can be processed simultaneously. We similarly assume that for each job T_q there is a singleton $B_j = \{T_q\}$ in $\mathcal{B}$.

* Received by the editors January 4, 1982, and in revised form March 7, 1983. This paper was written while the author was visiting the Technical University of Poznan (Poland) in September 1982. This research was supported by the Institute of Control Engineering of that University.

† Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, 1007 Lausanne, Switzerland.

All renewable resources $r_i^P \in R^P$ and $r_j^T \in R^T$ are distinct, but a job (or a processor) can use several resources since it can be included in several subsets B_j (or A_i).

We are interested in finding the smallest total completion time T; in order to obtain it, we may first define z_{pq} as the time during which job T_q is assigned to processor P_p and solve the linear programming (LP) problem

Min T

$$\sum_{p:\, P_p \in A_i} \sum_{q:\, T_q \in \tau_q} z_{pq} \leqq \alpha_i T \qquad (i \in I), \tag{2.1}$$

$$\sum_{q:\, T_q \in B_j} \sum_{p:\, P_p \in \Pi_q} z_{pq} \leqq \beta_j T \qquad (j \in J), \tag{2.2}$$

$$\sum_{p:\, P_p \in \Pi_q} \frac{z_{pq}}{t_{pq}} = 1 \qquad (q = 1, \cdots, n), \tag{2.3}$$

$$z_{pq} \geqq 0 \qquad (p = 1, \cdots, m, \quad q = 1, \cdots, n). \tag{2.4}$$

The constraints (2.1) express the fact that the sum of the processing times of the jobs using processors in A_i must not exceed the total availability of resource r_i^P, namely $\alpha_i T$. Similarly (2.2) says that the sum of the processing times of the jobs in B_j (i.e. using resource r_j^T) must not exceed the total availability $\beta_j T$ of resource r_j^T. Conditions (2.3) ensure that all jobs are completely processed. Since in $\mathscr{A}$ there is for each processor P_p a subset $A_i = \{P_p\}$ with $\alpha_i = 1$, then (2.1) says also that the total working time of P_p must not exceed T. Similarly since there is in $\mathscr{B}$ for each job T_q a subset $B_j = \{T_q\}$ with $\beta_j = 1$, then (2.2) amounts to saying that the total processing time of T_q must not exceed T.

This LP model is thus an extension of the model considered by Slowinski [4]; the analysis of complexity is the same and will not be discussed here. Solving this LP problem is the first stage of the method; the second stage will consist in finding a schedule using no more than T time units and based on the values z_{pq} previously obtained; this second phase consists in constructing a sequence of compatible flows as described in the proof of Theorem 3.1.

As a consequence, we obtain the following result: If in this scheduling problem, $\mathscr{A}$ and $\mathscr{B}$ have a special structure (noncrossing families), then the two-stage method (LP + flow) will find a schedule using no more than T time units (where T is given by the LP solution).

3. The main result. Given a set E, a family $\mathscr{F} = (F_i \mid i \in I)$ of subsets of E is *noncrossing* if $F_i \cap F_j \neq \varnothing$ implies $F_i \subseteq F_j$ or $F_j \subseteq F_i$ $(i, j \in I)$. Let $M = \{1, \cdots, m\}$, $N = \{1, \cdots, n\}$ and $\mathscr{A} = (A_i \mid i \in I)$, $\mathscr{B} = (B_j \mid j \in J)$ be 2 noncrossing families of subsets of M and N respectively. Let $\alpha_i (i \in I)$ and $\beta_j (j \in J)$ be positive integers; we may assume that $\alpha_i \leqq |A_i|$ $(i \in I)$ and $\beta_j \leqq |B_j|$ $(j \in J)$. Let Z be a matrix with nonnegative entries z_{pq} $(p \in M, q \in N)$ and define

$$a(i) = \frac{1}{\alpha_i} \sum_{p \in A_i} \sum_{q \in N} z_{pq},$$

$$b(j) = \frac{1}{\beta_j} \sum_{q \in B_j} \sum_{p \in M} z_{pq},$$

$$T = \max\left(\max_{i \in I} a(i), \max_{j \in J} b(j)\right).$$

A *subpermutation matrix* is an $(m \times n)$ matrix S with $s_{ij} = 0$ or 1 such that

$$\sum_{p \in A_i} \sum_{q \in N} s_{pq} \leqq \alpha_i \qquad (i \in I),$$

$$\sum_{q \in B_j} \sum_{p \in M} s_{pq} \leqq \beta_j \qquad (j \in J).$$

THEOREM 3.1. *Let $M, N, \mathscr{A}, \mathscr{B}, \alpha_i (i \in I), \beta_j (j \in J), Z, T$ be defined as before. Then Z is a linear combination of subpermutation matrices $S_1, \cdots, S_r$*

$$Z = \lambda_1 S_1 + \cdots + \lambda_r S_r$$

with $\lambda_i \geqq 0$ $(i = 1, \cdots, r)$ and $\lambda_1 + \cdots + \lambda_r = T$.

Proof. We have to show that we can find a subpermutation matrix S_1 and a positive number λ_1 such that $Z' = Z - \lambda_1 S_1$ is a nonnegative matrix for which the following conditions hold:

(1) If

$$a'(i) \equiv \frac{1}{\alpha_i} \sum_{p \in A_i} \sum_{q \in N} z'_{pq} \quad (i \in I), \qquad b'(j) \equiv \frac{1}{\beta_j} \sum_{q \in B_j} \sum_{p \in M} z'_{pq} \quad (j \in J),$$

then

$$T' \equiv \max\left(\max_{i \in J} a'(i), \max_{j \in J} b'(j)\right) = T - \lambda_1.$$

(2) Either Z' has one more zero than Z or Z' has at least as many zeros as Z but there is at least one more subset A_i such that $a'(i) = T'$ or one more subset B_j such that $b'(j) = T'$.

Since each one of these cases can occur only a finite number of times, the process is finite; we will necessarily end up with a matrix $Z^{(r)} = Z - \lambda_1 S_1 - \cdots - \lambda_r S_r$ consisting of zeros.

(a) *Construction of* S_1. Without any loss of generality we may assume that in $\mathscr{A}$ all subsets A_i of M are different, that there is in $\mathscr{A}$ a subset $A_0 = M$ and that all singletons are in $\mathscr{A}$. Similar hypotheses are made for $\mathscr{B}$.

We represent $\mathscr{A}$ by an arborescence with root A_0 ((A_i, A_j) is an arc if and only if $A_i \supset A_j$ and there is no $k \neq i, j$ with $A_i \supset A_k \supset A_j$); so for each A_i $(i \neq 0)$ there is one arc (A_j, A_i) going into A_i.

For $\mathscr{B}$ we use a similar representation but we reverse the orientation of all arcs (so for each B_j $(j \neq 0)$, there is one arc (B_j, B_i) going out of B_j). Also we link $A_i = \{p\}$ and $B_j = \{q\}$ by an arc (A_i, B_j) if $z_{pq} > 0$.

In the network $\mathscr{N}$ thus obtained we assign capacities $c(x, y)$ and lower bounds $1(x, y)$ to each arc (x, y) as in Table 3.1.

One verifies immediately that the values $f(x, y)$, given in the last column of Table 3.1, define a feasible flow in $\mathscr{N}$ from A_0 to B_0 with value $(1/T) \sum_{p \in M} \sum_{q \in N} z_{pq}$; this flow is not necessarily integer, but by the integer value theorem there exists an integer flow f' in $\mathscr{N}$; its values on the arcs (p, q) are 0 or 1; so the values $f'(p, q)$ define a subpermutation matrix S_1.

(b) *Computation of* λ_1. Let $E = \{(p, q) | f'(p, q) = 1\}$ and $z_{\min} = \min_{(p,q) \in E} z_{pq} > 0$.

TABLE 3.1
Description of network: capacities, lower bounds and flow.

(x, y)	$c(x, y)$	$l(x, y)$	$f(x, y)$
(A_j, A_i)	α_i	$\begin{cases}\alpha_i & \text{if } a(i)=T\\ 0 & \text{if } a(i)<T\end{cases}$	$\frac{1}{T}\sum_{p\in A_i}\sum_{q\in N} z_{pq}$
(B_j, B_i)	β_j	$\begin{cases}\beta_j & \text{if } b(j)=T\\ 0 & \text{if } b(j)<T\end{cases}$	$\frac{1}{T}\sum_{q\in B_j}\sum_{p\in M} z_{pq}$
(p, q)	1	0	$\frac{1}{T} z_{pq}$

For each A_i $(i \in I)$ define $m_i = |E \cap \{(p, q): p \in A_i\}|$ and for each B_j $(j \in J)$ define $n_j = |E \cap \{(p, q): q \in B_j\}|$. Furthermore,

$$z_{\max} = \min\left\{ \min_{\substack{i\in I\\ a(i)<T}} \frac{\alpha_i(T-a(i))}{\alpha_i - m_i},\ \min_{\substack{j\in J\\ b(j)<T}} \frac{\beta_j(T-b(j))}{\beta_j - n_j} \right\},$$

and if $a(i) = T$ for each $i \in J$ and $b(j) = T$ for each $j \in J$, then $z_{\max} = +\infty$.

We set $\lambda_1 = \min(z_{\min}, z_{\max})$. Clearly $\lambda_1 > 0$; it follows also from the choice of λ_1 that $Z' = Z - \lambda_1 S_1$ has nonnegative entries. We now have to show that in Z' the following holds:

$$T' \equiv \max\left(\max_{i\in I} a'(i), \max_{j\in J} b'(j) \right) = T - \lambda_1.$$

Observe that in each A_i with $a(i) = T$ (respectively in each B_j with $b(j) = T$) we have $m_i = \alpha_i$ (resp. $n_j = \beta_j$) so $a'(i) = a(i) - \alpha_i\lambda_1/\alpha_i = a(i) - \lambda_1 = T - \lambda_1$ (resp. $b'(j) = T - \lambda_1$). Furthermore if $a(i) < T$, then

$$a'(i) = a(i) - \frac{m_i\lambda_1}{\alpha_i} = a(i) - \lambda_1 + \frac{(\alpha_i - m_i)}{\alpha_i}\lambda_1 \leqq T - \lambda_1$$

(since by choice of λ_1 we have $(\alpha_i - m_i)\lambda_1/\alpha_i \leqq T - a(i)$). Similarly if $b(j) < T$ then $b'(j) \leqq T - \lambda_1$.

Finally if $\lambda_1 = z_{\min}$, there is one more zero entry in Z than in Z', and if $\lambda_1 = z_{\max}$ one more set A_i (or B_j) satisfies $a'(i) = T'$ (or $b'(j) = T'$). This ends the proof.

An example of a matrix Z with the families $\mathscr{A}$ and $\mathscr{B}$ and the integers α_i $(i \in I)$, β_j $(j \in J)$ is given in Fig. 3.1; the value obtained for T is 15.

The associated network $\mathcal{N}$ is represented on Fig. 3.2 where all flow values indicated should be divided by $T = 15$. Figure 3.3 shows the network with the lower bounds $l(x, y)$ and the capacities $c(x, y)$ indicated by $[l(x, y), c(x, y)]$ for all arcs (x, y) except those for which $l(x, y) = 0$ and $c(x, y) = 1$. An integer compatible flow from A_0 to B_0 with value 2 is represented and the corresponding subpermutation matrix S_1 is shown in Fig. 3.4. The computation of $z_{\max}$ is given in Fig. 3.5; one gets $z_{\max} = 5$; hence $\lambda_1 = \min(z_{\min}, z_{\max}) = 5$. The resulting matrix Z' is given in Fig. 3.6; one verifies that with $T' = T - \lambda_1 = 10$, one has now $a'(2) = T' = 10$ (initially we had $a(2) = 10 < T = 15$).

COROLLARY 3.1 (Birkhoff and von Neumann). *A bistochastic matrix is a convex combination of permutation matrices.*

Proof. Take $M = N$ and $A_i = \{i\}$, $\alpha_i = 1$ $(i \in M)$, $B_j = \{j\}$, $\beta_j = 1$ $(j \in N)$.

$$\begin{array}{ccc|c|c|c|c|c|}
 & & & \multicolumn{5}{c}{\overbrace{\qquad\qquad\qquad\qquad\qquad\qquad}^{B_0}} \\
 & & & & \multicolumn{4}{c}{\overbrace{\qquad\qquad\qquad\qquad}^{B_6}} \\
 & & & B_1 & B_2 & B_3 & B_4 & B_5 \\ \cline{4-8}
 & & A_1 & 4 & 0 & 0 & 8 & 0 \\ \cline{4-8}
 & A_5 & A_2 & 0 & 0 & 2 & 0 & 8 \\ \cline{4-8}
A_0 & & A_3 & 0 & 6 & 2 & 0 & 0 \\ \cline{4-8}
 & & A_4 & 0 & 0 & 4 & 0 & 0 \\ \cline{4-8}
\end{array}$$

$$Z$$

$M = A_0 = \{1, 2, 3, 4\} \alpha_0 = 4$

$A_i = \{i\}$ $(1 \leqq i \leqq 4)$

$\alpha_i = 1$

$A_5 = \{1, 2, 3\}$

$\alpha_5 = 2$

$N = B_0 = \{1, 2, 3, 4, 5\} \beta_0 = 4$

$B_j = \{j\}$ $(1 \leqq j \leqq 5)$

$\beta_j = 1$

$B_6 = \{2, 3, 4, 5\}$

$\beta_6 = 3$

i	$a(i)$
0	34/4
1	12
2	10
3	8
4	4
5	15

j	$b(j)$
0	34/4
1	4
2	6
3	8
4	8
5	8
6	10

$$T = \max\left(\max_{i \in I} a(i), \max_{j \in J} b(j)\right) = 15$$

FIG. 3.1

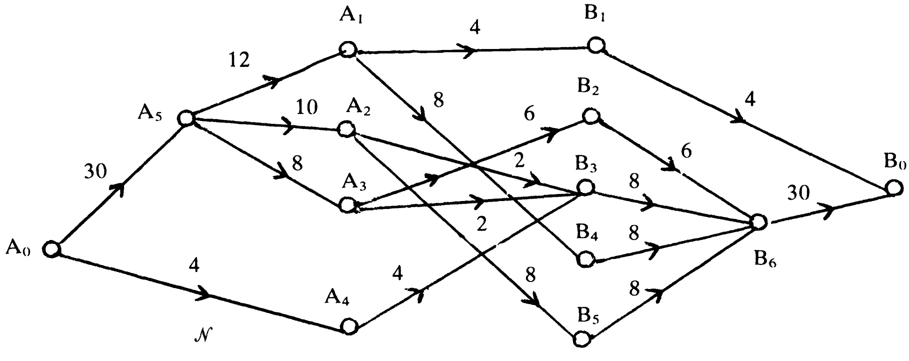


FIG. 3.2. *Network with flows in the arcs (all values are to be divided by $T = 15$).*

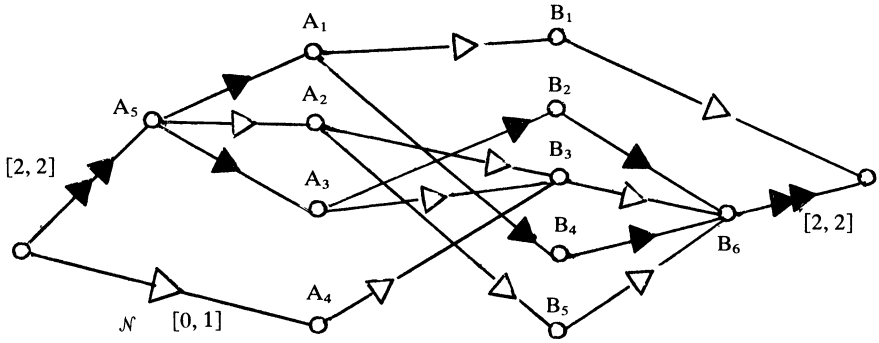


FIG. 3.3. *Network 𝒩 with compatible integer flow f' defining S_1 (the brackets indicate* $1(x, y)$, $c(x, y)$ *for arcs (x, y);* $1(x, y) = 0$ *and* $c(x, y) = 1$ *if there is no bracket).*

	B_1	B_2	B_3	B_4	B_5
A_1				1	
A_2					
A_3		1			
A_4					

$Z_{\min} = \min(Z_{14}, Z_{32}) = \min(8, 6) = 6$

FIG. 3.4. *Subpermutation matrix S_1 defined by the flow f' of Fig.* 2.3.

i	m_i	α_i	$a(i)$	$\alpha_i(T-a(i))/(\alpha_i-m_i)$
0	2	4	34/4	13
1	1	1	12	∞
2	0	1	10	$5 \leftarrow z_{\max} = 5$
3	1	1	8	∞
4	0	1	4	11
5	2	2	15	—

j	n_j	β_j	$b(j)$	$\beta_j(T-b(j))/(\beta_j-n_j)$
0	2	4	34/4	13
1	0	1	4	11
2	1	1	6	∞
3	0	1	8	7
4	1	1	8	∞
5	0	1	8	7
6	2	3	10	15

FIG. 3.5. *Computation of $Z_{\max}$.*

	B_1	B_2	B_3	B_4	B_5
A_1	4	0	0	3	0
A_2	0	0	2	0	8
A_3	0	1	2	0	0
A_4	0	0	4	0	0

Z'

FIG. 3.6

4. A simple example. We will describe the construction of the schedule in the special case considered by Slowinski [4]; the family $\mathscr{A}$ is $\mathscr{A}=(A_i|i\in M)$ with $A_i=P_p$ for $i=1,\cdots,m$; so (2.1) becomes

$$\sum_{q:\,T_q\in\tau_p} z_{pq}\leqq T \qquad (p=1,\cdots,m). \tag{4.1}$$

This means that no processor can work more than T time units and there are no renewable resources used by processors.

$\mathscr{B}$ contains a) disjoint subsets $B_1,\cdots,B_r$ of $\tau=\{T_1,\cdots,T_n\}$; B_j is the subset of jobs using resource r_j^T which is available in quantity β_j at any time; b) subsets $B_{r+j}=T_j$ for $j=1,\cdots,n$.

The constraints (2.2) now become

$$\sum_{q:\,T_q\in B_j}\ \sum_{p:\,P_p\in\Pi_q} z_{pq}\leqq\beta_jT \qquad (j=1,\cdots,r) \tag{4.2}$$

(the sum of the processing times of all jobs in B_j must not exceed the total availability β_jT of resource r_j^T),

$$\sum_{p:\,P_p\in\Pi_q} z_{pq}\leqq T \qquad (q=1,\cdots,n) \tag{4.3}$$

(the total processing time of job T_q must not exceed T). We may assume that dummy jobs have been introduced so that (4.1) consists in fact of equalities. One can obtain a subpermutation matrix S_1 by applying the flow method described above; this gives the first assignment of jobs to processors; in order to determine the time λ_1 during which this assignment will be used, one computes:

$$z_{\min}=\min_{p,q:\,s^1_{pq}=1} z_{pq}>0.$$

As before, once we have S_1, we denote by n_j the number of $s_{pq}=1$ of S_1 such that $T_q\in B_j$ and we compute for each j

$$b(j)=\frac{1}{\beta_j}\sum_{q:\,T_q\in B_j}\ \sum_{p:\,P_p\in\Pi_q} z_{pq},$$

$$z^1_{\max}=\min_{\substack{q\\ \sum_p z_{pq}<T}}\Big(T-\sum_{\substack{p\\ P_p\in\Pi_q}} z_{pq}\Big)\quad\text{with } z^1_{\max}=\infty\quad\text{if }\sum_p z'_{pq}=T\quad\text{for each } q,$$

$$z^2_{\max}=\min_{\substack{j\\ b(j)<T}}\frac{\beta_j(T-b(j))}{\beta_j-n_j}\quad\text{with } z^2_{\max}=\infty\quad\text{if } b(j)=T\quad\text{for each } j.$$

Finally $\lambda_1=\min(z_{\min}, z^1_{\max}, z^2_{\max})$.

As an example, consider the matrix Z in Fig. 4.1; there is only one renewable resource which is available in quantity $\beta_1 = 2$ and which is used by the jobs T_1, T_2, T_3. We have $T = 5$; if we have obtained by the flow technique the subpermutation matrix S_1 given in Fig. 4.1, then $n_1 = 1$ and $z_{\min} = 2$; we compute $z^1_{\max} = 2$ and $z^2_{\max} = 2(5-4.5)/(2-1) = 1$, so $\lambda_1 = 1$.

We now have one more constraint which is an equality, namely $b'(1) = 4 = T'$. The resulting matrix Z' is given in Fig. 4.2. We can continue the decomposition and we may finally get the schedule given in Fig. 4.3.

Remark 4.1. The computation of λ_1 given by Slowinski [4] is different from the one given here; for the example of Fig. 4.1, the value of λ_1 obtained by Slowinski is $\lambda_1 = \min(z_{\min}, T - z_{\max})$ where

$$z_{\max} = \max \max_{q:\, s^1_{pq}=0} \sum_p z_{pq}, \max_j \sum_{\substack{q \in B \\ s^1_{pq}=0}} \sum_p \frac{z_{pq}}{\beta_j};$$

this would have given $\lambda_1 = 2$, and for $Z' = Z - \lambda_1 S_1$ we would have $b'(1) = 7/2 > 3 = T' = T - \lambda_1$, and the construction could not be continued.

	T_1	T_2	T_3	T_4	T_5
P_1	4			1	
P_2		1	2	2	
P_3	1	1			3

Z
$T = 5$

$B_1 = \{T_1, T_2, T_3\}$ $\beta_1 = 2$

1					
			1		$m_1 = 1$
				1	

S_1

FIG. 4.1. *An example with one additional resource.*

P_1	3			1	
P_2		1	2	1	
P_3	1	1			2

B_1

$T' = T - \lambda_1 = 4$

$Z' = Z - \lambda_1 S_1$

FIG. 4.2

	1	3	4	5
P_1	T_1	T_1	T_1	T_4
P_2	T_4	T_3	T_4	T_2
P_3	T_5	T_5	T_2	T_1

FIG. 4.3. *Final schedule.*

5. General setting of the problem. More generally, we could formulate the problem of stage 2 as follows in terms of hypergraphs (see Berge [1] for definition of hypergraphs). Let Z be an $(m\times n)$ matrix with real nonnegative entries z_{pq}; let $X=\{(p,q)|z_{pq}>0\}$ and $\mathscr{E}=(E_i|i\in I)$ be a family of subsets of X; we are also given for each E_i a positive integer α_i (with $\alpha_i\leqq|E_i|$). Consider the hypergraph $H=(X,\mathscr{E})$ constructed on the node set X and the family $\mathscr{E}$ of edges. A subpermutation matrix S was defined as an $(m\times n)$ matrix with entries $s_{pq}=0$ or 1 such that for each E_i we have $|E_i\cap\{(p,q)|s_{pq}=1\}|\leqq\alpha_i$. S can be considered as the characteristic vector of a subset S^* of nodes of H such that $|E_i\cap S^*|\leqq\alpha_i$ for $i\in I$.

We have then the general result:

THEOREM 5.1. *Let Z, $\mathscr{E}$ and $\alpha_i (i\in I)$ be given as above. If the associated hypergraph $H=(X,\mathscr{E})$ is unimodular and if $T=\max_{i\in I}(1/\alpha_i)\sum_{(p,q)\in E_i}z_{pq}$, then there is a finite r such that $Z=\lambda_1S_1+\cdots+\lambda_rS_r$ where for each k $\lambda_k\geqq0$, S_k is a subpermutation matrix and $\lambda_1+\cdots+\lambda_r=T$.*

Sketch of proof. If A is the node-edge incidence matrix of H ($a_{ik}=1$ if edge E_i contains node k), one has to find a subset S^* of nodes (corresponding to a subpermutation matrix); such a set can be obtained by finding an integer solution of

$$\mathbf{l}\leqq A\mathbf{x}\leqq\mathbf{c},\qquad \mathbf{0}\leqq\mathbf{x}\leqq\mathbf{1} \tag{5.1}$$

where

$$l_i=\begin{cases}0 & \text{if } \dfrac{1}{T}\sum\limits_{(p,q)\in E_i}z_{pq}<\alpha_i,\\ \alpha_i & \text{if } \dfrac{1}{T}\sum\limits_{(p,q)\in E_i}z_{pq}=\alpha_i,\end{cases}\qquad (i\in I)$$

and $c_i=\alpha_i$. Such an integer solution can be found since A is totally unimodular, $\mathbf{l}$, $\mathbf{c}$ are integers and $x_j=(1/T)z_{pq}$ (if node j of H corresponds to entry (p,q) of Z) gives a feasible solution of (5.1). Notice that for ensuring the existence of a solution $\mathbf{x}\leqq\mathbf{1}$, it is sufficient (but not necessary) to assume that each pair (p,q) occurs in some subset E_r with $\varepsilon_r=1$. Then λ_1 is determined as in § 3 and the process is repeated with $Z'=Z-\lambda_1S_1$. This ends the proof.

For a general unimodular H, the second stage of the method would consist of solving a sequence of LP problems (i.e. finding a feasible integer solution to a system of inequalities with a totally unimodular matrix). If $\mathscr{E}$ is the union of 2 noncrossing families $\mathscr{A}$, $\mathscr{B}$, then stage 2 consists of solving a sequence of compatible flow problems as was shown here (see also de Werra [6]).

It is worth observing that in the first stage of the method one could add many other constraints (not involving T) in the LP problem; the second stage would not be changed since it depends only on the constraints involving the family $\mathscr{E}$ (or $\mathscr{A}$ and $\mathscr{B}$).

Acknowledgments. The author would like to express his gratitude to Professors J. Weglarz and R. Slowinski who introduced him to this problem and who made valuable comments on this paper.

REFERENCES

[1] C. BERGE, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1973.

[2] L. R. FORD AND J. FULKERSON, *Flows in Networks*, Princeton Univ. Press, Princeton, NJ, 1962.

[3] E. L. LAWLER AND J. LABETOULLE, *On preemptive scheduling of unrelated processors by linear programming*, J. Assoc. Comput. Mach., 25 (1978), pp. 612–619.

[4] R. SLOWINSKI, *L'ordonnancement des tâches préemptives sur les processeurs indépendants en présence de ressources supplémentaires*, RAIRO Inform., 15 (1981), pp. 155–166.

[5] R. SLOWINSKI AND J. WEGLARZ, *Minimalno-Czasowy Model Sieciowy z Roznymi Sposobami Wykonywania Czynnosci*, Przeglad Statyst, 24 (1977), pp. 409–415. (In Polish.)

[6] D. DE WERRA, *Some classes of hypergraphs occurring in chromatic scheduling*, Cahiers Centre Études Rech. Opér., 21 (1979), pp. 239–245.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0004 $01.25/0

A NEW FORMULATION FOR THE TRAVELLING SALESMAN PROBLEM*

A. CLAUS†

Abstract. The standard formulation of the travelling salesman problem on n nodes as an integer program involves use of 2^n subtour elimination constraints. In this paper we provide a set of on the order of n^3 constraints that define the same polytope. This is accomplished through introduction of additional variables. An additional set of $n(n-1)/2$ constraints is introduced and results in a polytope that is smaller than the subtour elimination polytope. The introduction of further variables as well as constraints results in an even smaller polytope.

1. Introduction. Many algorithms have been developed for the travelling salesman problem. These include edge switching heuristics, branch and bound procedures (Held and Karp [3]) integer programming approaches (Dantzig [1]) and other ingenious techniques (e.g. Christofides [2]).

The general problem is known to be NP complete (see Lawler [7]). A new approach to it is therefore of interest primarily for its value as a heuristic.

The approach described in this paper is an integer programming formulation that defines the same LP polytope as does the early subtour elimination approach of Dantzig, Fulkerson and Johnson or perhaps a smaller one. It however involves only a number of constraints proportional to the number of nodes times the number of edges of the graph, rather than an exponential number of constraints.

The method involves too large a number of variables and constraints to be practical as a heuristic for large problems, even if the LP formulation always gives rise to integer solutions. It may, however, have value in shedding light on the differences between the polytopes defined by travelling salesman solutions and by the subtour elimination constraints. When these constraints yield integer solutions this method yields exact solutions.

In any case this approach shows that one can find the minimum cost solution in the polytope defined by the subtour elimination (and continuity) constraints by a polynomial algorithm.

It also suggests an interesting problem discussed below.

2. The problem and formulation. The problem we consider is one of finding a minimum cost Hamiltonian path from one vertex V_S to another V_T in a given weighted directed graph.

The programming approach involves introducing a variable X_{ij} for each pair of nodes, that takes on value if the edge i to j is in the tour corresponding to the given choice of variables and 0 if it is not. (In the corresponding polytope these variables are constrained to be between zero and one.)

The continuity constraints, for all i,

$$\sum_j X_{ij} = \sum_j X_{ji} = 1,$$

$$\sum_j X_{sj} = 1, \qquad \sum X_{jT} = 1,$$

$$\sum_i X_{is} = 0, \qquad X_{Ti} = 0,$$

* Received by the editors October 26, 1982, and in revised form December 7, 1982. This research was supported in part by the National Science Foundation under contract 8202555.

† Ecole Supérieure de Commerce (ESCAE), Montpellier, France.

insure that, for an integral solution, exactly one arc enters and leaves each vertex. (The objective function, to be minimized, for arc cost C_{ij} is $\sum X_{ij}C_{ij}$.)

There are several ways to add additional constraints in order to guarantee that an integral solution corresponds to a Hamiltonian tour.

The Dantzig, Fulkerson and Johnson approach was to require that for every nontrivial subset X of the nodes (other than s), we have

$$\sum X_{ij} \geqq 1$$

for all i not in X and j in X. This insures that one can get from the source to any set of nodes in a tour corresponding to an integral solution.

These "subtour elimination constraints" have the difficulty that there are an exponential number of them.

Other sets of constraints having the same effect have been described (see for example Wagner [8] where constraints of the form $u_i - u_j + nX_{ij} \leqq n - 1$ are mentioned). However the sets we have seen that involve a polynomial number of constraints lead to a different polytope than the subtour elimination constraints, and tend to yield noninteger solutions peculiar to their own polytope.

We now introduce new variables and constraints that replace the subtour elimination constraints by a number of them that is proportional to the number of nodes times the number of finite cost arcs in the given weighted digraph.

The idea is to interpret the variables X_{ij} as defining capacity for the arc $\{i, j\}$. We then insist that the capacities be sufficient to support flow of one unit from the source to each vertex (and from thence to the sink X if we choose).

Thus for each vertex k other than s we define variables Y_{ijk} with the following properties:

Continuity at i: $\sum_j Y_{ijk} = \sum_j Y_{jik}$ for i other than s and k;

Capacity: $0 \leqq Y_{ijk} \leqq X_{ij}$;

Flow from s to k: $\sum_i Y_{ikk} = 1$, $\sum_i Y_{sik} = 1$, $\sum Y_{kik} = 0$.

The addition of these variables and constraints provide the new formulation under discussion here.

These additional constraints as noted contain the information that the capacities X_{ij} are sufficient to allow unit flow from the source to each node.

The following remarks immediately follow

1. Any travelling salesman tour corresponds to a set of X_{ij} admitting such flow.

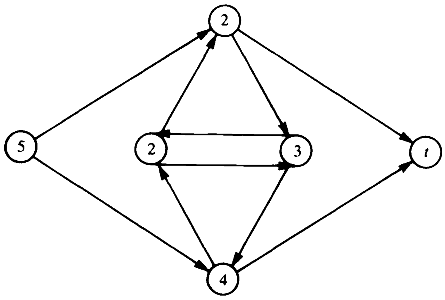
2. No other integral variable solution of the continuity constraints admits such flow.

3. Every solution of the linear program defined by these constraints obeys every one of the subtour elimination constraints. The polytope in the X_{ij} variables defined by these constraints is contained in that defined by the subtour elimination constraints.

3. Further constraints. Solutions $\{X_{ij}\}$ that do not have a Hamiltonian tour in their support can be constructed.

The smallest example shown below[1]

[1] Due to Peter Shor and Mark Haiman, Department of Mathematics, MIT.



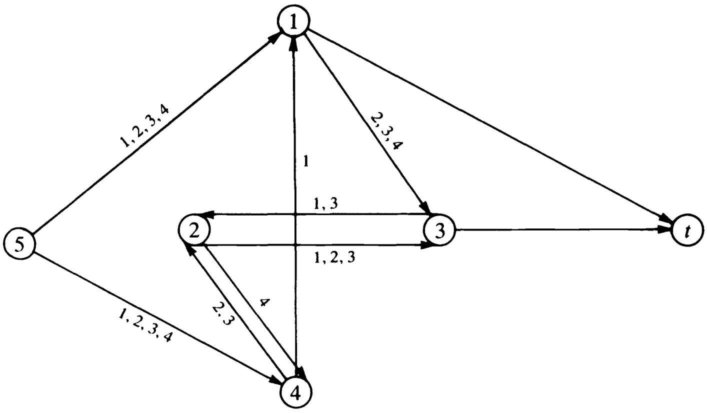
has arc capacities of $\frac{1}{2}$ units each and is a basic feasible solution to the subtour elimination polytope. However, the introduction of $n(n-1)/2$ additional constraints

$$\sum_j Y_{ikj} + \sum_f Y_{ijk} = 1, \qquad k, j \in N, \quad k \pm j,$$

would make the above solution infeasible.

These mutual flow constraints will always be satisfied by a Hamiltonian path, since along a Hamiltonian path only one node gets flow from the other.

A solution $\{Y_{ijk}\}$ with flows of $\frac{1}{2}$ units that satisfies the subtour elimination as well as the mutual flow constraints is shown below:



The mutual flow constraints for commodity 1 are partially satisfied by the circulation of commodity 1 between nodes 2 and 3.

To exclude these types of solutions we can introduce constraints that require the flow of any commodity into a node to originate from the source, thus avoiding the above types of circulations.

Introducing the quantities

$$Q^{l,k} = \sum_i Y_{ikl} \quad \text{(flow of commodity } l \text{ into node } k\text{)}$$

and

$$Z_{ij}^{kl} = \text{flow of commodity } l, k \text{ on the arc } i, j$$

permits us to express the above conditions with the following set of linear constraints (circulation elimination constraints):

$$\sum_i Z^{lk}_{si} = Q^{lk},$$

$$\sum_i Z^{lk}_{ik} = Q^{lk},$$

$$\sum_i Z^{lk}_{ij} - \sum_j Z^{kl}_{ij} = 0, \qquad i, j \neq s, k.$$

$$Z^{lk}_{ij} \leqq Y_{ijl},$$

$$Z^{lk}_{ij} \leqq Y_{ijk},$$

Since along a Hamiltonian path either $Q^{lk} = 0$ or $Q^{kl} = 0$ we can impose the stronger constraints

$$Z^{kl}_{ij} + Z^{lk}_{ij} \leqq Y_{ijk}, \qquad Z^{kl}_{ij} + Z^{lk}_{ij} \leqq Y_{ijl}.$$

If commodity l enters node k ($Q^{lk} > 0$) then we need a flow of commodity l from node k to node l. This may be expressed as follows:

$$\sum_i U^{lk}_{ki} = Q^{lk},$$

$$\sum_i U^{lk}_{il} = Q^{lk},$$

$$\sum_i U^{lk}_{ij} - \sum_j U^{lk}_{ij} = 0, \qquad i, j \neq k, l,$$

$$U^{lk}_{ij} \leqq Y_{ijl},$$

$$U^{lk}_{ij} \leqq X_{ij} - Y_{ijk},$$

In addition we have mutual flow constraints that are valid along a Hamiltonian path,

$$\left.\begin{array}{l} \sum_i (Z^{lk}_{ip} + Z^{kl}_{ip}) + \sum_i (Z^{pl}_{ik} + Z^{lp}_{ik}) + \sum_i (Z^{pk}_{il} + Z^{kp}_{il}) = 1, \\ \sum_i (U^{lk}_{ip} + U^{kl}_{ip}) + \sum_i (U^{pl}_{ik} + U^{lp}_{ik}) + \sum_i (U^{pk}_{il} + U^{kp}_{il}) = 1, \end{array}\right. \qquad p, l, k \in N,$$

as well as arc constraints,

$$Z^{lk}_{ij} + Z^{kl}_{ij} + U^{lk}_{ij} = Y_{ijl}, \qquad i, j, l, k \in N,$$

that are also valid along a Hamiltonian path.

Circulation elimination constraints for the Z and U commodities may also be added to avoid the mutual flow constraints for these commodities being satisfied by such circulations.

4. Conclusions. Introduction to these mutual flow and the circulation elimination constraints yields a polytope $\{P\}$ that is smaller than the subtour elimination polytope.

If the vertices of P were all travelling salesman tours then this formulation would provide a polynomial algorithm for the travelling salesman problem and by inference to all NP problems. Since this seems highly improbable there may be other constraints.

In particular if $P \neq NP$ it must be that there exist solutions $\{X_{ij}\}$ with $\{Y_{ijk}\}$ obeying all these constraints such that the arcs $\{i, j\}$ for which $X_{ij} \neq 0$ do not contain a Hamiltonian path from S to T. (If none such exists we have a polynomial algorithm for finding the existence of such a tour.) An interesting problem suggested by this work is to find such a solution $\{X_{ij}\}$ that, as noted here, does not have a Hamiltonian tour in this "support". Consideration of such configurations could conceivably shed light on the differences between the Travelling Salesman and the polytope P.

We are engaged in programming the algorithm suggested by this formulation. Application of the ellipsoid method [6] will either solve the TSP polynomially, or lead to examples of configurations as noted in § 3. We hope to be able to report on the rate of succcess of this approach at a later date.

Acknowledgments. The author would like to thank D. J. Kleitman and Peter Shor for editorial assistance and advice.

REFERENCES

[1] G. B. DANTZIG, D. R. FULKERSON AND S M. JOHNSON, *On a linear programming, combinatorial approach to the travelling salesman problem*, Oper. Res., 7 (1959), pp. 58–66.

[2] N. CHRISTOFIDES, *The shortest Hamiltonian chain of a graph*, SIAM J. Appl. Math., 19 (1970), pp. 689–696.

[3] M. HELD AND R. M. KARP, *The travelling salesman problem and minimum spanning trees*, Oper. Res., 18 (1970), pp. 1138–1162.

[4] M. W. PADBERG AND M. G. GROTSCHEL, *On the symmetric travelling salesman problem* I, II, Math. Programming, 16 (1979), pp. 265–302.

[5] A. CLAUS, *A simultaneous enumeration approach to the travelling salesman problem*, Stud. Appl. Math., 58 (1978), pp. 159–163.

[6] L. G. KHACHIYAN, *A polynomial algorithm in linear programming*, Soviet Math. Dokl, 20 (1979), pp. 191–194.

[7] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.

[8] H. WAGNER, *Principles of Operations Research*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[9] G. B. DANTZIG, D. R. FULKERSON AND S. M. JOHNSON, *Solution of a large scale travelling salesman problem*, Oper. Res., 2 (1954), pp. 393–410.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

CHARACTERIZATION OF POSITIVE DEFINITE AND SEMIDEFINITE MATRICES VIA QUADRATIC PROGRAMMING DUALITY*

S.-P. HAN†‡ AND O. L. MANGASARIAN†

Abstract. Positive definite and semidefinite matrices induce well-known duality results in quadratic programming. The converse is established here. Thus if certain duality results hold for a pair of dual quadratic programs, then the underlying matrix must be positive definite or semidefinite. For example, if a strict local minimum of a quadratic program exceeds or equals a strict global maximum of the dual, then the underlying symmetric matrix Q is positive definite. If a quadratic program has a local minimum, then the underlying matrix Q is positive semidefinite if and only if the primal minimum exceeds or equals the dual global maximum and $x^TQx=0$ implies $Qx=0$. A significant implication of these results is that the Wolfe dual may not be meaningful for nonconvex quadratic programs and for nonlinear programs without locally positive definite or semidefinite Hessians, even if the primal second order sufficient optimality conditions are satisfied.

Key words. positive definite matrices, quadratic programming, duality

AMS (MOS) subject classifications. 90C20, 15A63

1. Introduction. It is well known [3], [4], [11], [10] that the dual quadratic programs

$$\text{(1a)}\qquad \begin{aligned}&\underset{x}{\text{Minimize}}\quad \tfrac{1}{2}x^TQx+p^Tx\\ &\text{subject to}\quad Ax\leqq b,\qquad Cx=d,\end{aligned}$$

$$\text{(1b)}\qquad \begin{aligned}&\underset{x,u,v}{\text{Maximize}}\quad -\tfrac{1}{2}x^TQx-b^Tu-d^Tv\\ &\text{subject to}\quad Qx+A^Tu+C^Tv+p=0,\qquad u\geqq 0,\end{aligned}$$

where Q, A and C are given real matrices of order $n\times n$, $m\times n$ and $k\times n$ respectively, with $Q=Q^T$, and where p, b and d are given vectors in the real finite dimensional Euclidean spaces R^n, R^m and R^k respectively, possess many important relations when Q is positive semidefinite or positive definite. In this paper we are interested in the converse: What sort of duality relations between (1a) and (1b) induce positive definiteness or semidefiniteness in Q? A key role in deriving these converse relations is played by the following conjugate cone characterization of positive definite and semidefinite matrices [8].

THEOREM 1.1 [8]. *Let K be a nonempty convex polyhedral cone in R^n. The $n\times n$ real matrix P is positive semidefinite if and only if P is positive semidefinite plus on the cone K and positive semidefinite on the conjugate cone K^P; that is*:

$$\text{(2)}\qquad \begin{aligned}&x\in K\Rightarrow x^TPx\geqq 0,\\ &x^TPx=0,\ x\in K\Rightarrow (P+P^T)x=0,\end{aligned}$$

$$\text{(3)}\qquad y\in K^P:=\{y\,|\,y(P+P^T)x\leqq 0, \forall x\in K\}\Rightarrow y^TPy\geqq 0.$$

* Received by the editors July 6, 1982, and in revised form January 27, 1983. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041. This material is based on work supported by the National Science Foundation under grants ENG-7903881, MCS-8203603, MCS-7901066 and MCS-8200632.

† University of Wisconsin-Madison, Computer Sciences Department, Madison, Wisconsin 53706.

‡ University of Illinois, Department of Mathematics, Urbana, Illinois 61801.

THEOREM 1.2 [8]. *Let K be a nonempty convex polyhedral cone in R^n. The $n \times n$ real matrix P is positive definite if and only if P is positive definite on K and K^A, that is,*

$$0 \neq x \in K \Rightarrow x^TPx > 0,$$

$$0 \neq y \in K^P \Rightarrow y^TPy > 0.$$

With the help of these characterization theorems and the second order optimality conditions of quadratic programming [6], [9], [2], [1], we show for example in Theorem 3.5 that if a strict local minimum of a quadratic program exceeds or equals a strict global maximum of the dual, then the matrix Q must be positive definite. In Theorem 3.6 we show that if a quadratic program has a local minimum, then Q is positive semidefinite if and only if the primal minimum exceeds or equals the dual global maximum, and $Qx = 0$ whenever $x^TQx = 0$. In Corollary 3.7 we show that if the primal feasible and dual feasible sets are nonempty, and if the weak duality relation holds, that is, the primal objective exceeds or equals the dual objective over their respective feasible regions, and if $Qx = 0$ whenever $x^TQx = 0$, then Q is positive semidefinite. In [7] positive-definiteness of the Hessian of the Lagrangian of nonlinear programs was established under more restrictive assumptions.

The import of these and our other results is that when certain simple and desirable duality results are satisfied by a pair of dual quadratic programs, then the underlying matrix must be positive definite or semidefinite. This leads to the conclusion that the Dennis–Dorn–Wolfe quadratic dual programs [3], [4], [11] are meaningful only if the underlying matrix is positive definite or semidefinite. For example, even if the primal quadratic problem (1a) has a unique global minimum solution (thus satisfying the second order sufficient optimality condition), and if the underlying matrix is not positive semidefinite, then the dual quadratic problem (1b) may not have a solution. Thus the example,

$$\text{Minimize } x_1^2 - x_2^2, \quad \text{subject to } x_2 = 0,$$

has the unique global solution $x_1 = x_2 = 0$, but its dual,

$$\text{Maximize } x_1^2 - x_2^2 + vx_2, \quad \text{subject to } x_1 = 0, \qquad -2x_2 + v = 0,$$

is unbounded above. Similarly, the Wolfe dual for nonlinear programs may not be meaningful unless the Hessian of the Lagrangian is locally positive definite or semidefinite in the neighborhood of a stationary point of the primal problem [7]. Thus even if the second order sufficient optimality conditions are satisfied but the Hessian of the Lagrangian is not positive definite or semidefinite in a neighborhood of a local minimum solution, the dual problem may not have a solution.

We shall need second order optimality conditions for the dual quadratic programs (1a) and (1b) which have local and strictly local solutions. These results can be found in [9], [2], [1], which we summarize here in a convenient form. The points $(\bar{x}, \bar{u}, \bar{v}) \in R^{n+m+k}$ and $(\bar{x}, \bar{u}, \bar{v}, \bar{w}) \in R^{n+m+k+n}$ are Karush–Kuhn–Tucker points of (1a) and (1b), respectively, if they satisfy the following respective conditions [10]:

$$\begin{array}{llll}
(4a) & Q\bar{x} + A^T\bar{u} + C^T\bar{v} + p = 0, & (4b) & -Q\bar{x} + Q\bar{w} = 0, \\
 & A\bar{x} \leqq b, & & A\bar{w} - b \leqq 0, \\
 & C\bar{x} = d, & & C\bar{w} - d = 0, \\
 & \bar{u} \geqq 0, & & Q\bar{x} + A^T\bar{u} + C^T\bar{v} + p = 0, \\
 & \bar{u}^T(A\bar{x} - b) = 0, & & \bar{u} \geqq 0, \\
 & & & \bar{u}^T(A\bar{w} - b) = 0.
\end{array}$$

Note that if $(\bar{x}, \bar{u}, \bar{v})$ is a Karush–Kuhn–Tucker point of (1a), then $(\bar{x}, \bar{u}, \bar{v}, \bar{x})$ is a Karush–Kuhn–Tucker point of (1b). To characterize local solutions we need to define the following index sets associated with a Karush–Kuhn–Tucker point $(\bar{x}, \bar{u}, \bar{v})$ of (1a):

$$J := \{i \mid A_i\bar{x} = b_i, \bar{u}_i > 0\},$$
$$K := \{i \mid A_i\bar{x} = b_i, \bar{u}_i = 0\},$$
$$I := \{i \mid A_i\bar{x} < b_i, \bar{u}_i = 0\}.$$

The notation A_J will represent the rows A_i of A with $i \in J$. We can now state the following:

THEOREM 1.3 [2], [1] (characterization of local solutions of quadratic programs). *A point $\bar{x} \in R^n$ is a local minimum solution of the quadratic program* (1a) *if and only if $\bar{x}$ and some $(\bar{u}, \bar{v}) \in R^{m+k}$ satisfy the Karush–Kuhn–Tucker conditions* (4a) *and*

$$\text{(5a)} \qquad A_Jx = 0, \quad A_Kx \leqq 0, \quad Cx = 0 \Rightarrow x^TQx \geqq 0.$$

The Karush–Kuhn–Tucker point $(\bar{x}, \bar{u}, \bar{v})$ of (1a) *is a local maximum solution of the dual quadratic program* (1b) *if and only if*

$$\text{(5b)} \qquad Qx + A^Tu + C^Tv = 0, \quad u_K \geqq 0, \quad u_I = 0 \Rightarrow x^TQx \geqq 0.$$

THEOREM 1.4 [9], [2], [1] (characterization of strict local solutions of quadratic programs). *A point $\bar{x} \in R^n$ is a strict local minimum solution of the quadratic program* (1a) *if and only if $\bar{x}$ and some $(\bar{u}, \bar{v}) \in R^{m+k}$ satisfy the Karush–Kuhn–Tucker conditions* (4a) *and*

$$\text{(6a)} \qquad A_Jx = 0, \quad A_Kx \leqq 0, \quad Cx = 0, \quad x \neq 0 \Rightarrow x^TQx > 0.$$

The Karush–Kuhn–Tucker point $(\bar{x}, \bar{u}, \bar{v})$ of (1a) *is a strict local maximum solution of the dual quadratic program* (1b) *if and only if*

$$\text{(6b)} \qquad Qx + A^Tu + C^Tv = 0, \quad u_K \geqq 0, \quad u_I = 0, \quad (x, u, v) \neq 0 \Rightarrow x^TQx > 0.$$

In the next two sections we characterize positive definite and semidefinite problems in terms of equality-constrained quadratic programs (§ 2) and inequality-constrained quadratic programs (§ 3). This split into equality- and inequality-constrained problems permits the statement of somewhat sharper results for the former. For simplicity we confine the results of § 3 to inequality constraints only. Problems with both equality and inequality constraints can be handled in a straightforward extension of the results of § 3.

2. Equality-constrained quadratic programs. We specialize here the dual problems (1a) and (1b) to the following equality-constrained dual quadratic programs:

$$\text{(7a)} \quad \begin{array}{ll} \underset{x}{\text{Minimize}} & \frac{1}{2}x^TQx + p^Tx \\ \text{subject to} & Cx = d, \end{array} \qquad \text{(7b)} \quad \begin{array}{ll} \underset{x,v}{\text{Maximize}} & -\frac{1}{2}x^TQx - d^Tv \\ \text{subject to} & Qx + C^Tv + p = 0. \end{array}$$

We say that a problem is feasible if the set of points satisfying its constraints is nonempty.

THEOREM 2.1 (characterization of positive semidefinite and definite matrices). *Let* (7a) *be feasible.*

(i) *Let* (7b) *be feasible. A necessary and sufficient condition for Q to be positive semidefinite is that* (7a) *has a local minimum solution,* (7b) *has a local maximum solution and*

$$x^TQx=0,\ Cx=0\Rightarrow Qx=0. \tag{8}$$

(ii) *A sufficient condition for Q to be positive definite is that* (7a) *has a strict local minimum solution and* (7b) *has a strict local maximum solution. This condition is also necessary if C has linearly independent rows.*

Proof. (i) Necessity follows from existence and the duality theory of convex quadratic programming [5], [10]. We establish sufficiency now by means of Theorem 1.1. Define

$$K:=\{x\,|\,Cx=0\}. \tag{9}$$

Then

$$\begin{aligned} K^Q&:=\{y\,|\,y^TQx\leqq 0,\ \forall x\in K\}\\ &=\{y\,|\,y^TQx>0,\ Cx=0 \text{ has no solution } x\}\\ &=\{y\,|\,Qy+C^Tv=0\}. \end{aligned} \tag{10}$$

Since (7a) has a local minimum solution, it follows by Theorem 1.3, (5a) and (9) that

$$x^TQx\geqq 0 \quad \text{for } x\in K. \tag{11}$$

Since (7b) has a local maximum solution, it follows also by Theorem 1.3, (5b) and (10) that

$$y^TQy\geqq 0 \quad \text{for } y\in K^Q. \tag{12}$$

Hence by (11), (8), (12) and Theorem 1.1, Q is positive semidefinite.

(ii) *Necessity.* That both (7a) and (7b) have solutions follows from the feasibility of (7a) and the positive definiteness of Q. The uniqueness of solution for (7a) follows from the positive definiteness of Q. The uniqueness of solution for (7b) follows from the positive definiteness of Q, the linear independence of the rows of C and Theorem 1.4, (6b).

Sufficiency. We establish sufficiency by means of Theorem 1.2. Since (7a) has a strict local minimum solution, it follows by Theorem 1.4, (6a) and (9) that

$$x^TQx>0 \quad \text{for } 0\neq x\in K. \tag{13}$$

Since (7b) has a strict local maximum solution, it follows also by Theorem 1.4, (6a) that

$$x^TQx>0 \quad \text{for } Qx+C^Tv=0,\quad (x,v)\neq 0,$$

and hence by (10)

$$y^TQy>0 \quad \text{for } 0\neq y\in K^Q. \tag{14}$$

Hence by (13), (14) and Theorem 1.2, Q is positive definite. □

3. Inequality-constrained quadratic programs. We turn our attention now to the following inequality constrained dual quadratic programs:

$$\text{(15a)}\quad \underset{x}{\text{Minimize}}\ \tfrac{1}{2}x^TQx+p^Tx \qquad \text{(15b)}\quad \underset{x,u}{\text{Maximize}}\ -\tfrac{1}{2}x^TQx-b^Tu$$
$$\text{subject to}\ \ Ax\leqq b, \qquad\qquad \text{subject to}\ \ Qx+A^Tu+p=0,\quad u\geqq 0.$$

THEOREM 3.1 (characterization of positive semidefinite and definite matrices). *Let* (15a) *be feasible.*

(i) *Let* (15b) *be feasible. A necessary and sufficient condition for Q to be positive semidefinite is that* (15a) *has a local minimum solution $\bar{x}$ with multiplier $\bar{u}$, that $(\bar{x},\bar{u})$ is a local maximum solution of* (15b) *and*

$$\text{(16)}\qquad x^TQx=0,\quad A_Jx=0,\quad A_Kx\leqq 0\Rightarrow Qx=0.$$

(ii) *A sufficient condition for Q to be positive definite is that* (15a) *has a strict local minimum solution $\bar{x}$ with multiplier $\bar{u}$, and $(\bar{x},\bar{u})$ is a strict local maximum solution of* (15b). *If in addition the rows of A_J are linearly independent, and $A_Jx=0$, $A_Kx>0$ has a solution, then this condition is also necessary.*

Proof. (i) Necessity follows from existence and the duality theory of convex quadratic programs. We establish sufficiency now by means of Theorem 1.1. Define

$$\text{(17)}\qquad K:=\{x\,|A_Jx=0, A_Kx\leqq 0\}.$$

Then

$$K^Q=\{y\,|y^TQx\leqq 0, \forall x\in K\}=\{y\,|y^TQx>0, A_Jx=0, A_Kx\leqq 0, \text{has no solution } x\}$$
$$=\{y\,|Qy-A_J^Tu_J-A_K^Tu_K=0, u_K\geqq 0\}.$$

Therefore

$$\text{(18)}\qquad -K^Q=\{x\,|Qx+A^Tu=0, u_K\geqq 0, u_I=0\}.$$

Since $\bar{x}$ is a local minimum solution of (15a) with multiplier $\bar{u}$, it follows from Theorem 1.3, (5a) and (17) that

$$\text{(19)}\qquad x^TQx\geqq 0\quad \text{for } x\in K.$$

Since $(\bar{x},\bar{u})$ is also a local maximum solution of (15b), it follows from Theorem 1.3, (5b) and (18) that $x^TQx\geqq 0$ for $x\in -K^Q$, which is equivalent to

$$\text{(20)}\qquad x^TQx\geqq 0\quad \text{for } x\in K^Q.$$

Conditions (19), (16), (20) and Theorem 1.1 imply that Q is positive semidefinite.

(ii) *Necessity.* That both (15a) and (15b) have solutions follows from the feasibility of (15a) and the positive definiteness of Q. The uniqueness of the solution of (15a) follows from the positive definiteness of Q. The uniqueness of the solution of (15b) follows from the positive definiteness of Q, the linear independence of the rows of A_J, the existence of a solution to $A_Jx=0$, $A_Kx>0$ and Theorem 1.4, (6b).

Sufficiency. We establish sufficiency by means of Theorem 1.2. Since (15a) has a strict local minimum solution $\bar{x}$, it follows by Theorem 1.4, (6a) and (17) that

$$\text{(21)}\qquad x^TQx>0\quad \text{for } 0\neq x\in K.$$

Since $(\bar{x},\bar{u})$ is a strict local maximum solution of (15b), it follows from Theorem 1.4, (6b) and (18) that

$$\text{(22)}\qquad x^TQx>0\quad \text{for } 0\neq x\in K^Q.$$

Hence by (21), (22) and Theorem 1.2, Q is positive definite. □

COROLLARY 3.2 (globalization of local dual solutions).

(i) *Let $\bar{x}$ be a local minimum solution of* (15a) *with multiplier $\bar{u}$, let $(\bar{x}, \bar{u})$ be a local maximum solution of* (15b) *and let* (16) *hold. Then Q is positive semidefinite and hence $\bar{x}$ is a global minimum solution of* (15a), *and $(\bar{x}, \bar{u})$ is a global maximum solution of* (15b).

(ii) *Let $\bar{x}$ be a strict local minimum solution of* (15a) *with multiplier $\bar{u}$, and let $(\bar{x}, \bar{u})$ be a strict local maximum solution of* (15b). *Then Q is positive definite and hence $\bar{x}$ is a unique global minimum solution of* (15a), *and $(\bar{x}, \bar{u})$ is a global maximum solution of* (15b).

We note that condition (16) of Theorem 3.1 cannot be dispensed as shown by the following pair of dual programs:

$$\begin{array}{llll} \text{Minimize} & x_1x_2 & \text{Maximize} & -x_1x_2, \\ \text{subject to} & x_1 \geqq 0, & \text{subject to} & x_1 - u_2 = 0, \\ & x_2 \geqq 0, & & x_2 - u_1 = 0, \\ & & & (u_1, u_2) \geqq 0. \end{array}$$

Clearly $(x_1, x_2) = (0, 0)$ is a global solution to the primal problem, and $(x_1, x_2, u_1, u_2) = (0, 0, 0, 0)$ is a Karush–Kuhn–Tucker point for the primal problem as well as a global solution to the dual problem. However, the underlying matrix $Q = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ is not positive semidefinite, because condition (16) is violated.

We establish now other duality results which induce positive definiteness or semidefiniteness. We begin with two preliminary results.

LEMMA 3.3. *Let $(\bar{x}, \bar{u})$ satisfy the Karush–Kuhn–Tucker conditions of* (15a). *Then*

$$\tfrac{1}{2}\bar{x}^TQ\bar{x} + p^T\bar{x} \geqq -\tfrac{1}{2}x^TQx - b^Tu \tag{23}$$

implies that

$$-\tfrac{1}{2}\bar{x}^TQ\bar{x} - b^T\bar{u} \geqq -\tfrac{1}{2}x^TQx - b^Tu. \tag{24}$$

Proof. From the Karush–Kuhn–Tucker conditions of (15a) we have that

$$-\bar{x}^TQ\bar{x} - p^T\bar{x} - b^T\bar{u} = 0,$$

which when added to (23) yields (24). □

LEMMA 3.4. *Let $(\bar{x}, \bar{u})$ satisfy the Karush–Kuhn–Tucker conditions of* (15a) *such that for all (x, u) feasible for the dual quadratic program* (15b),

$$\tfrac{1}{2}\bar{x}^TQ\bar{x} + p^T\bar{x} \geqq -\tfrac{1}{2}x^TQx - b^Tu. \tag{25}$$

Then $(\bar{x}, \bar{u})$ solves (15b).

Proof. Since $(\bar{x}, \bar{u})$ is feasible for the dual quadratic program (15b), and since by (25) and Lemma 3.3

$$-\tfrac{1}{2}\bar{x}^TQ\bar{x} - b^T\bar{u} \geqq -\tfrac{1}{2}x^TQx - b^Tu$$

for all dual feasible (x, u), it follows that $(\bar{x}, \bar{u})$ solves (15b). □

THEOREM 3.5 (sufficient condition for positive definiteness). *If a strict local minimum of the quadratic program* (15a) *exceeds or equals a unique global maximum of the dual quadratic program* (15b), *then Q is positive definite.*

Proof. Let $\bar{u}$ be a multiplier associated with the strict local minimum solution of (15a). By Lemma 3.4, $(\bar{x}, \bar{u})$ is a global maximum solution of (15b). By assumption this global maximum is unique. Hence by Theorem 3.1(ii), Q is positive definite. □

THEOREM 3.6 (characterization of positive semidefinite matrices). *Let $\bar{x}$ be a local minimum solution of* (15a). *The matrix Q is positive semidefinite if and only if* (16) *holds, and for any dual feasible* (x, u),

$$\tag{26} \tfrac{1}{2}\bar{x}^TQ\bar{x}+p^T\bar{x} \geqq -\tfrac{1}{2}x^TQx-b^Tu.$$

Proof. Necessity follows from the duality theory of quadratic programming and from the fact that $x^TQx=0$ implies $Qx=0$ for any symmetric positive semidefinite matrix. To prove sufficiency, we note that there exists a $\bar{u}$ such that $(\bar{x}, \bar{u})$ is a Karush–Kuhn–Tucker point of (15a), and that by (26) and Lemma 3.4, $(\bar{x}, \bar{u})$ is a global maximum solution to (16b). Hence by Theorem 3.1(i), Q is positive semidefinite. □

A direct consequence of Theorem 3.6 is the following characterization of positive semidefinite matrices in terms of the weak duality [10] relation of quadratic programs.

COROLLARY 3.7 (positive semidefiniteness via weak duality). *Let the quadratic programs* (15a) *and* (15b) *be feasible. The matrix Q is positive semidefinite if and only if for all primal feasible x and all dual feasible* (y, u),

$$\tag{27} \tfrac{1}{2}x^TQx+p^Tx \geqq -\tfrac{1}{2}y^TQy-b^Tu$$

and

$$\tag{28} z^TQz=0 \Rightarrow Qz=0.$$

REFERENCES

[1] J. M. BORWEIN, *Necessary and sufficient conditions for quadratic minimality*, Dept. Mathematics Carnegie-Mellon Univ., Pittsburgh, November, 1981.

[2] L. CONTESSE, *Une caractérisation complète des minima locaux en programmation quadratique*, Numer. Math., 34 (1980), pp. 315–332, Theorem 1″.

[3] J. B. DENNIS, *Mathematical Programming and Electrical Networks*, Wiley, New York, 1959.

[4] W. S. DORN, *Duality in quadratic programming*, Quart. Appl. Math., 18 (1960), pp. 155–162.

[5] B. C. EAVES, *On quadratic programming*, Management Sci., 17 (1971), pp. 698–711.

[6] A. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[7] O. FUJIWARA, S.-P. HAN AND O. L. MANGASARIAN, *Local duality of nonlinear programs*, MRC Tech. Sum. Rep. 2329, Mathematics Research Center, Univ. Wisconsin, Madison, February, 1982, SIAM J. Control Optim., 22 (1984), pp. 162–169.

[8] S.-P. HAN AND O. L. MANGASARIAN, *Conjugate cone characterization of positive and semidefinite matrices*, Tech. Rep. 471, Computer Sciences Dept., Univ. Wisconsin, Madison, March, 1982, Linear Algebra and App., to appear.

[9] O. L. MANGASARIAN, *Locally unique solutions of quadratic programs, linear and nonlinear complementarity problems*, Math. Programming, 19 (1980), pp. 200–212.

[10] ———, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[11] P. WOLFE, *A duality theorem for nonlinear programming*, Quart. Appl. Math., 19 (1961), pp. 239–244.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0006 $01.25/0

A COMBINED DIRECT-ITERATIVE METHOD FOR CERTAIN *M*-MATRIX LINEAR SYSTEMS*

R. E. FUNDERLIC† AND R. J. PLEMMONS‡

Abstract. Large, sparse, irreducible singular (column diagonal dominant) M-matrices A occur in various applications including queueing networks, input-output analysis and compartmental analysis. Our splitting $A = M - N$ with the matrix M having symmetric zero structure is a regular splitting, and these splittings induce a combined direct-iterative solution to $Ax = 0$. A sparse LU factorization of a symmetric permutation of A can be obtained using a standard symmetric ordering scheme such as minimum degree. No pivoting for stability is necessary. Splitting strategies based on a tolerance factor are also discussed and some numerical experience is given.

Key words. M-matrix, queueing networks, homogeneous linear system, sparse matrix, preconditioning, regular splitting

1. Introduction.

This paper is concerned with regular splitting methods for solving linear systems $Ax = b$, where A is a certain M-matrix. M-matrices have many applications in the mathematical sciences (see e.g. Berman and Plemmons [1979, Chaps. 6–10] and Funderlic and Plemmons [1981, Introduction]). Since the method described here includes an iterative part, our emphasis is thus on solving large sparse systems. Such problems occur for example in queueing network (see e.g. Kaufman [1983]) and input-output economic problems (see e.g. Berman and Plemmons [1979, Chap. 9]). These problems give rise to singular systems $Ax = 0$, where A is an M-matrix with other useful properties described in § 2.

A real $n \times n$ *M-matrix* $A = (a_{ij})$ can be defined by the conditions

$$a_{ij} \leqq 0, \qquad i \neq j, \tag{1.1}$$

$$\operatorname{Re}[\lambda_i(A)] \geqq 0, \qquad \lambda_i(A) \text{ the eigenvalues of } A. \tag{1.2}$$

It follows that the diagonal elements a_{ii} of A are all nonnegative and A is nonsingular if and only if strict inequality holds in (1.2) for each i. A large number of alternative definitions are possible and M-matrices have many interesting properties (e.g., Berman and Plemmons [1979, Chap. 6]).

The regular splitting method is described in § 2. The theorem given there along with the related comments on error analysis put the method on practical ground. Section 3 contains a discussion of our initial numerical experience.

2. A combined direct-iterative method.

Although our results have applications for nonsingular problems, our main purpose is the computation of a vector $x = (x_i)$ such that

$$Ax = 0, \qquad x_i > 0, \quad i = 1, \cdots, n, \tag{2.1}$$

* Received by the editors December 10, 1982. The material in this paper was presented at the Fairfield Glade Sparse Matrix Symposium, Fairfield Glade, Tennessee, October 1982.

† Mathematics and Statistics Research Department, Computer Sciences, Union Carbide Corporation, Nuclear Division, Oak Ridge, Tennessee 37830. The research of this author was sponsored by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy under contract W-7405-eng-26 with the Union Carbide Corporation.

‡ Mathematics and Computer Sciences Departments, North Carolina State University, Raleigh, North Carolina 27650. The research of this author was supported in part by the U.S. Army Research Office under grant DAAG 29-81-k-0132 and by the National Science Foundation under grant MGS-8219500.

where A is an $n\times n$ singular, irreducible M-matrix. Since A has a one-dimensional null space, (2.1) has a unique solution whenever one of the components of x is fixed, that is, whenever x is scaled in some way. In this case we will call this unique x the *steady state vector* for A.

In addition to irreducibility and singularity we will also assume here that A is column diagonally dominant. This means that $e^TA\geqq 0$, $e^T=(1,\cdots,1)$ and thus since A is an irreducible, singular M-matrix, $e^TA=0$ so that

$$a_{jj}=-\sum_{i\neq j} a_{ij}, \qquad j=1,\cdots,n,$$

(see, e.g., Berman and Plemmons [1979, Chap. 6]). Such is the case, for example, if $A=I-Q^T$, where Q is the row stochastic matrix for an ergodic Markov chain. Here the steady state vector x solving (2.1) is normally scaled so that $\sum_{i=1}^n x_i=1$, and thus the x_i represent probabilities. The matrix A is also column diagonally dominant in problems associated with the compartmental analysis of tracer flows (see Funderlic and Mankin [1981]).

The purpose of this section is to discuss a certain regular splitting for A. This will lead to a combined direct-iterative method for computing the steady state vector x satisfying (2.1). General regular splittings are discussed in Rose [1984] (this issue, pp. 133–144).

Although the coefficient matrix A in (2.1) is singular, in principle a solution x can always be obtained in a stable way by Gaussian elimination on A, as described in Funderlic and Mankin [1981] and Funderlic and Plemmons [1981]. Indeed, for each permutation matrix P, PAP^T has an LU factorization

$$PAP^T=LU,$$

where L is an M-matrix with unit diagonal, $-l_{ij}\leqq 1$ for all $i\neq j$, and where U is an upper triangular M-matrix of rank $n-1$ with $u_{nn}=0$. Moreover, the growth factor g_A during the decomposition satisfies

$$g_A=\frac{\max\limits_{i,j,k}|a_{ij}^{(k)}|}{\max\limits_{i,j}|a_{ij}|}=1, \tag{2.2}$$

where $a_{ij}^{(k)}$ denotes the i, j element of the unreduced part of A before the kth step of Gaussian elimination (see Funderlic, Neumann and Plemmons [1982]). Here P can, for example, be chosen to reduce the fill-in during the elimination. To compute x, we then

1) Solve

$$\begin{bmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \vdots \\ & & u_{n-1,n} \\ & & 1 \end{bmatrix} y=\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \delta \end{bmatrix}, \qquad \delta>0,$$

2) Set $x=P^Ty$.

This method will be called the *direct factorization method* for solving (2.1). Harrod and Plemmons [1984] have examined the direct factorization method (together with some variations) in detail and have found it to be very effective for computing steady state vectors. The method.essentially corresponds to a specialized implementation of the first step in inverse iteration as considered by Wilkinson [1965, p. 619].

Indeed, suppose the direct factorization method is executed on a machine with unit roundoff error μ in floating point arithmetic. Harrod and Plemmons [1984] have shown, using a simple backward error analysis, that the computed solution $\hat{x}$ to the problem (2.1) by the direct factorization method satisfies exactly a homogeneous system

$$(A+E)\hat{x}=0,$$

where the unit roundoff error matrix $E=(e_{ij})$ satisfies

$$|e_{ij}|\leqq \mu i(3.02+1.01n)\max_t a_{tt} \tag{2.3}$$

for each i and j. In the particular case where $A=I-Q^T$, Q an irreducible stochastic matrix, (2.3) simplifies to

$$|e_{ij}|\leqq \mu i(3.02+1.01n).$$

However, the difficulty in applying the direct factorization method to the case of large sparse A is that one is generally limited to symmetric pivoting, PAP^T, in order to ensure stability in the factorization. But symmetric pivoting may often not be very effective in limiting fill-in. Alternatively, a nonsymmetric pivoting scheme, PAQ, combined with the use of a threshold pivoting criteria, is used in some sparse matrix packages (see Duff and Reid [1979]). But this scheme can be time consuming and, perhaps more importantly, such a method will generally require a dynamic storage scheme for U. Now if A has a (nearly) symmetric zero structure, i.e. if $a_{ij}=0$ if and only if $a_{ji}=0$ for most i and j, then the standard symmetric ordering schemes, such as those available in the package SPARSPAK (see George and Liu [1981]) can be modified and used quite effectively in reducing fill-in to U. Moreover, such schemes facilitate the use of a static storage scheme for U. Queueing networks where Q (and thus $A=I-Q^T$) has a nearly symmetric zero structure arise, for example, in certain traffic overflow systems (see Kaufman [1983]).

Iterative schemes for computing x can sometimes be very effective. Moreover, for very large-scale problems, storage considerations usually dictate the use of iterative schemes over direct methods. (See Kaufman [1983] for examples from queueing network analysis.) But iterative methods are sometimes plagued by slow convergence. It is in these cases that combined direct-iterative schemes may be attractive. In fact, the direct part of the scheme may be thought of as a method for accelerating the convergence of the iterative part. Various algorithms of this general type have been proposed and investigated in the literature for the special case of nonsingular systems of linear equations. In particular, we point out the work of Meijerink and Van der Vorst [1977], who considered nonsingular M-matrices. Before describing our direct-iterative algorithm, we develop the following terminology and notation.

An $n\times n$ matrix B is said to be *semiconvergent* if

$$\operatorname*{limit}_{k\to\infty} B^k \tag{2.4}$$

exists. If (2.4) is the zero matrix then B is *convergent.* It is well known that B is convergent if and only if the spectral radius, $\rho(B)$, satisfies $\rho(B)<1$, while B is semiconvergent (see, e.g., Neumann and Plemmons [1978]) if and only if $\rho(B)\leqq 1$ and if $\rho(B)=1$, then (a) 1 is an eigenvalue of B and (b) 1 is the only eigenvalue of B on the unit circle and (c) all the elementary divisors associated with 1 are linear.

If $\rho(B)=1$ and (a) and (c) are satisfied then it is easy to show that

$$B_\alpha=(1-\alpha)I+\alpha B \tag{2.5}$$

has no eigenvalues on the unit circle (and thus satisfies (b)) for any real $0<\alpha<1$ (see, e.g., Neumann and Plemmons [1978]). In particular then, B_α is semiconvergent.

Suppose B is semiconvergent. We define

$$\gamma(B)=\max\{|\lambda|\lambda \text{ is an eigenvalue of } B \text{ and } \lambda\neq 1\}.$$

In this case, $\gamma(B)$ is the controlling factor in the convergence of the powers of B in (2.5) (see Neumann and Plemmons [1978]).

Next,

$$A=M-N \tag{2.6}$$

is a *splitting* of A if M is nonsingular. It is called a *regular splitting* of A if M^{-1} and N are nonnegative matrices (see Varga [1962]). For the system (2.1), the splitting (2.6) induces the iteration scheme

$$Mx^{(k+1)}=Nx^{(k)}, \qquad k=0,1,\cdots. \tag{2.7}$$

It follows that the iterative scheme (2.7) produces a sequence of vectors which converges to a solution of (2.1) for each initial approximation vector $x^{(0)}$, if and only if $M^{-1}N$ is semiconvergent. In this case, the *asymptotic rate of convergence* is the parameter

$$-\ln[\gamma(M^{-1}N)].$$

Thus the smaller $\gamma(M^{-1}N)$ is, the faster the convergence of (2.7) can be expected to be. Thus, ideally, one would like to have good separation between 1 and the other eigenvalues (in absolute value) of $M^{-1}N$.

We are interested in particular splittings of A. By a *symmetric zero structure matrix* of A we mean a matrix S such that $s_{ii}=a_{ii}$ and the off-diagonal elements of $S=(s_{ij})$ are such that $s_{ij}=a_{ij}$ or $s_{ij}=0$. If $s_{ij}\neq 0$ then $a_{ij}a_{ji}\neq 0$ and both $s_{ij}=a_{ij}$ and $s_{ji}=a_{ji}$. In some cases the (entire) *symmetric zero structure matrix* of A is useful. This matrix S is defined from A by

$$s_{ij}=\begin{cases} a_{ij} & \text{if } a_{ij}a_{ji}\neq 0,\\ 0 & \text{otherwise.}\end{cases}$$

Observe that if A is a singular, irreducible M-matrix, for example if $A=I-Q^T$, Q an irreducible stochastic matrix, then a symmetric zero structure matrix S is such that $s_{ii}=a_{ii}>0, i=1,\cdots,n$, since A has all positive entries on its diagonal. Our purpose is to consider the splittings (2.6) with M a symmetric zero structure matrix of A, and to analyze the convergence of (2.7). The following theorem establishes the convergence results we need. Probably the most important practical point of the theorem is that if A is an irreducible singular M-matrix and M is a matrix obtained by setting any off-diagonal elements of A to zero, then M is nonsingular.

THEOREM 1. *Let A be an irreducible (possibly singular) M-matrix. If $A\leqq M\neq A$, then*

(1) *$A=M-N$ is a regular splitting, and*

(2) *for any $0<\alpha<1$, the iteration*

$$Mx^{(k+1)}=[(1-\alpha)M+\alpha N]x^{(k)} \tag{2.8}$$

converges to a solution x of (2.1) *where $x_i > 0$, $i = 1, \cdots, n$. Moreover if $M^{-1}N$ has no eigenvalue on the unit circle other than* 1, *then the iteration* (2.8) *converges with $\alpha = 1$.*

Proof. Statement (1) will follow from well-known (see Berman and Plemmons [1979, Chap. 6]) properties of M-matrices. Since $A \leqq M$ and A is an M-matrix, it follows that M is an M-matrix and that $0 \leqq \det A \leqq \det M$. But if $\det M = \det A$ and A is irreducible, then $M = A$. Since $M \neq A$ it follows that $\det M > 0$ and M is a nonsingular M-matrix. Thus M is nonsingular and the splitting is regular since $M^{-1} \geqq 0$ and $N = M - A \geqq 0$. Finally, statement (2) follows from Neumann and Plemmons [1978, § 4].

We point out that past experience has shown (see Kaufman [1983]) that in many practical situations, we can set $\alpha = 1$ in (2.8). It can be shown that such is the situation if $M^{-1}N$ has no zeros on its diagonal, or if M is irreducible (see Rose [1983]). However, if $M^{-1}N$ does have eigenvalues on the unit circle other than 1, then methods for choosing α, $0 < \alpha < 1$, to minimize

$$\gamma((1-\alpha)I + \alpha M^{-1}N)$$

are given for certain cases in Neumann and Plemmons [1978].

Of particular interest to us is the case where M is chosen to be a symmetric zero structure matrix S for A. Since $A \leqq S$ we have:

COROLLARY 1. *Let A be a singular, irreducible M-matrix and let S be a symmetric zero structure matrix for A. If $S \neq A$, then Theorem* 1 *holds for $M = S$.*

Let $N = S - A$ and assume for simplicity that $S^{-1}N$ has no eigenvalues other than 1 on the unit circle. We then consider the following algorithm.

SYMMETRIC ZERO SPLITTING ALGORITHM.

1) Factor

$$PSP^T = LU$$

where L and U are lower and upper triangular matrices, respectively, and where P is the permutation matrix reflecting the symmetric pivoting used to reduce the fill-in in L and U.

2) Choose an initial vector $x^{(0)}$ and for $k = 0, 1, \cdots$ solve

$$LUPx^{(k+1)} = (PN)x^{(k)} \tag{2.9}$$

by forward and backward substitution.

In general there are two extreme possibilities for S. If $S = A$ and is not too dense, then x is simply calculated using the direct factorization method as described earlier. The other extreme situation is where A has no symmetric zero structure, in which case $S = \operatorname{diag}(a_{11}, \cdots, a_{nn})$ and (2.9) reduces to the usual Jacobi method.

The results of some numerical experiments with this algorithm are described in the next section. Some comparisons are made with another direct-iterative method and with the Gauss–Seidel iterative method on a selection of test problems. The other direct-iterative method is based upon the following idea which has been used, for example, by Jennings [1981] for symmetric positive definite problems. A tolerance factor f is chosen and $M = (m_{ij})$ is defined as follows:

$$m_{ij} = \begin{cases} a_{ij} & \text{if } |a_{ij}| \geqq f \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

With $N = M - A$, the resulting splitting $A = M - N$ is called a *magnitude splitting* of A. Observe here that if $M \neq A$, then A and M satisfy the hypothesis of Theorem 1, and thus this splitting leads to a combined direct-iterative method as in (2.9). However, M may have no symmetric zero structure, so the use of a symmetric ordering scheme to reduce the fill-in in the LU-factorization of M may be ineffective. However, in addition to (2.10) one could require that only those nonzero m_{ij} be chosen such that $a_{ij}a_{ji} \neq 0$. In many cases, though, the basic structure of A will suggest an additional constraint on (2.10).

Of course in the Gauss–Seidel iterative method, M consists of the diagonal of A together with the lower (or upper) part of A. Thus M is already triangular and no LU-factorization is necessary.

3. Computational aspects. Our purpose here is not to give a comprehensive numerical treatment of the iterative method based on symmetric zero structure splittings. Our computational experience with this method is of a preliminary nature and therefore only sketchy comparisons are now available. Part of the reason for this lack of a more thorough treatment is that in most applications practitioners do not explicitly form the relevant matrices. Some exceptions are the work of Kaufman, e.g. [1983], and the work of W. Stewart, e.g. [1978]. However, the description of the A matrix is sometimes unwieldy and good examples in the queueing literature are as yet rare.

Most of our experience has been with contrived very small examples and one matrix of order 84. The latter's form arose from a queueing network analysis of a job line production model studied in the Industrial Engineering Department at North Carolina State University. The nonzero off-diagonal entries of the matrix (see Fig. 1)

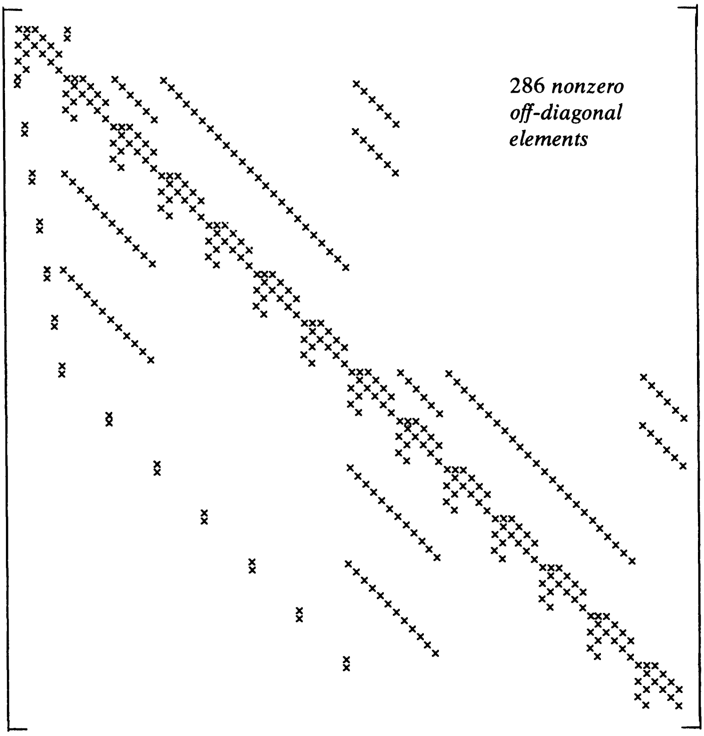
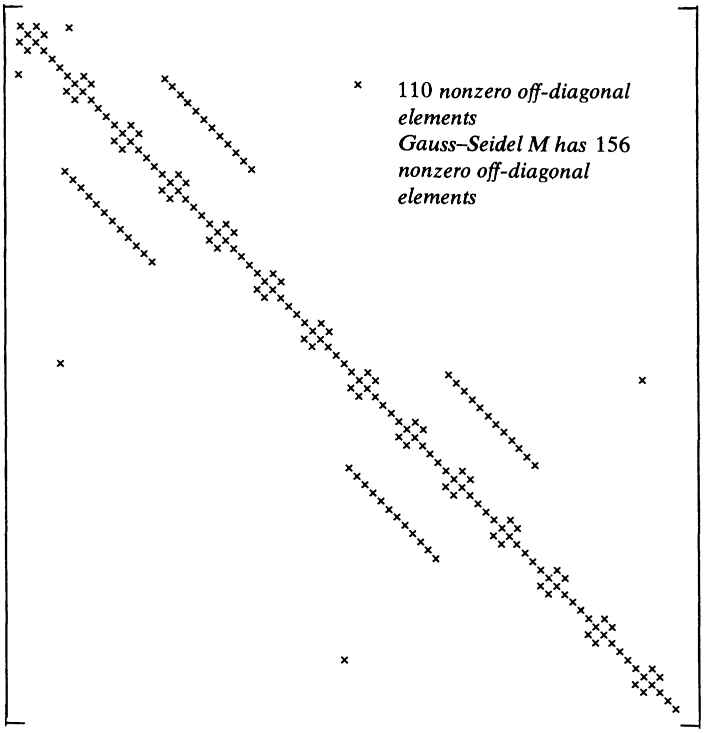


FIG. 1. *Job shop* 84×84 *matrix does not have significant nonzero symmetric structure.*

were chosen as random numbers between -1 and 0. The diagonal elements were chosen so that the column sums of the A matrix are zero. The jobshop matrix of Fig. 1 has 286 nonzero off-diagonal elements. The symmetric zero structure matrix, Fig. 2, has only 110 nonzero off-diagonal elements. A symmetric zero matrix M was chosen with 72 nonzero elements, the locations of which were chosen at random from the symmetric zero structure matrix given by Fig. 2. The splitting with more nonzero elements in M converged, as expected, in fewer iterations than the splitting with fewer elements in M; see Fig. 3. However, in contrast to the nonsingular case (see Varga [1962]), it is possible for a fuller M to cause slower convergence (see Kaufman [1983]). In fact, Schneider has given an example (see Buoni, Neumann and Varga [1982]) where the Jacobi method converges whereas the Gauss–Seidel method does not.



110 *nonzero off-diagonal elements*
Gauss–Seidel M has 156 *nonzero off-diagonal elements*

FIG. 2. *The zero-symmetric structure of job shop matrix.*

Comparison was made with the Gauss–Seidel method where M was chosen as the upper triangle of the job shop matrix. In this case the number of nonzero off-diagonal elements is 156 in contrast to 110 for the symmetric zero structure matrix. The Gauss–Seidel method converged in slightly fewer iterations than the method based on the symmetric zero structure matrix. In particular, after 80 iterations the latter method produced on average slightly more than 5 digits of accuracy in the solution and Gauss–Seidel slightly more than 6. Another experiment with the job shop problem of Fig. 1 was to give it additional nonzero symmetric structure. That is, positions were chosen at random such that in the original matrix $a_{ij} \neq 0$ but $a_{ji} = 0$.

These latter elements were set at nonzero and the diagonal elements were adjusted so that the column sums were zero. In this modified problem the method based on symmetric zero structure had 160 nonzero elements in its matrix M, an increase from 110, and the Gauss-Seidel method had 165, an increase from 156. Here the Gauss-Seidel method converged slower than the symmetric zero form method. This suggests that, for problems in which there are significantly more elements in M than in the Gauss-Seidel M and where fill-in and hence the cost of each iteration is not significant, the method of this paper may be attractive. In addition, there may be cases where the spectrum of the Gauss-Seidel iteration matrix $M^{-1}N$ is unfavorable (see Rose [1984]). An additional point of consideration is that the basic structure of the A matrix may suggest various symmetric zero structure matrices M. For example it is natural to exclude the four outliers in the matrix of Fig. 2.

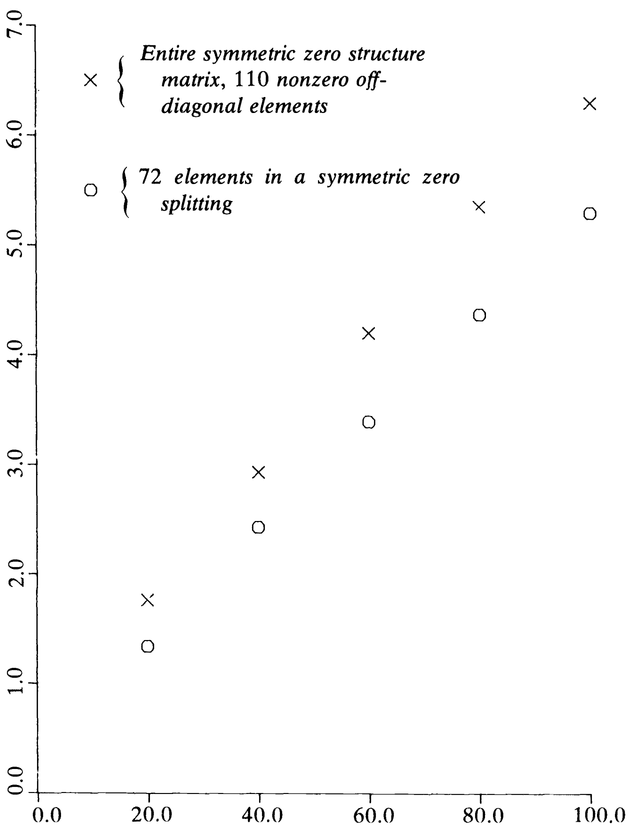


FIG. 3. *Average number of correct digits versus number of iterations for two symmetric zero splittings.*

It has been suggested by Jennings [1981] and others that the larger (in magnitude) elements of the A matrix could be split into the matrix M with benefit. Figure 4 shows a numerical comparison of the largest 113 off-diagonal elements of the job shop matrix that have been put in M versus the 113 smallest in M. The indication is that this type of splitting may have useful applications and may be combined with other splittings to provide effective direct-iterative algorithms for computing steady state vectors.

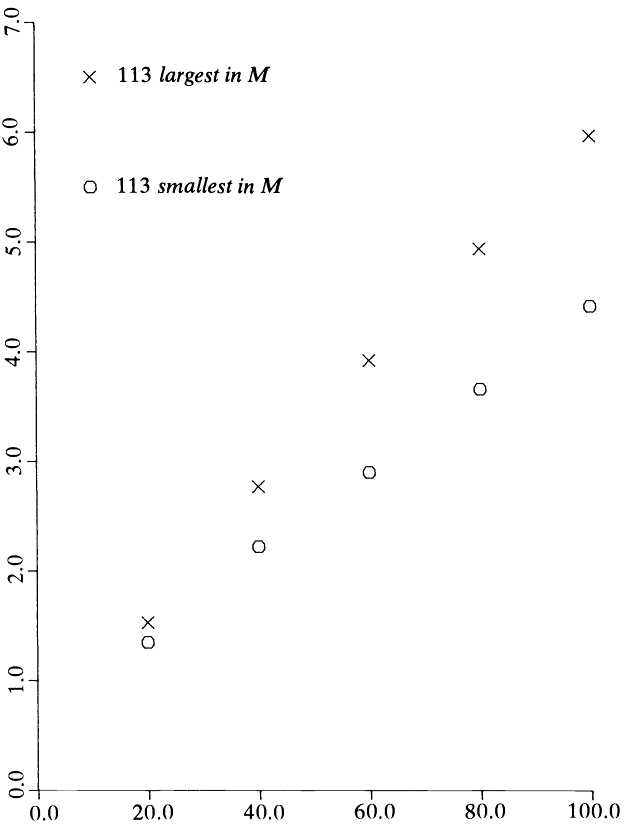


FIG. 4. *Average number of correct digits versus iterations for two magnitude splittings.*

REFERENCES

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

I. DUFF AND J. REID [1979], *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 2, pp. 18–35.

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

R. FUNDERLIC AND R. PLEMMONS [1981], *LU decomposition of M-matrices by elimination without pivoting*, Linear Algebra Appl., 41, pp. 99–110.

R. FUNDERLIC, M. NEUMANN AND R. PLEMMONS [1982], *LU decomposition of generalized diagonally dominant matrices*, Numer. Math., 40, pp. 57–69.

A. GEORGE AND J. LIU [1981], *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ.

W. HARROD AND R. PLEMMONS [1984], *Comparison of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Stat. Comp., to appear.

A. JENNINGS [1981], *Development of an* ICCG *algorithm for large sparse systems*, Preprint.

L. KAUFMAN [1983], *Matrix methods for queueing problems*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 525–552.

J. MEIJERINK AND H. VAN DER VORST [1973], *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31, pp. 148–162.

C. MEYER AND R. PLEMMONS [1977], *Convergent powers of a matrix with applications to iterative methods for singular linear systems*, SIAM J. Numer. Anal., 14, pp. 699–705.

M. NEUMANN AND R. PLEMMONS [1978], *Convergent nonnegative matrices and iterative methods for consistent linear systems*, Numer. Math., 31, pp. 265–279.

C. PAIGE, G. STYAN AND P. WACHTER [1975], *Computation of the stationary distributions of a Markov chain*, J. Statist. Comput. Simulation, 4, pp. 173–186.

R. PLEMMONS [1976], *Regular splittings and the discrete Neumann problem*, Numer. Math., 25, pp. 153–161.

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

G. W. STEWART [1980], *Computable error bounds for aggregated Markov chains*, Tech. Rept. 901, Computer Science Dept., Univ. of Maryland, College Park, MD.

W. J. STEWART [1978], *A comparison of numerical techniques in Markov modeling*, Comm. ACM, 21, pp. 144–151.

R. VARGA [1962], *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0007 $01.25/0

A HIERARCHICAL REPRESENTATION OF THE INVERSE FOR SPARSE MATRICES*

EUGENIUSZ TOCZYŁOWSKI†

Abstract. We present a representation of the inverse of a matrix for solving large sparse systems of linear algebraic equations $Ax = b$, that arises from the bordering method. We analyse the properties of the method for sparse matrices permuted to a suitable form. The relevant feature of the method is that it creates nonzero elements only in spike columns above the main diagonal. For an $n \times n$ matrix A with τ_0 nonzero elements, the number of nonzero elements in the representation of the inverse satisfies the inequality $\tau < \tau_0 + h \cdot n$, where h is a certain constant. It is proved that, under suitable assumptions, $h \leqq \log_2 s + 1$, where s is the number of spikes. The computation of the hierarchical form of the inverse requires at most $h \cdot \tau$ additions and $h \cdot \tau$ multiplications. For the known representation of the inverse, the solution of the system $Ax = b$ requires no more than $\tau - n$ additions and τ multiplications.

Key words. sparse matrices, Gaussian elimination, fill-in, bordered block lower-triangular form, hierarchical form of the inverse

1. Introduction.

Let us consider a general sparse system of n linear equations

$$Ax = b \tag{1}$$

where $x = (x_1, \cdots, x_n)^T$, $b = (b_1, \cdots, b_n)^T$, and $A = [a_{ij}]$ is an $n \times n$ nonsingular matrix. The best known methods for solving such a system decompose the matrix A into a product $A = P \cdot L \cdot U \cdot Q$, where P and Q are permutation matrices and L, U are lower and upper triangular matrices. This decomposition is equivalent to Gaussian elimination [15]. The permutation matrices P and Q are determined either prior to the start of the Gaussian elimination or in the course of the elimination, in order to ensure sparsity and numerical accuracy of the elimination form of the inverse. An extensive survey of sparse matrix techniques is contained in Duff [2].

In this paper we present an algorithm for solving sparse linear systems which is based on a hierarchical representation of the inverse of the coefficient matrix. The method originates from the bordering method; see Faddeeva [3]. It is apparently connected with nested methods for solving sparse systems of linear equations in both the symmetric case [4] and the nonsymmetric case [11]. In § 2 we present the method for a general system of equations. In the next sections we discuss the properties of the method in the case of sparse matrices. For the proposed method we define in § 3 the most desirable structure of sparse matrices, the hierarchical bordered block lower triangular form (HBBLT). In that section we also give some useful terminology which allows us to analyse the properties of different hierarchical structures of a matrix to obtain the most desirable hierarchical BBLT form. In § 4 we analyse the fill-in and number of computations of the method in the case when the sparse matrix A is in HBBLT form. We show that for a matrix A with τ_0 nonzero elements the fill-in in the hierarchical form of the inverse is at most $h \cdot n$, where h is the height of an appropriate tree of the HBBLT form of A. For the number of computations we show the bound $h \cdot \tau$, where τ is the number of nonzero elements in the representation of the inverse. In § 5 we discuss questions concerning an efficient implementation of the method, some effects of rounding errors and pivoting. In § 6 we give two illustrative examples.

* Received by the editors February 3, 1982, and in revised form April 19, 1983.

† Instytut Automatyki, Politechnika Warszawska, 00-665 Warszawa, Poland.

2. A representation of the inverse of the matrix A. In this section we make no use of possible sparsity of A. In order to compute a form of the inverse of the matrix A we assume that all leading principal submatrices of A, i.e.,

$$A_k = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} \tag{2}$$

are nonsingular, for $k = 1, 2, \cdots, n$. The discussion how to carry out the elimination process and how to realize the pivoting for an arbitrary nonsingular matrix A is given in § 5.

Now, for a given k, where $1 \leqq k \leqq n$, let us consider the system of the first k equations of (1)

$$a_{i1}x_1 + \cdots + a_{ik}x_k + \cdots + a_{in}x_n = a_{i,n+1}, \qquad i = 1, \cdots, k, \tag{3}$$

where, for convenience, we set $a_{i,n+1} = b_i$.

Since A_k is assumed to be nonsingular, premultiplying (3) by A_k^{-1} we obtain the reduced system

$$x_i + \sum_{j=k+1}^{n} a_{ij}^{(k)} \cdot x_j = a_{i,n+1}^{(k)}, \qquad i = 1, \cdots, k, \tag{4}$$

where

$$\begin{bmatrix} a_{1,j}^{(k)} \\ \vdots \\ a_{k,j}^{(k)} \end{bmatrix} = A_k^{-1} \cdot \begin{bmatrix} a_{1,j} \\ \vdots \\ a_{k,j} \end{bmatrix}, \qquad j = k+1, \cdots, n, n+1. \tag{5}$$

At the kth stage of the elimination process the nonzero elements in the kth row below the diagonal and in the kth column above the diagonal are eliminated. Thus the leading principal submatrix A_k is reduced to the identity matrix. We assume that at the end of the kth stage a form of the inverse of A_k is known. Initially for $k = 1$, $A_1 = [a_{11}]$ and $A_1^{-1} = [1/a_{11}]$. At the $(k+1)$st stage, after adding the $(k+1)$st equation of the system (1) to the system (4), we seek a form of the inverse of the principal submatrix A_{k+1}. The matrix A_{k+1} results from bordering the matrix A_k, i.e.,

$$A_{k+1} = \left[\begin{array}{c|c} A_k & h_{k+1} \\ \hline g_{k+1} & d_{k+1} \end{array}\right], \tag{6}$$

where $g_{k+1} = (a_{k+1,1}, \cdots, a_{k+1,k})$, $h_{k+1} = (a_{1,k+1}, \cdots, a_{k,k+1})^T$, $d_{k+1} = a_{k+1,k+1}$. It is easy to verify that

$$A_{k+1} = \begin{bmatrix} A_k & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ g_{k+1} & 1 \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & s_{k+1} \end{bmatrix} \cdot \begin{bmatrix} I_k & r_{k+1} \\ 0 & 1 \end{bmatrix}, \tag{7}$$

where I_k is the identity matrix of order k, $r_{k+1} = A_k^{-1} \cdot h_{k+1}$ and $s_{k+1} = d_{k+1} - g_{k+1} \cdot r_{k+1}$. It follows from (5) that $r_{k+1} = (a_{1,k+1}^{(k)}, \cdots, a_{k,k+1}^{(k)})^T$. Observe also that s_{k+1} is the Schur complement (A_{k+1}/A_k) of (6) [1]. For convenience we set $s_1 = a_{11}$.

It follows from (7) that

$$A_{k+1}^{-1} = \begin{bmatrix} I_k & -r_{k+1} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & s_{k+1}^{-1} \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ -g_{k+1} & 1 \end{bmatrix} \cdot \begin{bmatrix} A_k^{-1} & 0 \\ 0 & 1 \end{bmatrix}. \tag{8}$$

Notice that g_{k+1}, r_{k+1} and s_{k+1}^{-1} are sufficient to give a form of the inverse of the matrix A_{k+1}^{-1}, if a form of the inverse A_k^{-1} is known. To prove this, let us denote $w^{(k)}=(w_1,\cdots,w_k)^T$ and $v^{(k)}=(v_1,\cdots,v_k)^T$. Then, to compute

$$w^{(k+1)}=A_{k+1}^{-1}\cdot v^{(k+1)} \tag{9}$$

for any column vector $v^{(k+1)}=(v_1,\cdots,v_{k+1})^T$, it is sufficient to perform the following calculations

$$w^{(k)}:=A_k^{-1}v^{(k)}, \tag{10}$$

$$w_{k+1}:=(v_{k+1}-g_{k+1}\cdot w^{(k)})\cdot s_{k+1}^{-1}, \tag{11}$$

$$w^{(k)}:=w^{(k)}-w_{k+1}\cdot r_{k+1}. \tag{12}$$

The right side of (10) may be computed in a similar fashion, provided that A_{k-1}^{-1} is known.

Since we have $A_1^{-1}=[1/a_{11}]$, the formulae (10), (11), (12) define a recursive algorithm for the computation of $A^{-1}b=A_n^{-1}b$. Let us gather all the data necessary for the representation of the inverse A^{-1} in the following matrix:

$$T=\left[\begin{array}{c|c|c|c|c|c} s_1^{-1} & r_2 & \cdots & r_k & \cdots & r_n \\ \hline g_2 & s_2^{-1} & & & & \\ \hline \vdots & & \ddots & & & \\ \hline g_k & & & s_k^{-1} & & \\ \hline \vdots & & & & \ddots & \\ \hline g_n & & & & & s_n^{-1} \end{array}\right]. \tag{13}$$

The imbedding of one representation of the inverse in another, recursively applied, constitutes a hierarchical form of the inverse. Thus, for a matrix A, the representation (13) is called the *hierarchical form* of the inverse and is denoted by HFI(A).

THEOREM 1. (i) *Assume that for an $n\times n$ matrix A the form of the inverse* (13) *is given. Then the evaluation of the solution $x=A^{-1}b$ requires no more than $n(n-1)$ additions and n^2 multiplications.*

(ii) *The evaluation of the form of the inverse* (13) *requires n divisions and no more than $(n^3-n)/3$ multiplications and $n^3/3-n^2/2+n/6$ additions.*

Proof. We will prove (i) and (ii) by induction on the order n. For $n=1$ the theorem is true. For the induction step assume that the assertions of the theorem hold for some n. The matrix A_{n+1} of order $n+1$ may be considered as a result of bordering the matrix A_n of order n, i.e.,

$$A_{n+1}=\left[\begin{array}{c|c} A_n & h_{n+1} \\ \hline g_{n+1} & d_{n+1}\end{array}\right]. \tag{14}$$

To prove (i) let us assume that the form (13) of the inverse of A_{n+1} is given. Denote $b^{(n)}=(b_1,\cdots,b_n)$ and $x^{(n)}=(x_1,\cdots,x_n)$. It follows from (10), (11), (12) that $x^{(n+1)}=A_{n+1}^{-1}b^{(n+1)}$ may be computed by

$$x^{(n)}:=A_n^{-1}b^{(n)}, \tag{15}$$

$$x_{n+1}:=(b_{n+1}-g_{n+1}\cdot x^{(n)})\cdot s_{n+1}^{-1}, \tag{16}$$

$$x^{(n)}:=x^{(n)}-x_{n+1}\cdot r_{n+1}. \tag{17}$$

From the inductive assumption, the calculation of $x^{(n)}$ by (15) requires no more than n^2-n additions and n^2 multiplications. The calculations of x_{n+1} by (16) requires no more than n additions and $n+1$ multiplications. The calculation of $x^{(n)}$ by (17) requires no more than n additions and multiplications. This gives no more than $n(n+1)$ additions and $(n+1)^2$ multiplications for the computation of $x^{(n+1)}$, which completes the proof of (i).

To prove (ii) by induction, assume that (ii) holds for some n. It follows from (14) that to compute the form of the inverse A_{n+1}^{-1} it is sufficient to compute $r_{n+1}=A_n^{-1}h_{n+1}$ and s_{n+1}^{-1}, where $s_{n+1}=d_{n+1}-g_{n+1}\cdot r_{n+1}$. From (i), the evaluation of r_{n+1} requires no more than n^2-n additions and n^2 multiplications. The evaluation of s_{n+1} requires no more than n additions and multiplications. Together with the inductive assumption this gives no more than $(n+1)^3/3-(n+1)/3$ multiplications and no more than $(n+1)^3/3-(n+1)^2/2+(n+1)/6$ additions. □

COROLLARY. *The solution of the system* (1) *requires no more than* $n^3/3+n^2-n/3$ *multiplications and* $n^3/3+n^2/2-5n/6$ *additions* (*as in Gaussian elimination*).

The method presented here is a modification of the bordering method [3]. The regular bordering method, however, requires approximately twice as many multiplications and three times as many additions. Moreover, when used for sparse computations, it suffers heavy fill-in.

3. The hierarchical bordered block lower triangular form of sparse matrices. If the system (1) is solved by Gaussian elimination and the matrix A is sparse, sparse matrix techniques may be used to save storage and computational effort. The pivotal strategies which tend to keep fill-in small in the course of the elimination may be classified into two categories: first, those that involve arrangements of rows and columns at each stage of the elimination process [10], [14, p. 23], and second, those that involve a priori row and column permutations to transform the matrix A into various forms that are desirable for Gaussian elimination [14, pp. 33–80]. In this section we shall be concerned with a priori (preassigned) pivotal procedures. One of the most favorable forms for Gaussian elimination is the bordered block lower triangular (BBLT) form

$$(18)\qquad A=\begin{bmatrix} A_{11} & 0 & \cdots & 0 & A_{1p}\\ A_{21} & A_{22} & & \vdots & A_{2p}\\ \vdots & \vdots & \ddots & 0 & \vdots\\ A_{p-1,1} & A_{p-1,2} & & A_{p-1,p-1} & A_{p-1,p}\\ A_{p,1} & A_{p,2} & \cdots & A_{p,p-1} & A_{pp}\end{bmatrix},$$

where the square submatrices on the diagonal, A_{ii}, $i=1,2,\cdots,p-1$, are nonsingular and at least one submatrix A_{ip}, $i\neq p$, is a nonzero matrix.

If in (18), $A_{ip}=0$, $i\neq p$, and the matrices A_{ii}, $i=1,2,\cdots,p$ are nonsingular, then A is said to be in block lower triangular (BLT) form.

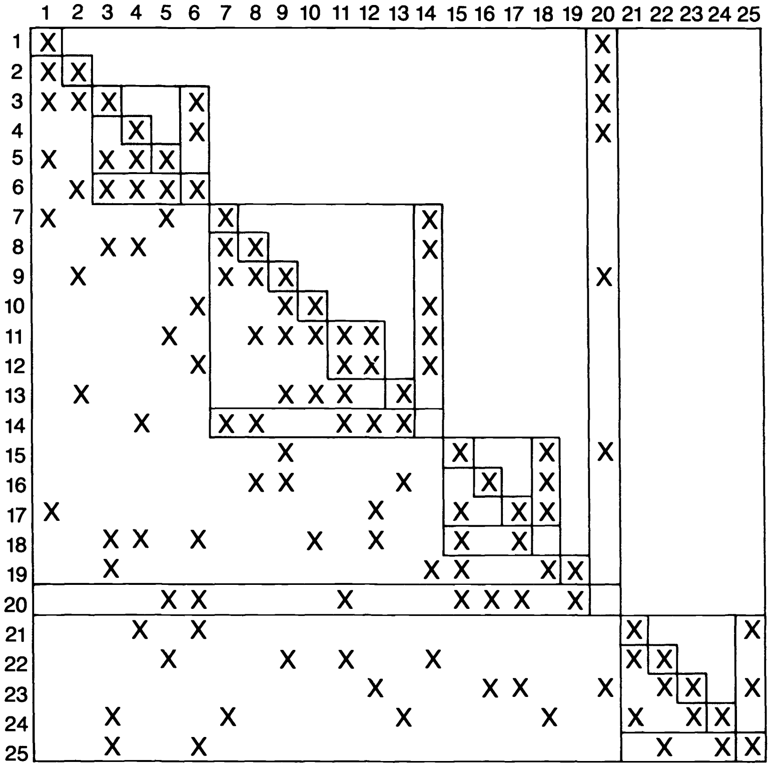
DEFINITION. A nonsingular matrix A is in *hierarchical* BBLT form, if one of the following conditions holds:

(i) A is a nonzero matrix of order one.

(ii) A is in the BBLT form, where the matrix A_{pp} is of order one and each square submatrix on the diagonal, A_{ii}, $i=1,2,\cdots,p-1$, is in the hierarchical BBLT form.

(iii) A is in the BLT form and each square submatrix A_{ii}, $i=1,2,\cdots,p$ is in the hierarchical BBLT form.

An example of a hierarchical structure of a matrix A is shown in Fig. 1. The hierarchical BBLT form is a very desirable form for the elimination method presented in this paper. Such a structure may be identified as follows. We apply the hierarchical partition procedure [9] or the P^4 procedure [5] (or any other procedure which makes use of the Steward partitioning and tearing concepts [12], [8], [6]) to a given matrix $\hat{A}$ without any particular structure. As the result, permutation matrices P and Q are obtained such that the matrix $A = P\hat{A}Q$ is in BLT form with relatively few columns, called spikes, which contain nonzero elements above the main diagonal.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	X																			X					
2	X	X																		X					
3	X	X	X			X														X					
4				X		X														X					
5	X		X	X	X																				
6		X	X	X	X	X																			
7	X				X		X							X											
8			X	X			X	X						X											
9		X					X	X	X											X					
10						X			X	X				X											
11					X			X	X	X	X	X		X											
12						X					X	X		X											
13		X							X	X	X		X												
14				X			X	X			X	X	X												
15									X						X			X		X					
16								X	X				X			X		X							
17	X											X			X		X	X							
18			X	X		X				X		X			X		X								
19			X											X	X			X	X						
20					X	X					X				X	X	X		X						
21				X		X															X				X
22					X				X		X			X							X	X			
23												X				X	X			X		X	X		X
24			X				X						X					X			X		X	X	
25			X			X																X		X	X

FIG. 1. *A matrix A in the hierarchical* BBLT *form* (*x represents nonzero*).

To define the most desirable hierarchical BBLT form of a matrix A let us introduce some terminology. Let $V = \{1, 2, \cdots, n\}$ be the index set for columns (rows) of the matrix $A = [a_{ij}]$. The set V will be identified with the set of vertices of an appropriate graph. The set $V_s \subset V$, $V_s = \{k : a_{ik} \neq 0$ for some $i < k\}$ corresponds to the columns which are spikes. For any column $k \in V$ we recursively define the set of ancestor spikes as follows. A spike $m \in V_s$, $m > k$, is an *ancestor* of k either if there exists a row i such that $i \leqq k$ and $a_{im} \neq 0$ or if m is an ancestor of a spike l and l is an ancestor of k. If m is an ancestor of k, k is called *descendant* of m. The spike $l = \min\{i : i$ is an ancestor of $k\}$ is called the *father* of k. Conversely, k is a *son* of l. If two columns have the same father, they are called *brothers*. Each column $k \in V$ has no more than one father. A column which has no father is called a *root*. A matrix A may contain many root columns.

The father–son relationship defines an oriented graph (V, H), where V is the set of vertices and $H \subset V \times V$ is the set of arcs defined by $H = \{(l, k) : l$ is the father of $k\}$. The graph (V, H) is the *forest*, i.e., the set of disconnected subgraphs called rooted trees, where each rooted tree has a designated root vertex such that there is a unique path from the root to any other vertex in the tree. Let, for any spike vertex $l \in V_s$, the ordered set of the vertices $S(l) = (k_1, \cdots, k_m)$ denote the ordered list of the sons of the lth vertex, where the ordering is defined as follows: $k_i < k_j$ if and only if $i < j$. The ordering of the sons defines the ordering of the tree. To define the ordering in the forest (V, H) which is the set of trees, additionally assume that the root vertices of the trees are ordered in the same way. Figure 2 shows the ordered forest for the matrix A of Fig. 1. The roots are placed at the top of the figure. They are ordered from left to right. Then the vertices adjacent from the root vertices are placed one level below in the same order, etc., as in Fig. 2. The *level number* of a vertex k in the forest (V, H) is the length of the path from the ancestor root to k. The vertices of (V, H) with no sons are called *leaves*. Leaves correspond to these columns of the matrix A which have only zero elements above the main diagonal. Notice that a root vertex may be also a leaf. All other vertices are called *internal* and correspond to spikes.

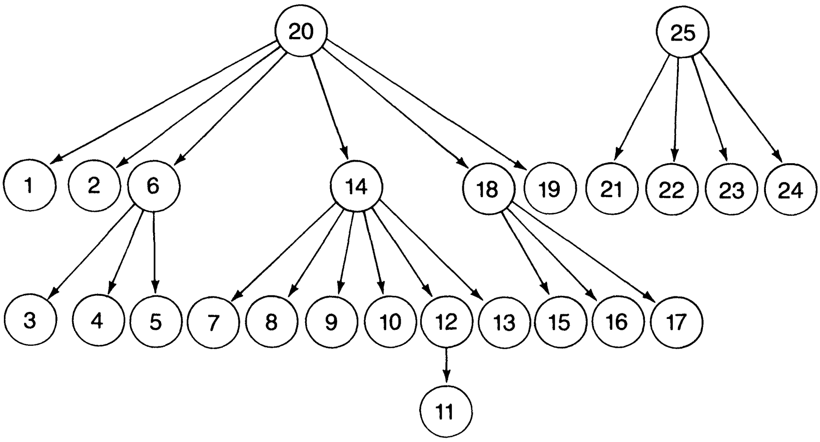


FIG. 2. *The forest of the matrix from Fig.* 1.

Let the function ldsc: $V_s \to V$ denote for each spike $k \in V_s$ its lowest descendant, i.e., $\text{ldsc}(k) = \min\{i: i \text{ is descendant of } k\}$. Each spike column $k \in V_s$ isolates in the matrix A a square diagonal block B_k, with elements a_{ij}, $i, j = \text{ldsc}(k), \cdots, k$. By the definition of ldsc(k), nonzero elements in the matrix A cannot occur in the columns ldsc$(k), \cdots, k$ above the submatrix B_k. The matrix B_k is called a *bump*. Any diagonal block which is positioned inside a larger diagonal block we call an *internal* block. Each nonspike column $k \in V \backslash V_s$ isolates a diagonal block $B_k = [a_{kk}]$ of order one. A diagonal block which is not contained in a larger diagonal block is called *external*. Figure 1 presents a matrix A with two external bumps B_{20} and B_{25}. The bump B_{20} is broken up into 6 internal blocks B_1, B_2, B_6, B_{14}, B_{18} and B_{19}. The bumps B_6, B_{14} and B_{18} are divided in a similar way.

The *height* of a tree T is the length of the longest path in the tree. The height of the forest (V, H) is the height of the highest tree in (V, H). In the next section it will become evident that the forest (V, H) should have as small a height as possible. There is a class of balanced forests for which the height is a logarithmic function of the number of spikes. We call a tree T of height h *balanced* if each spike vertex at levels $0, 1, \cdots, h-3$ has at least two internal sons. Any tree of height less than 3 is balanced. We call the forest (V, H) *balanced* if at least one of the highest trees of (V, H) is balanced.

THEOREM 2. *A balanced tree T with s internal vertices has height $h \leqq \lfloor \log_2 s \rfloor + 1$.*

Proof. If $h=1$, then $s=1$. If $h=2$, then $s \geqq 2$. Thus for $h \leqq 2$ the theorem is true. Assume that $h \geqq 3$. The root vertex must have at least two sons which are spikes. Generally, there are at least 2^l spikes at the level l, $l=0, 1, \cdots, h-2$. Thus there are at least $2^{h-1}-1$ spikes at levels $0, 1, \cdots, h-2$. Moreover, there is at least one spike at level $h-1$. Consequently $s \geqq 2^{h-1}$ and $h \leqq \log_2 s + 1$. □

COROLLARY. *If the forest (V, H) is balanced, its height $h \leqq \lfloor \log_2 s \rfloor + 1$, where s is the number of spikes.*

Let us define the function ldir: $V_s \rightarrow V$, $\text{ldir}(k) = \min_i \{i: a_{ik} \neq 0\}$. In general, $\text{ldsc}(k) \leqq \text{ldir}(k)$ for each $k \in V_s$. If m is an ancestor of k, we say that m and k are *properly positioned* if $\text{ldir}(m) \leqq \text{ldir}(k)$. In order to reduce the height of the forest (V, H) and to reduce fill-in, the spikes should be properly *nested*, i.e., $\text{ldir}(k) = \text{ldsc}(k)$ for each $k \in V_s$. Let us assume that for given permutation matrices P and Q the matrix $A = P \cdot \hat{A} \cdot Q$ is in the hierarchical BBLT form with some spikes initially not properly positioned. Then rearrangements of spike columns to provide proper nesting of spikes may be easily done by an appropriate sorting.

4. Properties of the hierarchical form of the inverse for sparse matrices. In this section we shall discuss the properties of the form of the inverse (13) in the case where the matrix A is sparse and is in the HBBLT form. Notice first, that in the course of the elimination only $s_1, r_2, s_2, \cdots, r_n, s_n$ are computed; the nonzero elements under the main diagonal remain unchanged, and thus new nonzero elements may be created only on and above the main diagonal in the spike columns only.

Observe that, if at the $(k+1)$st stage, the $(k+1)$st column is a nonspike column, i.e., $h_{k+1} = r_{k+1} = 0$; then, when calculating (9) for any $v^{(k+1)}$, the formula (12) disappears. Furthermore, if h_{k+1} is a spike column which has its first l coefficients equal to zero, $l \leqq k$, then the recursive algorithm for computation of $r_{k+1} = A_k^{-1} \cdot h_{k+1}$ by (9), (10), (11), (12) starts from the $(l+1)$st step, i.e., $w^{(l)} = A_l^{-1} \cdot v^{(l)} = 0$.

The property that fill-in does not occur in columns which are not spikes may be generalized as follows. If k is a spike column and B_k is the bump matrix associated with k, fill-in does not occur above the bump B_k. To prove it, notice that the kth leading principal submatrix A_k has the BLT form

$$A_k = \begin{bmatrix} A_l & \\ C & B_k \end{bmatrix} \tag{19}$$

where A_l is the leading principal submatrix of order l, $l = \text{ldsc}(k) - 1$, and C is an appropriate rectangular matrix. Thus the above property follows directly from (19) and (5). We also deduce the following.

THEOREM 3. *If the kth column of the matrix A is a spike which has β_k nonzero elements above and on the main diagonal, then the number of new nonzero elements created in the kth column does not exceed $\tau_k = n_k - \beta_k$, where $n_k = k + 1 - \text{ldsc}(k)$ is the order of the matrix B_k.*

Figure 3 shows the difference in the creation of nonzeros by the Gaussian and our methods. Shaded areas denote possible fill-in in spike columns. Let k be a spike column, $l_k = \text{ldir}(k)$, and β'_k be the number of nonzero elements in the kth column of A. Then the Gaussian elimination process may produce $n + 1 - l_k - \beta'_k$ new nonzero elements in the kth column. If s is the number of spikes, $s = |V_s|$, the bound for possible fill-in in Gaussian elimination equals $s(n+1) - \sum_{k \in V_s} (l_k + \beta'_k)$.

THEOREM 4. *If the forest (V, H) of the matrix A has height h, then the number of new nonzero elements created during the elimination process is less than or equal to*

$h \cdot n - \beta$, where β is the number of nonzero elements in spike columns of the matrix A on and above the main diagonal.

Proof. Consider the set of spikes $S_i = \{k : k$ has level number $i\}$, where $0 \leqq i < h$. The sum of orders of the matrices B_k, $k \in S_i$, is not greater than n, thus the fill-in in spike columns $k \in S_i$ is less than or equal to $n - \sum_{k \in S_i} \beta_k$. The spike vertices are only at the levels $i = 0, 1, \cdots, h-1$; therefore the overall bound is $h \cdot n - \beta$, where $\beta = \sum_{k \in V_s} \beta_k$. □

COROLLARY. *If the forest (V, H) is balanced, then the fill-in is less than or equal to $n(\log_2 s + 1) - \beta$.*

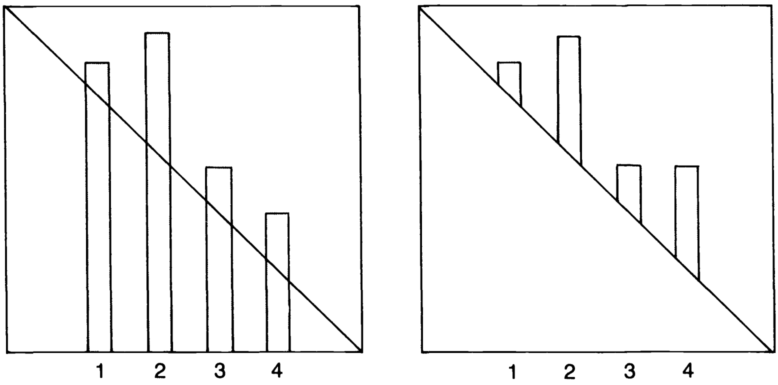


FIG. 3. *Fill-in during* (a) *the Gaussian elimination,* (b) *the hierarchical elimination.*

Notice in Fig. 3 that spikes 3 and 4 are improperly positioned. If we interchange them, fill-in will not occur in spike 4 above the row $i =$ ldir (4). There is a further way to reduce fill-in. Assume that we solve the system $Ax = b$ but the inverse of the matrix A is not required. At the kth stage of the elimination process only nonzero coefficients in (4) are necessary. Thus the storage space associated with nonzero elements of HFI (A_k) are released and can be used for storing the items associated with new nonzero elements created at the $(k+1)$st stage.

Notice that for any column k, the diagonal block B_k contains the full information necessary to compute r_k and s_k in (13). This follows immediately from (19) and (5). This property makes it possible to exploit the hierarchical structure of the matrix A in parallel processing and in alternating sequential/parallel processing, since the inverses of the disjoint diagonal blocks may be computed independently.

If a diagonal block B_k lies inside a block B_l, the representation of the inverse of B_k forms a part of the representation of the inverse of B_l. Let us consider an example. For the matrix in Fig. 1 the inverses of the 1×1 matrices $B_1, \cdots, B_5, B_7, \cdots, B_{11}$, $B_{13}, B_{15}, \cdots, B_{17}, B_{19}, B_{21}, \cdots, B_{24}$ may be computed independently at the first stage. In the next stages the inverse of B_{12} should be computed before B_{14}^{-1} since B_{12} is imbedded in B_{14}, and finally, the inverses of B_6, B_{14} and B_{18} should be computed before B_{20}^{-1}, since B_6, B_{14}, B_{18} are imbedded in B_{20}. The hierarchical form of the inverse of the matrix A contains the disjoint representations of the inverses of B_{20} and B_{25}; the representation of the inverse of B_{20} contains the inverses of B_1, B_2, B_6, B_{14}, B_{18} and B_{19}; the representation of the inverse of B_{14} contains the inverses of

B_7, B_8, B_9, B_{10}, B_{12} and B_{13}. Such a repeated partitioning is continued to the level where only 1×1 blocks remain.

Notice that the forest (V, H) of a matrix A is also the forest of the matrix T which is the hierarchical form of the inverse of A. For the instance matrix discussed in the example this may be verified in Fig. 2.

Now we shall analyse the number of computations for the solution of the system $Ax = b$.

THEOREM 5. *Let for an $n \times n$ nonsingular matrix A the* HFI (13) *have τ nonzero elements. Then the evaluation of the solution $x = A^{-1}b$ requires no more than $\tau - n$ additions and τ multiplications.*

Proof. The proof is almost identical to the proof of Theorem 1(i). The main difference consists in considering the actual fill-in of matrices and vectors involved instead of the maximal fill-in.

Assume that for the matrix A_{n+1} in (14) the matrix A_n has τ nonzero elements in the inverse representation (13), g_{n+1} has δ nonzero elements and $r_{n+1} = A_n^{-1}h_{n+1}$ has ρ nonzero elements. Thus A_{n+1} has $\tau' = \tau + \delta + \rho + 1$ nonzero elements in the representation of the inverse (13). The evaluation of (16) requires no more than δ additions and $\delta + 1$ multiplications. The evaluation of (17) requires no more than ρ additions and multiplications. Hence, together with the bound resulting from the inductive assumption on A_n, we have the bounds $\tau' - (n + 1)$ for additions and τ' for multiplications (if b is sparse, the number of computations may be lower). □

From the proof of the above theorem we also deduce that each nonzero element in (13) is involved in no more than one addition and multiplication. Now let us estimate the complexity bounds for computation of the hierarchical form of the inverse (13).

DEFINITION. A nonzero element t_{ij} in the HFI (13) is called *adjacent* to the mth level of the forest (V, H) if:

(i) it is located in a matrix B_k associated with the vertex k at level m, but it is not located in the kth column above the main diagonal;

(ii) m is the maximal level for which (i) holds.

THEOREM 6. *Let $m < h$. Any element t_{ij} adjacent to the mth level of the forest (V, H) is involved in no more than $(m + 1)$ additions and multiplications during the computation of the* HFI (13).

Proof. Let t_{ij} be adjacent to the mth level and located in the matrix B_k associated with the vertex k at the level m. Thus the vertex k has m ancestors. If l is an ancestor of k, t_{ij} is used in computing r_l. From Theorem 5 this requires no more than one addition and multiplication for each ancestor. Also during the computation of r_k and s_k, t_{ij} is involved in no more than one addition and one multiplication. Since t_{ij} is not involved in any other computations, this ends the proof. □

COROLLARY. *The computation of the* HFI (13) *requires no more than $h \cdot \tau$ additions and multiplications.*

The analogous bound in Gaussian elimination is higher, since the computation of some elements in the factorization may even require up to n additions and multiplications.

5. Further remarks. Now let us discuss some questions concerning whether the proposed method can be implemented efficiently, since eventual complexity in the data structure of matrix nonzeros may mitigate other advantages of the method. Moreover, interchange strategies to avoid numerical singularities and to mitigate the effects of rounding errors in the arithmetic operations should also be realized efficiently.

We advise implementing the hierarchical elimination algorithm by making use of row-linked lists where direct access to data via rows is provided [14], [7]. The memory requirements may not exceed τ list items and $4n+h$ integers. Each item contains the nonzero value of the element, the corresponding column index and the pointer to the next item in the same row. $2n$ integers are used for an array RS that points to the start of row-linked lists and for an auxiliary array RSAUX that points to the start of row-linked lists in the transformed subsystem (4). The remaining $2n$ integers are used for representing the permutation matrices P and Q. Additionally only h integers are used for a working stack array which contains the indices of current ancestor spike columns. In order to ensure the proper nesting of spikes, the function values ldsc (k) and ldir (k) should be stored, but they may be packed in the free space of RSAUX.

At the $(k+1)$st stage of the elimination process the coefficients $a_{ij}^{(k+1)}$, $i \leqq k+1$, $j=k+2, \cdots, n$ in (5) may be computed according to (9)–(12) by scanning appropriate row linked lists and performing the elimination steps for which the basic operation is to substract a multiple of one row from another.

The data structure discussed above allows for easy choice of a pivotal element by interchanging an ancestor spike column with the current column. The following pivoting strategy may be easily performed. At the $(k+1)$st stage let the Schur complements $s_{k+1,j} = a_{k+1,j} - g_{k+1} \cdot w_j^{(k)}$ be computed for each current spike column j, $j \geqq k+1$, where $w_j^{(k)} = (a_{1j}^{(k)}, \cdots, a_{kj}^{(k)})^T$. Then we may choose as the pivot column the lth spike column with the maximal value of ldir (l) among those columns which satisfy the inequality

$$|s_{k+1,l}| \geqq u \cdot \max_{k+1 \leqq j \leqq n} |s_{k+1,j}|,$$

where u $(0 < u \leqq 1)$ is an input parameter. The value of u settles a compromise between rounding errors and fill-in. If $u=1$ then the multiplier w_{k+1} in (12) is less or equal to one at each step of the elimination process. Notice that at the $(k+1)$st step there must exist a column l for which $s_{k+1,l}$ is nonzero, since otherwise the first $k+1$ rows are linearly dependent.

When considering rounding errors, it seems that for nonsparse matrices the hierarchical elimination method with partial pivoting strategy $(u=1)$ is comparable to the Gaussian elimination method with partial pivoting since, for the same pivot sequence, the elements $a_{k,j}^{(k)}$ computed at the kth stage of the Gaussian elimination are equal to the Schur complements $s_{k+1,j}$ computed in the hierarchical elimination method. It is very difficult to analyse the propagation of numerical errors in the case of sparse matrices. We expect that, for a matrix A in hierarchical BBLT form with a reasonable depth of nesting, the hierarchical elimination method may also be advantageous with respect to numerical errors, since then relatively few nonzero elements are involved in a relatively few computations. Observe also that, while computations in the Gaussian elimination are distributed, in the hierarchical method we basically compute scalar products for which higher numerical accuracy may be obtained.

Other hierarchical approaches for inverting large sparse matrices have been proposed before. George [4] developed a nested dissection method for sparse symmetric matrices. McBride [11] developed a nested method which permits solving nonsymmetric sparse systems by repeated partitioning of the matrix and by computing a series of miniinverses (minikernels). In his approach, however, hierarchy involves exponential growth of the computations in function of the depth of nesting. The nested algorithms combined with sparse Gaussian elimination require increased program overhead and storage management that mitigate the possible advantage of low operation or fill-in count over conventional methods. In the hierarchical elimination algorithm the hierarchical structure of matrices is utilized in a natural way with no additional program and storage overhead.

6. Illustrative examples.

Example 1. Let us consider the system

$$\begin{bmatrix} 2 & 0 & 1 & 0 & 0 & 0 & 1 \\ 4 & 2 & 0 & 0 & 0 & 0 & 2 \\ 1 & -2 & 2 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 2 & 0 & 1 & 2 \\ 1 & 0 & -2 & 1 & 1 & 1 & 0 \\ 0 & 2 & -1 & 1 & 1 & 0 & 2 \\ -1 & 0 & 2 & 0 & -1 & 0 & 1 \end{bmatrix} \cdot x = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \\ -1 \\ 2 \\ 2 \end{bmatrix}, \tag{20}$$

where $x=(x_1,\cdots,x_7)^T$ is the unknown vector. We shall describe the elimination process for the system (20). At the first stage we consider the first equation of (20). We premultiply it by the inverse of the element $a_{11}=2$, which is equal to the first leading principal submatrix A_1 of A. For $k=1$ we have the reduced equation

$$[1 \quad 0 \quad 0.5 \quad 0 \quad 0 \quad 0 \quad 0.5]\cdot x = 1. \tag{21}$$

The HFI of A_1 equals 0.5. At the second stage we add the second equation of (20) to (21). Notice that x_2 is a nonspike variable. Thus at the second stage, the first equation (21) remains unchanged. To recompute the coefficients in the second equation, we use formula (11) only for columns 3, 7 and the free column. Columns 4, 5 and 6 remain equal to zero. Hence, for $k=2$, the reduced system is

$$\begin{bmatrix} 1 & 0 & 0.5 & 0 & 0 & 0 & 0.5 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{22}$$

and

$$\mathrm{HFI}\,(A_2) = \begin{bmatrix} 0.5 & 0 \\ 4 & 0.5 \end{bmatrix}. \tag{23}$$

At the third stage we add the third equation of (20) to (22). Now x_3 is a spike variable. Using formula (11) we recompute the coefficients in the third row only in the free column and in column 7. Coefficients in columns 4, 5 and 6 remain equal to zero. Then we recompute the coefficients in the first and second row in the free column and in column 7 by (12). Hence, for $k=3$, the reduced system is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot x = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \qquad \mathrm{HFI}\,(A_3) = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 4 & 0.5 & -1 \\ 1 & -2 & -2 \end{bmatrix}.$$

Continuing this process we obtain the reduced systems and the hierarchical form of the inverses as follows:

$$k=4 \qquad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0.5 & 1.5 \end{bmatrix} \cdot x = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1.5 \end{bmatrix},$$

$$\mathrm{HFI}\,(A_4)=\begin{bmatrix}0.5 & 0 & 0.5 & 0\\ 4 & 0.5 & -1 & 0\\ 1 & -2 & -2 & 0\\ 1 & -1 & 0 & 0.5\end{bmatrix},$$

$$k=5\qquad \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1 & 0 & 0.5 & 1.5\\ 0 & 0 & 0 & 0 & 1 & 0.5 & 0.5\end{bmatrix}\cdot x=\begin{bmatrix}0\\ 1\\ 2\\ 1.5\\ 1.5\end{bmatrix},$$

$$\mathrm{HFI}\,(A_5)=\begin{bmatrix}0.5 & 0 & 0.5 & 0 & 0\\ 4 & 0.5 & -1 & 0 & 0\\ 1 & -2 & -2 & 0 & 0\\ 1 & -1 & 0 & 0.5 & 0\\ 1 & 0 & -2 & 1 & 1\end{bmatrix},$$

$$k=6\qquad \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 1\end{bmatrix}\cdot x=\begin{bmatrix}0\\ 1\\ 2\\ 1\\ 1\\ 1\end{bmatrix},$$

$$\mathrm{HFI}\,(A_6)=\begin{bmatrix}0.5 & 0 & 0.5 & 0 & 0 & 0\\ 4 & 0.5 & -1 & 0 & 0 & 0\\ 1 & -2 & -2 & 0 & 0 & 0\\ 1 & -1 & 0 & 0.5 & 0 & 0.5\\ 1 & 0 & -2 & 1 & 1 & 0.5\\ 0 & 2 & -1 & 1 & 1 & -1\end{bmatrix},$$

and finally, for $k=7$, we obtain the solution $x=(0,0,1,0,1,0,1)^T$ and the hierarchical form of the inverse of the matrix A,

$$\mathrm{HFI}\,(A)=\begin{bmatrix}0.5 & 0 & 0.5 & 0 & 0 & 0 & 0\\ 4 & 0.5 & -1 & 0 & 0 & 0 & 1\\ 1 & -2 & -2 & 0 & 0 & 0 & 1\\ 1 & -1 & 0 & 0.5 & 0 & 0.5 & 1\\ 1 & 0 & -2 & 1 & 1 & 0.5 & 0\\ 0 & 2 & -1 & 1 & 1 & -1 & 1\\ -1 & 0 & 2 & 0 & -1 & 0 & -1\end{bmatrix}.$$

Figure 4 shows the tree of the hierarchical BBLT form of the matrix A which is also the tree of the hierarchical BBLT form of HFI (A).

Example 2. Let us consider a matrix A of order 25 which has the structure presented in Fig. 1. Comparison of the fill-in and the number of computations in the Gaussian and our elimination methods for the system $Ax = b$ is given in Table 1.

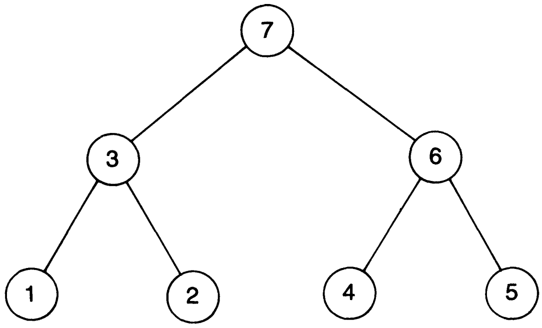


FIG. 4

TABLE 1

		factorization of A		solution of the system $Ax = b$	
method	fill-in	additions	multiplications	additions	multiplications
Gaussian elimination method	43	165	271	307	438
our elimination method	21	105	134	225	279

7. Conclusions. The hierarchical form of the inverse of a sparse matrix allows us to exploit the hierarchical BBLT structure of the matrix in a very efficient way. The hierarchical BBLT structure may be identified by the commonly used heuristic preassigned pivotal procedures. It should be stressed, however, that these procedures tend to minimize the number of spikes rather than the depth of nesting, and the latter appears to be a better criterion. A new preassigned pivotal algorithm, which is tailored for the hierarchical elimination method was developed in [13].

A full evaluation of the hierarchical elimination method requires extensive performance comparisons with other available methods for sparse matrices of greater order. We have begun a program of comparisons and plan to publish the results in a separate paper.

Acknowledgments. The author is grateful to A. Ruszczyński and K. Kiwiel for reading the manuscript and for valuable discussions. The author would also like to thank the referee for valuable comments and suggestions which helped to refine the paper.

REFERENCES

[1] R. W. COTTLE, *Manifestation of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189–211.

[2] I. S. DUFF, *A survey of sparse matrix research*. Proc. IEEE, 65 (1977), pp. 500–535.

[3] V. N. FADDEEVA, *Computational Methods of Linear Algebra*, C. D. Benster, trans., Dover, New York, 1959. Translated from the 1950 Russian ed: *Vitchislitielnye Metody Linieynoy Algebry.*

[4] A. GEORGE, *Nested dissection for a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[5] E. HELLERMAN AND D. C. RARICK, *The partitioned preassigned pivot procedure* (P4), in Sparse Matrices and Their Applications, D. J. Rose and R. A. Willoughby, eds., Plenum Press, New York, 1972, pp. 67–76.

[6] A. K. KEVORKIAN AND J. SNOEK, *Decomposition in large scale systems: theory and applications in solving large sets of non-linear simultaneous equations*, in Decomposition of Large-Scale Problems, D. Himmelblau, ed., North-Holland, Amsterdam, 1973, pp. 491–515.

[7] D. E. KNUTH, *The Art of Computer Programming, Vol.* 1, *Fundamental Algorithms*, Addison-Wesley, Reading, MA, 1969.

[8] W. P. LEDET AND D. M. HIMMELBLAU, *Decomposition procedures for the solving of large scale systems*, Adv. Chem. Eng., (1970), pp. 185–253.

[9] T. D. LIN AND R. S. H. MAH, *Hierarchical partition—a new optimal pivoting algorithm*, Mathematical Programming, 12 (1977), pp. 260–278.

[10] H. M. MARKOWITZ, *The elimination form of the inverse and its application to linear programming*, Management Sci., 3 (1957), pp. 255–269.

[11] R. D. MCBRIDE, *A bump triangular dynamics factorization algorithm for the simplex method*, Math. Programming, 18 (1980), pp. 41–61.

[12] D. V. STEWARD, *On the approach to techniques for the analysis of the structure of large systems of equations*, SIAM Rev., 4 (1962), pp. 321–342.

[13] E. TOCZYŁOWSKI, *An algorithm for finding nested bordered block lower triangular form of sparse matrices*, in preparation.

[14] R. P. TEWARSON, *Sparse Matrices*, Academic Press, New York and London, 1973.

[15] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0008 $01.25/0

SPECTRA OF SOME GRAPHS*

KAI WANG†

Abstract. In this paper a method for computing the characteristic polynomial of a graph possessing an involutary automorphism from the characteristic polynomials of two smaller associated graphs is presented.

1. Introduction. A graph G consists of 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. If G is a graph with $V(G)=\{v_1, \cdots, v_n\}$, then the adjacency matrix of G is an $n\times n$ matrix $A(G)=[w(v_i, v_j)]$ where $w(v_i, v_j)=1$ if $(v_i, v_j)\in E(G)$ and $=0$ otherwise. The characteristic polynomial of $A(G)$ is called the characteristic polynomial of G and is denoted by $\varphi(G; x)$. We refer to [1] for basic knowledge and related topics.

A permutation τ on $V(G)$ is an automorphism of G if τ preserves adjacency and τ is involutary if $\tau^2=\text{identity}$. In this paper, we will show how to reduce the adjacency matrix of a graph possessing an involutary automorphism to the direct sum of two submatrices which can be easily read off from the given graph. We refer to later sections for the main result of this paper and its applications.

We are very grateful to a referee for the suggestions on an early version of this paper.

2. Graphs with involutions. In this section, we will study the characteristic polynomials of graphs possessing involutary automorphisms. For our purpose, we will consider pseudographs. A *weighted pseudograph* K consists of a finite set V and a map $w: V\times V\to\mathbb{C}$, denoted by $K=(V, w)$. V is called the vertex set and is denoted by $V(K)$ if necessary. The value $w(u, v)$ is called the weight of the edge (u, v). For convenience we assume that $w(u, v)=w(v, u)$ for all $u, v\in V$. For a labelled weighted pseudograph K, let $A(K)=[w(u_i, u_j)]$ be its adjacency matrix and let $\varphi(K, x)=\det(xI-A(K))$ be the characteristic polynomial of K. An involution τ on K is a permutation on V so that $\tau^2=\text{identity}$ and $w(u, v)=w(\tau(u), \tau(v))$ for all $u, v\in V$. Let $F=\{v\in V\,|\,\tau(v)=v\}$, the set of fixed vertices of τ. For $v\in V$, let $v^*=\{v, \tau(v)\}$ if $v\notin F$ or $v^*=\{v\}$ if $v\in F$ be the orbit of v. Let $V^*=\{v^*\,|\,v\in V\}$. Let $f=|F|$ and let $|V^*|=f+t$. Then $|V|=f+2t$. Let $V=\{v_1, \cdots, v_{f+2t}\}$ be so labelled that $F=\{v_1, \cdots, v_f\}$, $V^*=\{v_1^*, \cdots, v_{f+t}^*\}$ and $v_{f+t+i}=\tau(v_{f+i})$ for $1\leqq i\leqq t$. Let

$$\begin{aligned}
a_{ij}&=w(v_i, v_j) && \text{for } 1\leqq i, j\leqq f,\\
b_{ij}&=w(v_{f+i}, v_{f+j}) && \text{for } 1\leqq i, j\leqq t,\\
c_{ij}&=w(v_{f+t+i}, v_{f+j}) && \text{for } 1\leqq i, j\leqq t,\\
d_{ij}&=w(v_i, v_{f+j}) && \text{for } 1\leqq i\leqq f,\quad 1\leqq j\leqq t.
\end{aligned}$$

Then the corresponding adjacency matrix for K is given by

$$A(K)=\begin{bmatrix} Z & T & T\\ T^t & X & Y\\ T^t & Y & X\end{bmatrix}$$

* Received by the editors November 16, 1981, and in revised form February 22, 1983.

† Department of Mathematics, Wayne State University, Detroit, Michigan 48202.

where $Z=[a_{ij}]$, $X=[b_{ij}]$, $Y=[c_{ij}]$ and $T=[d_{ij}]$. Note that if $F=\varnothing$, then

$$A(K)=\begin{bmatrix} X & Y \\ Y & X \end{bmatrix}.$$

Let $M=I_f \oplus (\Omega_2 \otimes I_t)$ where Ω_2 is the Fourier matrix of order 2 [2] given by

$$\Omega_2=\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Then M is unitary and $M^*=M$. An easy computation shows that

$$M^{-1}A(K)M=MA(K)M$$

$$=\begin{bmatrix} I_f & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}}I & \frac{1}{\sqrt{2}}I_t \\ 0 & \frac{1}{\sqrt{2}}I_t & \frac{-1}{\sqrt{2}}I_t \end{bmatrix}\begin{bmatrix} Z & T & T \\ T^t & X & Y \\ T^t & Y & X \end{bmatrix}\begin{bmatrix} I_f & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}}I_t & \frac{1}{\sqrt{2}}I_t \\ 0 & \frac{1}{\sqrt{2}}I_t & \frac{-1}{\sqrt{2}}I_t \end{bmatrix}$$

$$=\begin{bmatrix} Z & \sqrt{2}\,T & 0 \\ \sqrt{2}\,T^t & X+Y & 0 \\ 0 & 0 & X-Y \end{bmatrix}.$$

Note that

$$\begin{bmatrix} Z & \sqrt{2}\,T \\ \sqrt{2}\,T^t & X+Y \end{bmatrix}$$

is the adjacency matrix for the orbit graph K^* with vertex set V^* and $w^*(u^*,v^*)$ given by

$$w^*(u^*,v^*)=\begin{cases} w(u,v) & \text{if } u,v\in F, \\ w(u,v)+w(u,\tau(v)) & \text{if } u,v\notin F, \\ \sqrt{2}\,w(u,v) & \text{if } u\in F, v\notin F \text{ or } u\notin F, v\in F, \end{cases}$$

and $X-Y$ is the adjacency matrix for a *cograph* to F in K^*, denoted by F^c, which is a weighted pseudograph with vertex set V^*-F and weight function w' defined by $w'(u^*,v^*)=w(u,v)-w(u,\tau(v))$. Notice that although F^c depends on the choice of orbit representatives, the characteristic polynomial of $X-Y$ does not. With the above notation, we can state our main result in the following theorem.

THEOREM 2.1. *Let K be a weighted pseudograph possessing an involuntary automorphism. Let K^* be its orbit graph and let F^c be the cograph of the fixed point set in K^*. Then*

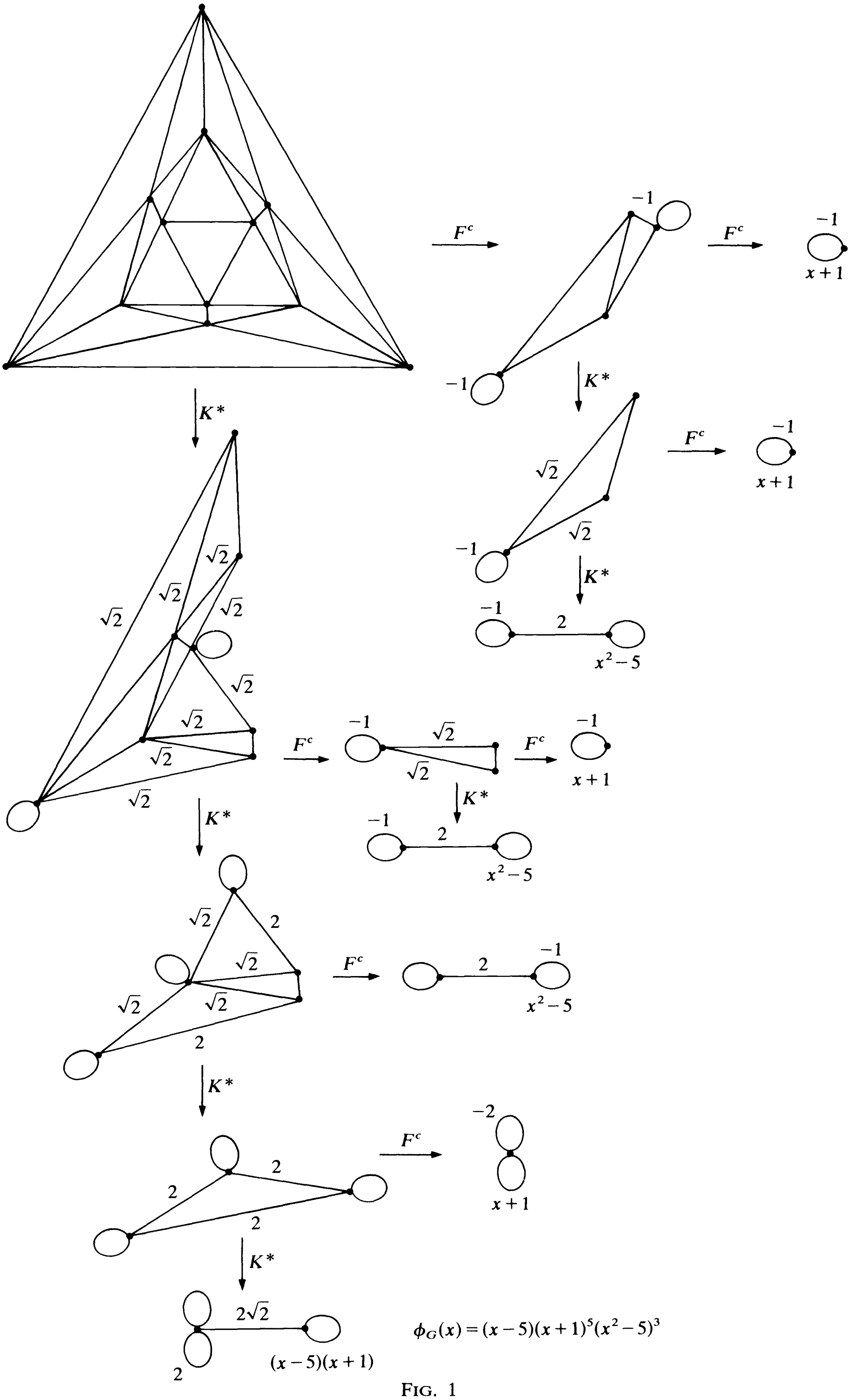
$$\varphi(K;x)=\varphi(K^*;x)\varphi(F^c;x).$$

3. An example. In practice, we may construct K^* as follows.

1. Start with a convenient induced subgraph of K such that its vertex set is a set of representatives for the orbits.

2. Multiply the weight of each edge which joins a fixed vertex to a nonfixed vertex by $\sqrt{2}$.

3. Construct a new edge or loop for each unordered pair (u,v) of nonfixed vertices with weight equal to $w(u,\tau(v))$.



$$\phi_G(x) = (x-5)(x+1)^5(x^2-5)^3$$

FIG. 1

The resulting weighted pseudograph is K^*. For F^c we proceed as follows.

1. Take the complement of F in K^*.
2. Multiply the weight of each *new* edge or loop in K^* by -1.

The result is F^c. Usually, we simplify the edges or loops with same end vertices. We remark that another method to compute $p(G; x)$ is given in [1]. Their method is too long to be discussed here. We plan to study the relation between these two methods in a future paper. In the examples, shown in Fig. 1 we use diagrams to show how to find the characteristic polynomial of a given graph with our method. For convenience, edges without specific weight have unit weight and edges not shown have weight zero.

REFERENCES

[1] D. M. CVETKOVIC, M. DOOB AND H. SACHS, *Spectra of Graphs, Theory and Applications*, Academic Press, New York, 1980.

[2] P. J. DAVIS, *Circulant Matrices*, John Wiley, New York, 1979.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0009 $01.25/0

ON THE SIZE OF SEPARATING SYSTEMS AND FAMILIES OF PERFECT HASH FUNCTIONS*

MICHAEL L. FREDMAN† AND JÁNOS KOMLÓS†

Abstract. This paper presents two applications of an interesting information theoretic theorem about graphs. The first application concerns the derivation of good bounds for the function $Y(b, k, n)$, which is defined to be the minimum size of a family of functions such that for every subset of size k from an n element universe, there exists a perfect hash function in the family mapping the subset into a table of size b. The second application concerns the derivation of good bounds for the function $M(i, j, n)$, which is defined to be the minimum size of an (i, j)-separating system.

1. Introduction. Two problems involving the minimum size of families of partitions satisfying certain separating conditions are considered in this paper. An interesting information theoretic technique is developed and applied to these problems. Our first problem is motivated by hashing. Let $U=\{1, 2, \cdots, n\}$ and let P be a partition of U into b (possibly empty) blocks. A subset S of U with at most b elements is *separated* by P if every block of P contains at most one element of S. A family F consisting of partitions of U into b blocks is called a (b, k)-system provided that every subset S with k elements is separated by at least one partition in F. Our first problem is to derive bounds for the function $Y(b, k, n)$ which is defined to be the minimum size of any (b, k)-system. A function h from U into $\{1, \cdots, b\}$ is said to be a perfect hash function with respect to a subset S provided that h is one-to-one on S. A collection C of functions from U into $\{1, \cdots, b\}$ is called a (b, k)-family of perfect hash functions provided that for each subset S of size k, there is a function h in C which is perfect with respect to S. A (b, k)-family of perfect hash functions provides a means for storing subsets of size k into tables with b cells. Given the obvious correspondence between partitions into b blocks and functions with range $\{1, \cdots, b\}$, it is clear that $Y(b, k, n)$ gives the minimum size of any (b, k)-family of perfect hash functions.

Our second problem concerns the minimum size of (i, j)-separating systems. As defined in Friedman, Graham, and Ullman [2], an (i, j)-separating system is a family F of partitions of U into two blocks, (P_k, Q_k), $(P_k \cup Q_k = U)$ satisfying the following condition. For each pair of disjoint subsets, S and T, of size i and j, respectively, there is at least one partition (P_k, Q_k) in F such that $S \subseteq P_k$ and $T \subseteq Q_k$, or $T \subseteq P_k$ and $S \subseteq Q_k$. Friedman et al. [2] show that (i, j)-separating systems are motivated by considerations involving asynchronous sequential circuits. We define $M(i, j, n)$ to be the size of a mimimal (i, j)-separating system.

In § 2 we describe some techniques which produce initial estimates for the functions $Y(b, k, n)$ and $M(i, j, n)$. A discussion of the relationships between these techniques and initial estimates motivates an information theoretic result which is established in § 3. This information theoretic result is applied in the final sections of the paper to improve upon the initial estimates.

Throughout this paper all logarithms are assumed to be to the base 2.

2. Initial estimates. Two related techniques which lead to lower and upper bounds for these problems are, respectively, the volume bound and the random bound. We

* Received by the editors December 7, 1982. This research was supported in part by the National Science Foundation under grant MCS-8204031.

† Department of Electrical Engineering and Computer Sciences, University of California, San Diego, La Jolla, California 92093.

illustrate these techniques in the context of the (b, k)-systems problem, and remark that they similarly apply to the (i, j)-separating systems problem.

The volume bound considers the maximum possible number v of sets which can be separated by any fixed partition. Since $N = \binom{n}{k}$ sets must be separated in all, we deduce that

$$H(b, k, n) \geqq \frac{N}{v}. \tag{1}$$

We refer to the number of sets separated by a given partition as the volume of the partition. Let G be the family of all partitions having maximal volume. The random bound considers the necessary length of a sequence of random independently chosen partitions from G in order for the sequence to comprise a (b, k)-system with positive probability. The probability that such a sequence of length m *fails* to comprise a separating system is at most

$$\left(1-\frac{v}{N}\right)^m N. \tag{2}$$

If m is chosen so that the expression in (2) is less than 1, then $Y(b, k, n) \leqq m$. Hence we obtain the bound

$$Y(b, k, n) = O\left(\frac{\log N}{-\log(1-v/N)}\right). \tag{3}$$

It is clear that a partition of maximal volume will have block sizes which are as uniform as possible. In particular, if $n > b^{2+\varepsilon}$, $\varepsilon > 0$, then

$$v \sim \left(\frac{n}{b}\right)^k \binom{b}{k}. \tag{4}$$

Combining (1), (3), and (4), we obtain (for $n > b^{2+\varepsilon}$)

$$\Omega\left(\frac{b^k}{b^{\underline{k}}}\right) = Y(b, k, n) = O\left(-k \log n/\log\left(1-\frac{b^{\underline{k}}}{b^k}\right)\right) \quad \text{where } b^{\underline{k}} = b!/(b-k)! \tag{5}$$

A particularly striking discrepancy between the lower and upper bounds in (5) is the fact that the upper bound depends on n while the lower bound does not. Our primary interest throughout the remainder of this paper is an analysis of this discrepancy. Let us assume that $k = \lfloor \alpha b \rfloor$ for fixed α, $0 < \alpha \leqq 1$ (corresponding to hash tables with load factor α). Then $b^k/b^{\underline{k}}$ grows exponentially in b, and taking logarithms, our bound in (5) becomes

$$\begin{aligned} bg(\alpha) + O(\log b) &\leqq \log Y(b, k, n) \\ &\leqq bg(\alpha) + \log\log n + O(\log b) \\ (g(\alpha) = (1-\alpha)\log(1-\alpha) + \alpha \log e). \end{aligned} \tag{6}$$

A lower bound for $Y(b, k, n)$ which grows as $\Omega(\log n)$ can be derived by appealing to a simple information theoretic argument. We assign a numbering to the partitions of a (b, k)-system and we number the blocks within partitions, so that we may speak of the ith block of the jth partition. Then a (b, k)-system with m partitions induces the following assignment of m-dimensional b-ary vectors to the elements of $U = \{1, 2, \cdots, n\}$: The value of the jth component of the vector v_t assigned to $t \in U$

is i provided that t is in the ith block of the jth partition. Observe that $v_t \neq v_{t'}$ for $t \neq t'$ since t is separated from t' by at least one partition in the (b, k)-system. Therefore, $\{v_t | t \in U\}$ is a set of n distinct b-ary vectors, implying that

$$Y(b, k, n) \geqq \frac{\log n}{\log b}. \tag{7}$$

Combining (6) and (7), we obtain

$$\begin{aligned} &\max\,(g(\alpha)b, \log\log n) - O(\log b) \\ &\qquad \leqq \log Y(b, k, n) \leqq g(\alpha)b + \log\log n + O(\log b). \end{aligned} \tag{8}$$

For large b the lower and upper bounds in (8) differ by a factor ranging between 1 and 2, depending on n.

The reader should observe that our interest actually centers on the function $Y(b, k, n)$ rather than $\log Y(b, k, n)$. The discrepancy between the lower and upper bounds in (8) translates into an extremely large gap: a sum versus a product.

To improve our lower bound, we seek a method which combines the volume bound argument and the simple information theoretic argument. The volume bound involves the counting of subsets and the information theoretic argument assigns distinct vectors to the elements of U. A combined argument assigns vectors to subsets of U. In effect, we show that a typical subset must be separated about $\log n$ times in a (b, k)-system. Before presenting this argument, we need to establish an information theoretic result which is interesting in its own right.

Yao [5] mentions that R. Graham observed that the randomization technique yields a good estimate for the size of (b, k)-systems. Melhorn [4] has derived all of the estimates presented in this section. Berman et al. [1] have also derived some of these bounds. Friedman et al. [2] have used the random bound method to derive an upper bound for the size of (i, j)-separating systems.

3. An information theoretic inequality. Let V be a set of the form $W \cup \{*\}$ where W is a finite set of integers, and let V^d denote the set of d-dimensional vectors over V. Given two vectors u_1 and u_2 in V^d, we say that u_1 and u_2 *strongly* differ provided that in some position they assume different integer values. (We interpret $*$ as meaning "don't care", so that two vectors, which differ only in positions in which one of the vectors has a $*$, do not strongly differ.)

The concept of strongly differing vectors arises naturally. For example, if we pad the words of a binary prefix code with $*$'s so that the resulting words are of the same length, then these padded words pairwise strongly differ.

Let G be a finite undirected graph. A coloring of the vertices of G with colors chosen from the elements of V^d is said to be *strong* provided that the colors of any two vertices connected by an edge strongly differ.

Let w be a strong coloring of a graph G with colors from V^d. For each i, $1 \leqq i \leqq d$, and each $e \in V$ we let p_{ie} denote the fraction of vertices whose colors contain an e in the ith coordinate. We let $p(w, i)$ denote the probability that the ith coordinate is an integer: $p(w, i) = \sum_{e \neq *} p_{ie}$. If e is an integer, we let q_{ie} denote the conditional probability that the ith coordinate is e given that it is an integer; namely, $q_{ie} = p_{ie}/p(w, i)$. The conditional entropy of the ith coordinate, denoted $H(w, i)$, is defined by

$$H(w, i) = -\sum_{e \neq *} q_{ie} \log q_{ie}.$$

A set of vertices in G is *independent* provided that no two vertices in the set are joined by an edge. The *independence number* $\alpha(G)$ of G is the maximum size of any independent set.

THEOREM 1. *Let G be a graph on n vertices $\{v_1, \cdots, v_n\}$ with independence number α, and let w be a strong coloring of G with colors from V^d. Let $p(w, i)$ and $H(w, i)$ be the quantities defined above. Then*

$$\log\left(\frac{n}{\alpha}\right) \leqq \sum_{i=1}^{d} p(w, i)H(w, i). \tag{9}$$

Remark. If G is a clique and our coloring contains no $*$'s, then (1) reduces to the classical inequality $\log n \leqq \sum_i H(X_i)$, where X_i is the ith coordinate of a random vector which has a uniform distribution on n values. The proof of Theorem 1 refines the method in Fredman [3, Thm. 3], which can be regarded as an application of Theorem 1 in a setting involving $*$'s and a clique.

Proof of Theorem 1. For a large integer k, we let C_k denote the collection of all sequences of vertices in G of length nk, which contain exactly k occurrences of each of the n vertices. We define the following graph G_k whose vertices are the sequences in C_k. Two sequences, $\langle v_{i_r}\rangle$ and $\langle v_{j_r}\rangle$ are joined by an edge provided that at the first position in which the two sequences differ, say position s, the two vertices v_{i_s} and v_{j_s} are joined by an edge in G. The number of vertices in this graph, $|C_k|$, is given by the multinomial coefficient $\binom{nk}{k,\cdots,k}$, which is approximated by (n fixed, k large)

$$2^{kn}\log n + O(\log k). \tag{10}$$

We show below (Lemma 1) that the independence number $\alpha(G_k)$ of G_k cannot exceed

$$\alpha(G)^{kn}. \tag{11}$$

For any graph G' on h vertices, the chromatic number $X(G')$ satisfies $X(G') \geqq h/\alpha(G')$, i.e.

$$X(G_k) \geqq |C_k|/\alpha(G_k). \tag{12}$$

We deduce from (10), (11), and (12) that

$$\liminf_{k\to\infty} \frac{1}{nk}\log X(G_k) \geqq \log\left(\frac{n}{\alpha(G)}\right). \tag{13}$$

Next, we proceed to show that

$$\sum_{i=1}^{d} p(w, i)H(w, i) \geqq \liminf_{k\to\infty} \frac{1}{nk}\log X(G_k). \tag{14}$$

The theorem then follows by combining (13) and (14). To establish (14), we construct a coloring w_k of G_k using at most

$$\theta_k = \prod_{i=1}^{d} \binom{p(w, i)nk}{p_{i1}nk, p_{i2}nk, \cdots} \tag{15}$$

colors. Since

$$\lim_{k\to\infty} \frac{1}{nk}\log\theta_k = \sum_{i=1}^{d} p(w, i)H(w, i),$$

this coloring suffices to establish (14). Our coloring w_k will be defined so as to color each sequence in C_k with a d-dimensional vector whose entries consist of integer sequences. Given a vertex v in G, let $w(v, j)$ denote the jth component of the color of v under the strong coloring w. Then the jth component of the color of $\langle v_{i_r}\rangle \in C_k$ under the coloring w_k is the compressed sequence obtained by deleting all $*$'s from the sequence $\langle w(v_{i_r}, j)\rangle$. Observe that for each integer m, this sequence contains $p_{jm}nk$ occurrences of m. It follows that w_k utilizes at most θ_k (defined in (15)) possible colors. Lemma 2(below) establishes that w_k is indeed a proper coloring of G_k, completing the proof.

LEMMA 1. *The independence number $\alpha(G_k)$ of G_k satisfies $\alpha(G_k) \leqq \alpha(G)^{nk}$.*

Proof. Consider an independent set I of vertices in G_k. Recall that each vertex in G_k is a sequence of vertices in G of length nk. Suppose that there are t vertices in I which have a common prefix of length (say) $r-1$, and which pairwise differ in the rth position. Then the t vertices which comprise the rth positions of these t sequences form (by definition of G_k) an independent set of G. Thus $t \leqq \alpha(G)$. We can now argue by induction on $m \geqq 0$, that there can be at most $\alpha(G)^m$ sequences in I which have a common prefix of length $nk-m$. Setting $m = nk$, we conclude that $|I| \leqq \alpha(G)^{nk}$, completing the proof.

LEMMA 2. *The coloring w_k of G_k is proper.*

Proof. Let $\langle v_{i_r}\rangle$ and $\langle v_{j_r}\rangle$ be two sequences in C_k joined by an edge. Assume that these sequences have a common prefix of length $s-1$ and differ in position s, so that v_{i_s} and v_{j_s} are joined by an edge in G. Then for some l, $* \neq w(v_{i_s}, l) \neq w(v_{j_s}, l) \neq *$ since w is a strong coloring. Since $v_{i_r} = v_{j_r}$ for $r < s$, the prefixes, $\langle w(v_{i_r}, l); 1 \leqq r < s\rangle$ and $\langle w(v_{j_r}, l); 1 \leqq r < s\rangle$, are identical; and in particular both contain an equal number (say u) of occurrences of $*$. Thus, the colors of $\langle v_{i_r}\rangle$ and $\langle v_{j_r}\rangle$ differ in the lth component at position $s-u$. This implies that w_k is proper, completing the proof.

4. Lower bound for $Y(b, k, n)$. We now proceed to combine the information theoretic and volume bounds as discussed at the end of § 2. We define the following graph $G = G(b, k, n)$. The vertices of G consist of all pairs of the form (x, R), where R is a subset of U of size $k-2$ and x is an element of U which is *not* contained in R. Two vertices, (x, R) and (x', R'), are joined by an edge provided that $R = R'$. The number N of vertices in G is $n\binom{n-1}{k-2}$. Observe that G consists of disjoint cliques of size $n-k+2$, one clique for each subset R of U of size $k-2$. The independence number $\alpha = \alpha(G)$ is simply the number of these cliques, and so $N/\alpha = n-k+2$.

Let C be a (b, k)-system. We use C to induce a strong coloring of G as follows. The vertices of G are colored with vectors over $\{*, 1, 2, \cdots, b\}$. These vectors are composed of blocks of coordinates, each block containing $\binom{b}{k-2}$ coordinates. The jth partition P_j in C defines the values of the coordinates in the jth block of these vectors. Number the $\binom{b}{k-2}$ subsets of $\{1, \cdots, b\}$ of size $k-2$, and let Z_i be the ith subset. Then the value v_{ij} of the ith coordinate of the jth block of the vector (color) associated with the vertex (x, R) is given by

$$v_{ij} = \begin{cases} l & \text{if } x \text{ is in the } l\text{th block of } P_j,\ l \notin Z_i, \text{ and the } u\text{th block} \\ & \text{of } P_j \text{ contains exactly one element of } R \text{ for each } u \text{ in } Z_i, \\ * & \text{otherwise.} \end{cases}$$

Since v_{ij} cannot assume a value in Z_i, it must assume one of the remaining $b-k+2$ integer values or the value $*$.

To see that this is a strong coloring of G, consider two joined vertices (x, R) and (y, R). Let P_j be a partition which separates the set $\tilde{R} = \{x, y\} \cup R$ (which contains k elements), and let Z_i be the set of size $k-2$ consisting of the $k-2$ indices of the blocks of P_j containing the $k-2$ elements of R. Assume block l of P_j contains x and block l' of P_j contains y. Since $\tilde{R}$ is separated by P_j, we have that $l \neq l', l \notin Z_i$ and $l' \notin Z_i$. Thus, for the vertex (x, R), $v_{ij} = l$; and for (y, R), $v_{ij} = l'$. In other words, the colors of these two vertices strongly differ.

THEOREM 2. *For* $n > b^{2+\varepsilon}$ $(\varepsilon > 0)$,

$$Y(b, k, n) = \Omega\left(\frac{b^{k-1} \log n}{b^{\underline{k-1}} \log(b-k+2)}\right).$$

Proof. Given a (b, k)-system C, consider the coloring w of $G(b, k, n)$ induced by C as described above. We apply Theorem 1 and show (below) that each partition in C contributes at most

$$\sigma = \frac{\binom{b}{k-2}(b-k+2)(n/b)^{k-1} \log(b-k+2)}{\binom{n}{k-2}(n-k+2)} \tag{16}$$

to the r.h.s. of the inequality in (9). Since the l.h.s. of (9) equals $\log(n-k+2) = \Omega(\log n)$, the number of partitions in C must then be at least $\Omega((\log n)/\sigma)$, which establishes the desired bound.

Now consider the partition P_j of C. The components v_{ij}, $1 \leqq i \leqq \binom{b}{k-2}$, induced by P_j under the coloring w, assume $b-k+2$ possible non-$*$ values. Therefore, the conditional entropies (referred to in (9)) of these components cannot exceed $\log(b-k+2)$. If we count the number L of non-$*$ values contributed by these components to all of the vertices in G, we easily see that this quantity is maximized if P_j has block sizes which are as uniform as possible. Since the sum of the conditional probabilities in (9), associated with the components v_{ij}, is given by L/N, we conclude that the total contribution of any P_j in C to the r.h.s. of (9) cannot exceed (16), completing the proof.

The ratio of the random (upper) bound (5) to the lower bound in Theorem 2 is bounded by

$$O\left(\frac{bk}{(b-k+1)}\right) \log(b-k+2) \qquad (n > b^{2+\varepsilon} \text{ for some fixed } \varepsilon > 0). \tag{17}$$

In particular, the gap in (8) is resolved in favor of the upper bound.

5. Bounds for $M(i, j, n)$. In this section we derive bounds for the minimum size of (i, j)-separating systems, denoted by $M(i, j, n)$. Define

$$Z(i, j) = \frac{(i+j)^{i+j}}{i^i j^j}.$$

For large n (i and j fixed) the volume and random bounds give

$$\Omega(Z(i, j)) = M(i, j, n) = O((i+j)Z(i, j) \log n). \tag{18}$$

(Note. Again the lower bound in (18) does not depend on n.) We now proceed to improve the lower bound by applying Theorem 1. This time we use the graph $\Gamma = \Gamma(i, j, n)$ defined as follows. The vertices of Γ consist of all triples of the form (x, R, S), where R is a subset of U of size $i-1$, S is a subset of U of size $j-1$ disjoint from R; and x is an element of U not in R or S. Two vertices (x, R, S) and (x', R', S') are joined by an edge provided that $R = R'$ and $S = S'$.

Let C be an (i, j)-separating system. We use C to induce a strong coloring of Γ as follows. The vertices of Γ are colored with vectors over $\{*, 0, 1\}$. The mth partition (P_m, Q_m) in C defines the value of the mth coordinate v_m of these vectors; namely,

$$v_m = \begin{cases} 0 & \text{if } \{x\} \cup R \subseteq P_m \text{ and } S \subseteq Q_m, \text{ or } \{x\} \cup S \subseteq P_m \text{ and } R \subseteq Q_m, \\ 1 & \text{if } \{x\} \cup R \subseteq Q_m \text{ and } S \subseteq P_m, \text{ or } \{x\} \cup S \subseteq Q_m \text{ and } R \subseteq P_m, \\ * & \text{otherwise.} \end{cases}$$

To see that this is a strong coloring of Γ, consider two joined vertices, (x, R, S) and (y, R, S). Because C is an (i, j)-separating system and $|\{x\} \cup R| = i$ and $|\{y\} \cup S| = j$, there is some partition (P_j, Q_j) in C such that (say) $\{x\} \cup R \subseteq Q_j$ and $\{y\} \cup S \subseteq P_j$. Thus, for the vertex (x, R, S), $v_j = 1$; and for (y, R, S), $v_j = 0$. In other words, the colors of these two vertices strongly differ.

THEOREM 3. *For fixed $i \leqq j$ and large n,*

$$M(i, j, n) = \Omega\left(Z(i, j) \frac{\log n}{\log((i+j)/i)}\right).$$

Proof. Our proof closely follows the proof of Theorem 2, except that this time the conditional entropies on the r.h.s. of (9) play a more significant role. Since Γ consists of disjoint cliques of size $n+2-i-j$ (one clique for each choice for R and S), we have that $N/\alpha = \log(n+2-i-j)$. Thus, the l.h.s. of (9) is $\Omega(\log n)$. Next we consider the contribution of a given partition (P_m, Q_m) to the r.h.s. of (9). Assume that $|P_m| = t$ and $|Q_m| = n-t$. Then the total number L_0 of 0's assumed by the mth coordinates of the colors of the vertices in Γ is given by

$$L_0 = t\left[\binom{t-1}{i-1}\binom{n-t}{j-1} + \binom{t-1}{j-1}\binom{n-t}{i-1}\right]. \tag{19}$$

The total number L_1 of 1's assumed by the mth coordinates is given by

$$L_1 = (n-t)\left[\binom{t}{i-1}\binom{n-t-1}{j-1} + \binom{t}{j-1}\binom{n-t-1}{i-1}\right]. \tag{20}$$

The total number N of vertices in Γ is given by

$$N = n\binom{n-1}{i-1}\binom{n-i}{j-1}. \tag{21}$$

Let $h(x)$ denote the entropy function $-x \log x - (1-x)\log(1-x)$. The contribution of (P_m, Q_m) to the r.h.s. of (9) is given by

$$c_m = \frac{L_0 + L_1}{N} h\left(\frac{L_0}{L_0 + L_1}\right). \tag{22}$$

Substituting (19), (20), and (21) into (22), writing $t = \theta n$, and assuming that n is large, we obtain

$$c_m = 0(h(\theta)[\theta^{i-1}(1-\theta)^{j-1} + \theta^{j-1}(1-\theta)^{i-1}]). \tag{23}$$

Therefore, we conclude from (9) and (23) that

$$M(i, j, n) = \Omega\left(\log n \cdot \min_{0 < \theta \leqq 1/2} \left[(\theta^i(1-\theta)^j + \theta^j(1-\theta)^i) \cdot \log \frac{1}{\theta}\right]^{-1}\right). \tag{24}$$

Because $i \leqq j$, we have $\theta^i(1-\theta)^j \geqq \theta^j(1-\theta)^i$ when $\theta \leqq 1/2$. Therefore, Theorem 3 is an easy consequence of (24).

The significance of the entropy term in (23) is particularly apparent in the case $i = 1$. Without the entropy term, our lower bound for M would only be $\Omega(\log n)$, with it, the bound becomes $\Omega((j/\log j) \log n)$.

6. Remark. Given a vector over $V = \{0, 1, *\}$, we define its *real weight* to be the total number of 0's and 1's among its components. Given a strong coloring w of a graph G with vectors over V, we define content (G, w) to be the average real weight (over the vertices in G) of the colors of the vertices. We define content$(G) = \min_w$ content(G, w). Referring to the inequality (9) of Theorem 1, noting that $H(w, i) \leqq 1$, we observe that $\log N/\alpha \leqq$content(G). On the other hand, we have content$(G) \leqq \lceil \log X \rceil$ (which can be attained by a coloring without any $*$ components). Since for typical graphs it is known that N/α is a good estimate for X, we conclude that Theorem 1 provides a good estimate for content(G) for most graphs.

REFERENCES

[1] F. Berman, M. E. Bock, E. Dittert, M. J. O'Donnell and D. Plank, *Collections of functions for perfect hashing*, Purdue Univ. Technical Report CSD-TR-408, W. Layfayette, IN, 1982.

[2] A. D. Friedman, R. L. Graham and J. D. Ullman, *Universal single transition time asynchronous state assignments*, IEEE Trans. Comput., C-18 (1969), pp. 541–547.

[3] M. L. Fredman, *The complexity of maintaining an array and computing its partial sums*, J. Assoc. Comput. Mach., 29 (1982), pp. 250–260.

[4] K. Melhorn, *On the program size of perfect and universal hash functions*, Proc. 23rd Annual IEEE Symposium on Foundations of Computer Science, 1982.

[5] A. C. Yao, *Should tables be sorted*, J. Assoc. Comput. Mach., 28 (1981), pp. 615–628.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0010 $01.25/0

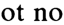
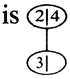
ON THE MOST PROBABLE SHAPE OF A SEARCH TREE GROWN FROM A RANDOM PERMUTATION*

HOSAM MAHMOUD† AND BORIS PITTEL‡

Abstract. The shape of a sequence of trees grown by the progressive ordering of the elements of an infinite random permutation is studied. L_n, the length of the path to the node containing the $(n+1)$st element, is shown to grow, in probability, as $\ln n/(1/2+\cdots+1/m)$, $m-1$ being the capacity of a node. Furthermore, C_n, the number of comparisons needed to insert the $(n+1)$st element grows, in probability, as $L_n(m-1)(m+2)/2m$. We also show that, almost surely, $L_n/\ln n \in [\alpha_1-\varepsilon, \alpha_2+\varepsilon]$, $C_n/\ln n \in [\beta_1-\varepsilon, \beta_2+\varepsilon]$ for large enough $n(\varepsilon>0)$. Here α_i, β_i are roots of two certain equations.

Key words. random search trees, permutations, path lengths, generating functions, distributions, asymptotics, convergence in probability, almost surely

1. Introduction. Consider a sequence $w=(w(1), w(2), \cdots)$ of distinct numbers. One may think of w as a stream of input records for a computer. The computer reads the records, each in turn, and constructs a sequence of trees $\{t_n\}_{n=0}^{\infty}$, such that each $t_n=t_n(w^{(n)})$, $(w^{(n)}=(w(1), w(2), \cdots, w(n)))$ is an m-way search tree defined recursively, essentially in the same way as in [1]. Namely, an m-way search tree on n elements t_n is either empty or each node contains at most $m-1$ elements (keys) $K_1, K_2, \cdots, K_i$, $1 \leqq i \leqq m-1$, sorted left to right $(K_1<K_2<\cdots<K_i)$. Furthermore, (i) if $n<m$ then all the elements are in the root; (ii) if $n \geqq m$ then the first $m-1$ elements are in the root, and the remaining $n-m+1$ elements are distributed among subtrees $T_{n1}, \cdots, T_{nm}$ subject to the following conditions: the elements of T_{n1} are less than all the elements in the root, the elements of T_{ni}, $1<i<m$, lie properly between keys K_{i-1} and K_i of the root, and the elements of T_{nm} are greater than all the elements of the root; $T_{n1}, \ldots, T_{nm}$ are also m-ary search trees.

In what follows we informally describe an algorithm to construct t_n from $w^{(n)}$ by an example. Consider the permutation $w^{(7)}=(4, 2, 3, 5, 7, 6, 1)$. Let us construct a ternary search tree $(m=3)$. Four appears first so it goes into the root and the tree becomes . Two comes along and $2<4$ and the root node is not filled yet. Thus, 2 also goes into the root, and it pushes the 4 to the right, so that the labels of the root are sorted left to right. The tree now is . Three is next; $2<3<4$ so it goes into the middle subtree. The tree now is . The construction proceeds in this manner until we end up with the full tree shown in Fig. 1.

It is interesting to notice that more than one sequence may give rise to the same tree under this construction scheme. For example, the sequence $\tilde{w}^{(7)}=(4, 2, 1, 5, 7, 6, 3)$ will give the same tree as in Fig. 1. In fact, two n-long sequences will give rise to the same *step-by-step* realization of the algorithm provided that their integer-valued vectors of sequential ranks are identical. ($r^{(n)}=(r(1), \cdots, r(n))$ is the vector of sequential ranks of a sequence $w^{(n)}$ if $w(j)$ is the $r(j)$th largest in comparison with $w(k)$, $1 \leqq k \leqq j$.) When t_n is constructed there are exactly $n+1$ available positions for the $(n+1)$st element to be put in, counting vacant slots in partially filled nodes of t_n plus new (still empty) nodes adjacent to the old nodes. The choice of this position is uniquely determined by the value of $r(n+1)$.

* Received by the editors July 19, 1982, and in revised form March 8, 1983.

† Computer Science Department, The Ohio State University, Columbus, Ohio 43210.
‡ Mathematics Department, The Ohio State University, Columbus, Ohio 43210.

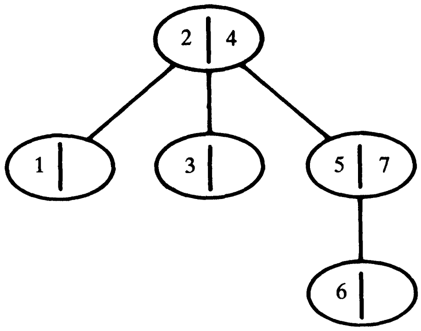


FIG. 1.

Certain characteristics of the tree t_n, such as its depth, turn out to be very instrumental for the problem of estimating the time needed to search for a record already present in the tree, or expand it to insert a new record. Our goal is to study the asymptotical behavior of related characteristics of $\{t_n\}_{n=1}^{\infty}$ under the assumption that w is an infinite random permutation. By the infinite random permutation we mean a random sequence whose vector of sequential ranks satisfies:

a) $r(j)$ is uniformly distributed on $R_j = \{1, \cdots, j\}$,

b) $r(1), r(2) \cdots$ are independent.

Notation. Let L_n stand for the length of the path from the root to the node which will contain $w(n+1)$, and let H_n, h_n denote the length of the longest and the shortest paths respectively from the root to the $(n+1)$ potential positions for $w(n+1)$. We shall also be interested in C_n, the number of comparisons required to find the place for $w(n+1)$. It should be clear that C_n depends on how the current number $w(n+1)$ is compared with the elements in each node along the path to its destination. We assume that these comparisons are made sequentially from left to right. It can be easily seen that

$$0 \leqq h_n \leqq \lfloor \log_m (n+1) \rfloor, \qquad \lceil \log_m (n+1) \rceil \leqq H_n \leqq \left\lfloor \frac{n}{m-1} \right\rfloor,$$

and all bounds are attainable. Thus, these characteristics may vary widely from one sequence w to the other, which means it is natural to study their distributions for large n.

For the special case of binary trees t_n $(m = 2)$, the distribution of L_n has long been known, Lynch [2], Knuth [3]; namely,

$$P(L_n = k) = 2^k S(n, k)/(n+1)!, \qquad 1 \leqq k \leqq n,$$

where $S(n, k)$ are the Stirling numbers of the first kind [4]. Knuth also showed that $E(L_n)$, the expected value of L_n, is $2 \ln n + O(1)$ as $n \to \infty$.[1] Later Robson [5] proved, for the same tree, that its height H_n with high probability is also logarithmically bounded as $n \to \infty$.

Ruskey [6] studied expected lengths of the paths leading from the root to all external nodes in the extended binary tree (see [3] for definitions and properties). But the random tree he considered is completely different, namely each of $\binom{2n}{n}/(n+1)$ binary trees with n internal nodes is equally likely. Among other results we should mention Rényi–Szekeres [7], Stepanov [8], De Bruijn, Knuth and Rice [9]; the former

[1] *Added in proof*: Using algebraic properties of the numbers $S(n, k)$, we have shown recently that L_n is asymptotically Gaussian with mean and variance both equal to $2 \ln n$.

two contain the limiting distribution of the height of the unordered random tree, the latter gives asymptotical value of the expected height for the ordered random tree, and in both models all the feasible trees are equally likely. (For a very comprehensive survey of numerous other results the reader is referred to Moon [10] and Flajolet and Odlyzko [11].)

We now can formulate our main results.

THEOREM 1. *In probability,*

$$\lim_{n\to\infty}\frac{L_n}{\ln n}=a(m)=\left(\frac{1}{2}+\cdots+\frac{1}{m}\right)^{-1}, \tag{1.1}$$

$$\lim_{n\to\infty}\frac{C_n}{\ln n}=b(m)=\left[\frac{(m-1)(m+2)}{2m}\right]a(m). \tag{1.2}$$

Remarks. It will follow from the proof of (1.1) that, with high probability, all but a negligible fraction of $n+1$ paths leading to possible locations for $w(n+1)$ have lengths between $(a(m)-\varepsilon)\ln n$ and $(a(m)+\varepsilon)\ln n$, $\varepsilon>0$. Also, observe that $a(2)=b(2)=b(3)=2>a(3)=1.2$. Thus cramming two elements into each node, instead of just one, leads, with high probability, to noticeably shorter trees while the typical values of C_n, the number of necessary comparisons, are kept essentially the same ($\sim 2\ln n$).

While Theorem 1 describes the property of almost all trees t_n for a fixed (but large) n, Theorems 2 and 3 concern the asymptotical behavior of almost all *infinite* sequences of the trees t_n.

THEOREM 2. *With probability one,*

$$\liminf_{n}\frac{L_n}{\ln n}\geqq\alpha_1, \tag{1.3}$$

$$\limsup_{n}\frac{L_n}{\ln n}\leqq\alpha_2; \tag{1.4}$$

here

$$\alpha_i=\left[\sum_{j=0}^{m-2}(x_i+j)^{-1}\right]^{-1}, \tag{1.5}$$

and $0<x_1<x_2<\infty$ *are the roots of the equation*

$$f(x)=x+\left[\sum_{j=0}^{m-2}(x+j)^{-1}\right]^{-1}\sum_{k=0}^{m-2}\ln[(k+2)(k+x)^{-1}]-1=0. \tag{1.6}$$

THEOREM 3. *With probability one,*

$$\liminf_{n}\frac{C_n}{\ln n}\geqq\beta_1, \tag{1.7}$$

$$\limsup_{n}\frac{C_n}{\ln n}\leqq\beta_2; \tag{1.8}$$

here $\beta_i=\beta(x_i)=y(x_i)/y'(x_i)$, $y=y(x)$ *being the positive root of the equation*

$$\sum_{j=1}^{m-1}y^j+y^{m-1}=\langle x\rangle_{m-1}\qquad\left(\langle x\rangle_\nu\overset{\text{def}}{=}x(x+1)\cdots+(x+\nu-1)\right), \tag{1.9}$$

and $0<x_1<x_2<\infty$ *are the roots of the equation*

$$g(x)=x-\beta(x)\ln y(x)-1=0. \tag{1.10}$$

Loosely speaking, these statements mean that both L_n and C_n considered as random functions of n are almost surely bounded above and below by logarithmic functions of n.

Notes. (a) Consider the special case of the binary trees t_n, i.e., $m=2$. Then $C_n=L_n$ and, as should be expected, $\beta_1=\alpha_1, \beta_2=\alpha_2$, where (see (1.5), (1.6)) $0<\alpha_1<\alpha_2<\infty$ are the roots of the equation

$$\alpha+\alpha\ln\left(\frac{2}{\alpha}\right)-1=0,$$

$\alpha_1\approx 0.37, \alpha_2\approx 4.31$. The number α_2 already appeared in [5], where H_n, the height of t_n, was shown not to exceed, with high probability, $(\alpha_2+\varepsilon)\ln n, \varepsilon>0$. There were also given in [5] some ingenious, though incomplete, arguments intended to prove that $E(H_n)\geqq 3.63\ln n+o(\ln n)$.

One of the authors [12] recently proved the existence of two constants $c_1\in[0.37, 0.50]$ and $c_2\in[3.58, 4.32]$ such that, with probability one,

$$\lim_{n\to\infty}\frac{h_n}{\ln n}=c_1, \qquad \lim_{n\to\infty}\frac{H_n}{\ln n}=c_2;$$

(remember, h_n is the length of the shortest path from the root of t_n to the possible location of $w(n+1)$). As $h_n\leqq L_n\leqq H_n$, and $L_n=h_n$ and $L_n=H_n$ infinitely often almost surely, we conclude that in this case, with probability one,

$$\liminf_n\frac{L_n}{\ln n}=c_1, \qquad \limsup_n\frac{L_n}{\ln n}=c_2.$$

It is worth remembering (Theorem 1) that still in probability

$$\lim_{n\to\infty}\frac{L_n}{\ln n}=2\in(c_1,c_2).$$

(b) Table 1 is the table of α_i, β_i, $(i=1,2)$, rounded to three decimal places for some small values of m.

The table shows that $\alpha_1=\beta_1$ and $\alpha_2=\beta_2$ for $m=2$; it must be expected as $C_n=L_n$ in this case. What is more surprising is the fact that β_1 and β_2 coincide for $m=2$ and

TABLE 1

m	α_1	α_2	β_1	β_2
2	0.373	4.311	0.373	4.311
3	0.318	4.490	0.373	4.311
4	0.287	4.636	0.350	4.552
5	0.267	4.759	0.328	4.865
6	0.253	4.867	0.310	5.202
7	0.242	4.963	0.297	5.548
8	0.233	5.049	0.286	5.897
9	0.226	5.128	0.277	6.243
10	0.219	5.200	0.270	6.585

$m=3$. This is a direct consequence of a stronger result which follows from Lemma 1 below: the distributions of C_n for $m=2$ and $m=3$ are the same.

2. Proofs.

Notation. Let X_{nk} be the random number of possible positions for $w(n+1)$ at distance k from the root; clearly, $X_{0k}=\delta_{0k}$. Let Y_{nk} be the random number of possible positions for $w(n+1)$ such that to reach one of them $w(n+1)$ has to be compared with exactly k elements of $w^{(n)}=(w(1),\cdots,w(n))$. Denote $F_{nk}=E(X_{nk})$, $G_{nk}=E(Y_{nk})$. The numbers F_{nk}, G_{nk} are closely associated with the distributions of L_n and C_n. Namely, according to the distribution of $r=\{r(n)\}_{n=1}^{\infty}$,

$$P(L_n=k|t_n)=\frac{X_{nk}}{n+1},\qquad P(C_n=k|t_n)=\frac{Y_{nk}}{n+1};$$

so averaging over t_n we have

$$P(L_n=k)=\frac{F_{nk}}{n+1},\qquad P(C_n=k)=\frac{G_{nk}}{n+1}. \tag{2.1}$$

Introduce the generating functions of F_{nk}, G_{nk}:

$$\begin{aligned} &F_k(x)=\sum_{n\geqq 0}F_{nk}x^n,\quad F_n(y)=\sum_{k\geqq 0}F_{nk}y^k,\quad F(x,y)=\sum_{n\geqq 0,k\geqq 0}F_{nk}x^ny^k,\\ &G_k(x)=\sum_{n\geqq 0}G_{nk}x^n,\quad G_n(y)=\sum_{k\geqq 0}G_{nk}y^k,\quad G(x,y)=\sum_{n\geqq 0,k\geqq 0}G_{nk}x^ny^k, \end{aligned} \tag{2.2}$$

LEMMA 1. $F(x,y)$, $G(x,y)$ *satisfy*

$$(1-x)^{m-1}\left[\frac{\partial^{m-1}F}{\partial x^{m-1}}\right]=(ym!)F, \tag{2.3}$$

$$\frac{\partial^iF(0,y)}{\partial x^i}=(i+1)!,\qquad 0\leqq i\leqq m-2, \tag{2.4}$$

and

$$(1-x)^{m-1}\left[\frac{\partial^{m-1}G}{\partial x^{m-1}}\right]=(\rho_{m-1}(y)(m-1)!)G, \tag{2.5}$$

$$\frac{\partial^iG(0,y)}{\partial x^i}=i!\rho_i(y),\qquad 0\leqq i\leqq m-2, \tag{2.6}$$

$$\rho_0(y)\equiv 1,\qquad \rho_t(y)=\sum_{1\leqq s\leqq t}y^s+y^t\quad \textit{for } t\geqq 1. \tag{2.7}$$

Proof. (a) Observe first that

$$F_{n0}=n+1\quad\text{if } 0\leqq n\leqq m-2,\qquad F_{n0}=0\quad\text{if } n\geqq m-1, \tag{2.8}$$

$$F_{nk}=0\quad\text{if } 0\leqq n\leqq m-2 \text{ and } k\geqq 1. \tag{2.9}$$

Let $n\geqq m-1$ and $T_{n1},\cdots,T_{nm}$ be the subtrees of t_n ordered from left to right, whose roots are adjacent to the root of t_n (see Introduction). If τ_{nj} stands for the size of T_{nj}, i.e. the number of elements of $w^{(n)}$ contained in its nodes, then

$$\sum_{j=1}^{m}\tau_{nj}=n-m+1. \tag{2.10}$$

Let us show by induction that the random vector $\tau_n = \{\tau_{nj}\}_{j=1}^m$ is uniformly distributed on the set Ω_{nm} of nonnegative integer-valued solutions of (2.10), in other words that

$$P(\tau_{nj} = i_j,\ 1 \leqq j \leqq m) = |\Omega_{nm}|^{-1} = \binom{n}{m-1}^{-1}, \tag{2.11}$$

if

$$\sum_{j=1}^m i_j = n - m + 1, \qquad i_j \geqq 0.$$

It is obviously true for $n = m - 1$; assume that it holds for $n = \nu \geqq m - 1$. As the sequential rank $r(\nu+1)$ of $w(\nu+1)$ is independent of $r(1), \cdots, r(\nu)$ and uniformly distributed on $\{1, \cdots, \nu+1\}$, by the induction hypothesis, we have: if $\sum_{j=1}^m i_j = (\nu+1) - m + 1$, then

$$\begin{aligned} &P(\tau_{\nu+1,j} = i_j,\ 1 \leqq j \leqq m) \\ &\quad = \sum_{s:\, i_s \geqq 1} \left(\frac{i_s}{\nu+1}\right) P(\tau_{\nu j} = i_j,\ 1 \leqq j \leqq m \text{ and } j \neq s,\ \tau_{\nu s} = i_s - 1) \\ &\quad = \binom{\nu}{m-1}^{-1} (\nu+1)^{-1} \sum_{s=1}^m i_s = \binom{\nu}{m-1}^{-1} (\nu+1)^{-1}(\nu+2-m) = \binom{\nu+1}{m-1}^{-1}. \end{aligned}$$

Furthermore, for $k \geqq 1$ (and $n \geqq m - 1$)

$$X_{nk} = \sum_{j=1}^m \mathscr{X}_{k-1}^{(j)}, \tag{2.12}$$

where $\mathscr{X}_{k-1}^{(j)}$ is the number of positions available for $w(n+1)$ in T_{nj}, which are $(k-1)$ steps apart from its root. Notice that

$$P(\mathscr{X}_{k-1}^{(j)} = a \mid \tau_{nj} = b) = P(X_{b,k-1} = a), \qquad 1 \leqq j \leqq m, \tag{2.13}$$

for all a and b. Taking expectations of both sides of (2.12), and invoking (2.11), (2.13), leads to

$$F_{nk}\binom{n}{m-1} = \sum_{i_1, \cdots, i_m} \sum_{j=1}^m F_{i_j, k-1}, \qquad n \geqq m-1, \quad k \geqq 1; \tag{2.14}$$

here the outer sum extends over all solutions of (2.10).

Therefore, using the notation $(n)_{m-1} = n(n-1) \cdots (n-m+2)$, we arrive at

$$\begin{aligned} F_k^{(m-1)}(x) &= \sum_{n \geqq m-1} (n)_{m-1} F_{nk} x^{n-m+1} \\ &= (m-1)! \cdot \sum_{i_s \geqq 0, 1 \leqq s \leqq m} \prod_{t=1}^m x^{i_t} \left(\sum_{j=1}^m F_{i_j, k-1}\right) \\ &= m! \left(\sum_{i \geqq 0} x^i F_{i,k-1}\right)\left(\sum_{\nu \geqq 0} x^\nu\right)^{m-1}, \end{aligned} \tag{2.15}$$

or

$$(1-x)^{m-1} F_k^{(m-1)}(x) = m! F_{k-1}(x), \qquad k \geqq 1. \tag{2.16}$$

Since, in view of (2.8),

$$F_0(x)=\sum_{n\geqq 0} F_{n0}x^n=\sum_{0\leqq n\leqq m-2}(n+1)x^n, \tag{2.17}$$

the relation (2.16) implies

$$(1-x)^{m-1}\partial^{m-1}F(x,y)/\partial x^{m-1}=(1-x)^{m-1}\sum_{k\geqq 1}F_k^{(m-1)}(x)y^k=(m!\,y)F(x,y).$$

Initial conditions (2.4) follow directly from (2.8), (2.9).

(b) As in (2.12), for $k\geqq 1$ and $n\geqq m-1$,

$$Y_{nk}=\sum_{1\leqq j\leqq m-1}\mathcal{Y}^{(j)}_{k-j}+\mathcal{Y}^{(m)}_{k-m+1},$$

where, for $s\geqq 0$, $\mathcal{Y}^{(j)}_s$ is the number of positions available for $w(n+1)$ in T_{nj} which are s comparisons away from the root of T_{nj}, and $\mathcal{Y}^{(j)}_s=0$ for $s<0$ $(1\leqq j\leqq m)$.

Define $G_{nk}=0$ for $n\geqq 0$, $k<0$. Then, as in (2.15),

$$\begin{aligned} G_k^{(m-1)}(x)&=(m-1)!\cdot\sum_{i_s\geqq 0,1\leqq s\leqq m}\left(\prod_{t=1}^{m}x^{i_t}\right)\left(\sum_{j=1}^{m-1}G_{i_j,k-j}+G_{i_m,k-m+1}\right)\\ &=(m-1)!\,(1-x)^{-m+1}\left[\sum_{j=1}^{m-1}\left(\sum_{i\geqq 0}G_{i,k-j}x^i\right)+\sum_{i\geqq 0}G_{i,k-m+1}x^i\right],\end{aligned}$$

or

$$(1-x)^{m-1}G_k^{(m-1)}(x)=(m-1)!\left[\sum_{j=1}^{m-1}G_{k-j}(x)+G_{k-m+1}(x)\right] \tag{2.18}$$

(in the right-hand expression, $G_s(x)\equiv 0$ for $s<0$).

By definition of G_{nk}, we also have $G_{0k}=\delta_{0k}$, and, for $1\leqq n\leqq m-2$,

$$G_{nk}=\begin{cases}0 & \text{if } k=0 \text{ or } k\geqq n+1,\\ 1 & \text{if } 1\leqq k\leqq n-1,\\ 2 & \text{if } k=n.\end{cases} \tag{2.19}$$

In particular, $G_0(x)\equiv 1$, so according to (2.18),

$$\begin{aligned}(1-x)^{m-1}&\frac{\partial^{m-1}G(x,y)}{\partial x^{m-1}}\\ &=(m-1)!\sum_{k\geqq 1}y^k\left[\sum_{j=1}^{m-1}G_{k-j}(x)+G_{k-m+1}(x)\right]\\ &=(m-1)!\left[\sum_{j=1}^{m-1}y^j\sum_{k\geqq j}G_{k-j}(x)y^{k-j}+y^{m-1}\sum_{k\geqq m-1}G_{k-(m-1)}(x)y^{k-(m-1)}\right]\\ &=(m-1)!\left(\sum_{j=1}^{m-1}y^j+y^{m-1}\right)G(x,y)=(m-1)!\,\rho_{m-1}(y)G(x,y).\end{aligned}$$

Conditions (2.6) follow from (2.19). The lemma is thus proven.

LEMMA 2. *Let $y>0$ and $\lambda=\lambda(y)$, $\sigma=\sigma(y)$ be the positive roots of the equations*

$$\langle z\rangle_{m-1}=ym!,\qquad \langle z\rangle_{m-1}=\rho_{m-1}(y)(m-1)! \tag{2.20}$$

($\langle\xi\rangle_\mu=\xi(\xi+1)\cdots(\xi+\mu-1)$, for $\mu\geqq1$). *Then*

$$F_n(y)\leqq\gamma\langle\lambda\rangle_n/n!,\qquad \gamma=\gamma(\lambda)=\max\left[1,\frac{(m-1)(m-2)}{\lambda}\right],\tag{2.21}$$

$$G_n(y)\leqq\delta\langle\sigma\rangle_n/n!,\qquad \delta=\delta(m)\qquad(\delta(2)=1),\tag{2.22}$$

and (2.21), (2.22) *become identities for* $m=2$.

Remark. In the binary case ($m=2$), we have $\lambda=\sigma=2y$ and

$$F_n(y)=G_n(y)=\langle 2y\rangle_n/n!.$$

So, using a well-known identity [4]

$$\langle\xi\rangle_\mu=\sum_{\nu=0}^{\mu}S(\mu,\nu)\xi^\nu,$$

where $S(\mu,\nu)$ are the Stirling numbers of the first kind, we get (see (2.1) (2.2))

$$F_{nk}=G_{nk}=\frac{2^kS(n,k)}{n!}$$

and

$$P(L_n=k)=\frac{2^kS(n,k)}{(n+1)!},$$

which is the Lynch–Knuth formula mentioned in the Introduction.

Proof of Lemma 2. As the arguments for proving (2.21) and (2.22) are quite similar, we shall prove only (2.21). Notice first that

$$\bar F(x,y)=(1-x)^{-\lambda}$$

is a solution of (2.3) which satisfies the conditions

$$\partial^i\bar F(0,y)/\partial x^i=\langle\lambda\rangle_i,\qquad 0\leqq i\leqq m-2,\tag{2.23}$$

($\langle\lambda\rangle_0=1$, by definition). In view of (2.4) then

$$F(x,y)=\bar F(x,y)$$

for $m=2$, and subsequently

$$F_n(y)=\text{coeff}_{x^n}\,(1-x)^{-\lambda}=\langle\lambda\rangle_n/n!=\langle 2y\rangle_n/n!.$$

Let now $m\geqq3$. Simple arguments based on (2.20) and comparison of (2.4) and (2.23) show that

$$\partial^i\tilde F(0,y)/\partial x^i\geqq\partial^iF(0,y)/\partial x^i,\qquad 0\leqq i\leqq m-2,\tag{2.24}$$

where

$$\tilde F(x,y)=\gamma\bar F(x,y)=\gamma(1-x)^{-\lambda},\qquad \gamma=\max_{0\leqq i\leqq m-2}[(i+1)!/\langle\lambda\rangle_i].$$

Now, if

$$\tilde F(x,y)=\sum_{n\geqq0}\tilde F_n(y)x^n,$$

then ($F(x,y)=\sum_{n\geqq0}F_n(y)x^n$), (2.24) is equivalent to

$$F_n(y)\leqq\tilde F_n(y),\qquad 0\leqq n\leqq m-2.\tag{2.25}$$

Next, we show that (2.25) is also true for $n \geqq m-1$. Both $F(x, y)$ and $\tilde{F}(x, y)$ as functions of x are solutions of the equation

$$\frac{d^{m-1}\psi}{dx^{m-1}} = (1-x)^{-m+1}(ym!)\psi.$$

So, using Taylor's expansions of F and $\tilde{F}$ about $x = 0$ and the binomial series

$$(1-x)^{-m+1} = \sum_{\mu=0}^{\infty} \binom{m+\mu-2}{m-2} x^{\mu},$$

we obtain the recurrence relations: for $n \geqq m-1$,

$$(n)_{m-1}\psi_n = (ym!) \cdot \sum_{j=0}^{n-m+1} \binom{n-1-j}{m-2} \psi_j, \qquad \psi_n = F_n(y) \text{ or } \tilde{F}_n(y).$$

Invoking positivity of $(n)_{m-1}$ and $\binom{n-1-j}{m-2}$ for $n \geqq m-1$ and (2.25), by induction we conclude that

$$(2.26) \qquad F_n(y) \leqq \tilde{F}_n(y),$$

for all $n \geqq 0$. The lemma is thus proved.

Proof of Theorem 1. (a) by Lemma 2 (see (2.21),

$$F_{nk} \leqq \gamma \frac{\langle \lambda \rangle_n}{y^k n!}, \qquad y > 0,$$

or, equivalently,

$$(2.27) \qquad F_{nk} \leqq \gamma \frac{\langle x \rangle_n}{y^k n!}, \quad y = (m!)^{-1}\langle x \rangle_{m-1}, \quad \gamma = \gamma(x), \quad x \in (0, \infty).$$

Observe that $y = y(x)$ is strictly increasing, $y(0) = 0$, $\lim_{x\to\infty} y(x) = \infty$,

$$(2.28) \qquad y(0+) = 0, \quad y(2) = 1, \quad \lim_{x\to\infty} y(x) = \infty.$$

Introduce the function

$$(2.29) \qquad \alpha = \alpha(x) = (x^{-1} + \cdots + (x+m-2)^{-1})^{-1}, \qquad x \in (0, \infty);$$

obviously, $\alpha(x)$ is also strictly increasing,

$$\lim_{x\to 0+} \alpha(x) = 0, \qquad \lim_{x\to\infty} \alpha(x) = \infty.$$

Denote also $k(x, n) = \alpha(x) \ln n$. We want to show that

$$(2.30) \qquad \sum_{k \geqq k(x,n)} F_{nk} \leqq \gamma_1(x)(1-y^{-1})^{-1} \exp[f(x) \ln n] \quad \text{if } x > 2,$$

$$(2.31) \qquad \sum_{k \geqq k(x,n)} F_{nk} \leqq \gamma_2(x)(1-y)^{-1} \exp[f(x) \ln n] \quad \text{if } x < 2,$$

where $f(x) = x - 1 - \alpha \ln y$, and $y = y(x)$ is defined in (2.27).

Consider, for example, $x > 2$. Clearly $y = y(x) > 1$, so by (2.27)

$$\sum_{k \geqq k(x,n)} F_{nk} \leqq \gamma(1-y^{-1})^{-1} \frac{\langle x \rangle_n}{y^{k(x,n)} n!}.$$

By the Stirling formula for the gamma-function,

$$\Gamma(z+1) = \sqrt{2\pi z} \left(\frac{z}{e}\right)^z (1 + o(1)), \qquad z \to \infty,$$

we have

$$\frac{\langle x\rangle_n}{n!}=\Gamma(x+n)\Gamma^{-1}(x)\Gamma^{-1}(n+1)$$
$$=\Gamma^{-1}(x)\exp[(x-1)\ln n](1+o(1)).$$

Therefore,

$$\sum_{k\geqq k(x,n)} F_{nk}\leqq\tilde{\gamma}(x)(1-y^{-1})^{-1}\exp[f(x)\ln n].$$

The case $x<2$ is treated similarly.

It is easy to check that

$$\lim_{x\to0+}f(x)=-1,\quad f(2)=1,\quad \lim_{x\to\infty}f(x)=-\infty, \tag{2.32}$$

and

$$f'(x)=-(\ln y)\alpha^2(x)[x^{-2}+\cdots+(x+m-2)^{-2}]; \tag{2.33}$$

so $f(x)$ is unimodal and

$$\max\{f(x):0<x<\infty\}=f(2)=1. \tag{2.34}$$

Then, in view of (2.1), (2.30)–(2.34), for all $0<x'<2<x''<\infty$,

$$\begin{aligned}&P(L_n\leqq\alpha(x')\ln n,\text{ or }L_n\geqq\alpha(x'')\ln n)\\ &\leqq n^{-1}\Bigl(\sum_{k\leqq k(x',n)}F_{nk}+\sum_{k\geqq k(x'',n)}F_{nk}\Bigr)\\ &\leqq\gamma(x',x'')\exp[(\ln n)\max(f(x')-1,f(x'')-1)]\to0,\end{aligned} \tag{2.35}$$

as $n\to\infty$. Since

$$\alpha(2)=(2^{-1}+\cdots+m^{-1})^{-1},$$

it follows then that, in probability,

$$\frac{L_n}{\ln n}\to\alpha(2). \tag{2.36}$$

(b) Similarly to (2.27),

$$G_{nk}\leqq\delta\frac{\langle x\rangle_n}{y^k n!},\qquad x>0, \tag{2.37}$$

where y is the positive root of

$$\rho_{m-1}(y)(m-1)!=\langle x\rangle_{m-1},\qquad \rho_{m-1}(y)=\sum_{j=1}^{m-1}y^j+y^{m-1}. \tag{2.38}$$

This shows that $y=y(x)$ is strictly increasing and satisfies (2.28). It can be also verified that $y'(x)\in[C_1,C_2]$ where $C_1<C_2$ are two positive constants. Introduce the function

$$\beta(x)=y(x)/y'(x); \tag{2.39}$$

clearly, $\beta(0+)=0$, $\lim_{x\to\infty}\beta(x)=\infty$. Then, exactly as in (a), one can show that

$$\sum_{k\geqq\chi(x,n)} G_{nk} \leqq \delta_1(x)(1-y^{-1})^{-1}\exp[g(x)\ln n] \quad \text{if } x>2, \tag{2.40}$$

$$\sum_{k\leqq\chi(x,n)} G_{nk} \leqq \delta_2(x)(1-y^{-1})^{-1}\exp[g(x)\ln n] \quad \text{if } x<2, \tag{2.41}$$

$$\chi(x,n)=\beta(x)\ln n, \qquad g(x)=x-1-\beta(x)\ln y. \tag{2.42}$$

Now, as in (2.32),

$$g(0+)=-1, \quad g(2)=1, \quad \lim_{x\to\infty} g(x)=-\infty.$$

Let us demonstrate also that, like $f(x)$, $g(x)$ is unimodal. We have, (see (2.42)),

$$g'(x)=1-\beta'(x)\ln y(x)-\beta(x)\frac{y'(x)}{y(x)}=-\beta'(x)\ln y(x).$$

Since $y(x)=1$ iff $x=2$, unimodality will follow if we prove that $\beta'(x)>0$ for $x\in(0,\infty)$. Taking the logarithmic derivative of both sides of (2.38) leads to, (see (2.39)),

$$\beta(x)=[x^{-1}+\cdots+(x+m-2)^{-1}]^{-1}(y\rho'_{m-1}(y)/\rho_{m-1}(y)). \tag{2.43}$$

Notice that $y'(x)>0$, and $[x^{-1}+\cdots+(x+m-2)^{-1}]^{-1}$ has a positive derivative, too. Now, as

$$\rho_{m-1}(y)=\sum_{j=1}^{m-1}\omega_j y^j, \qquad \omega_j>0,$$

we also have

$$\frac{d(y\rho'_{m-1}(y)\rho^{-1}_{m-1}(y))}{dy}=[y\rho^2_{m-1}(y)]^{-1}\left[\left(\sum_{j=1}^{m} j^2\omega_j y^j\right)\left(\sum_{j=1}^{m}\omega_j y^j\right)-\left(\sum_{j=1}^{m} j\omega_j y^j\right)^2\right]>0,$$

by the Cauchy–Schwarz inequality. Thus $\beta'(x)>0$, $g(x)$ is unimodal and

$$\max\{g(x):0<x<\infty\}=g(2)=1.$$

The rest of the proof goes exactly as in (a), and we get: in probability,

$$\frac{C_n}{\ln n}\to\beta(2),$$

where (as $y(2)=1$)

$$\begin{aligned}\beta(2)&=[2^{-1}+\cdots+m^{-1}]^{-1}\rho'_{m-1}(1)/\rho_{m-1}(1)\\ &=[2^{-1}+\cdots+m^{-1}]^{-1}(m-1)(m+2)(2m)^{-1}.\end{aligned}$$

Theorem 1 is proven.

Proof of Theorem 2. Notice first that

$$h_n\leqq L_n\leqq H_n. \tag{2.44}$$

(a) According to the proof of Theorem 1, $f(x)$ is unimodal, and there exist two positive roots $0<x_1<2<x_2<\infty$ of $f(x)=0$, so that $f(x)>0$ for $x\in(x_1,x_2)$ and $f(x)<0$ for $x\in(x_1,x_2)^c$.

Given $\varepsilon > 0$, introduce $k(\varepsilon, n) = \alpha(x_2 + \varepsilon) \ln n$, $\alpha(x)$ being defined in (2.29). As $x_2 > 2$, by (2.30) we have

$$P(H_n \geqq k(\varepsilon, n)) = P\Big(\bigcup_{k \geqq k(\varepsilon,n)} (X_{nk} > 0)\Big)$$
$$\leqq \sum_{k \geqq k(\varepsilon,n)} P(X_{nk} > 0) \leqq \sum_{k \geqq k(\varepsilon,n)} E(X_{nk}) = \sum_{k \geqq (\varepsilon,n)} F_{nk}$$
$$\leqq c \exp[f(x_2 + \varepsilon) \ln n] = cn^{-c_1}, \qquad c = c(\varepsilon) > 0, \quad c_1 = c_1(\varepsilon) > 0.$$

So

$$P(H_n \geqq \alpha(x_2 + \varepsilon) \ln n) \leqq cn^{-c_1}. \tag{2.45}$$

Similarly,

$$P(h_n \leqq \alpha(x_1 - \varepsilon) \ln n) \leqq dn^{-d_1}, \quad d = d(\varepsilon) > 0, \quad d_1 = d_1(\varepsilon) \geqq 0. \tag{2.46}$$

(b) Let us show that (2.45) implies that

$$P\Big(\limsup_n \frac{H_n}{\ln n} \leqq \alpha_2\Big) = 1,$$

in other words that, for each $\alpha > \alpha_2$,

$$P(H_n \geqq \alpha \ln n \text{ infinitely often}) = 0.$$

To this end it would suffice to show (Borel–Cantelli lemma) that

$$\sum_{n=1}^{\infty} P(H_n \geqq \alpha \ln n) < \infty. \tag{2.47}$$

But (2.47) does not follow from (2.45) as c_1 depends on ε, and, in fact, goes to 0 as ε gets smaller. Still, a simple idea, (see also [13], [14]), based on the observation that H_n increases with n, and $\ln n$ is slowly varying, helps to overcome this obstacle.

Choose an integer k so large that $kc_1 > 1$. Then

$$\sum_{\nu=1}^{\infty} P(H_{\nu^k} \geqq \alpha(x_2 + \varepsilon) \ln(\nu^k)) \leqq c \sum_{\nu=1}^{\infty} \nu^{-kc_1} < \infty, \qquad \varepsilon > 0,$$

so that (Borel–Cantelli!)

$$P\Big(\limsup_\nu \frac{H_{\nu^k}}{\ln(\nu^k)} \leqq \alpha_2\Big) = 1, \qquad \alpha_2 = \alpha(x_2).$$

Further, given n, let $\nu(n)$ be determined by

$$\nu^k(n) \leqq n < (\nu(n) + 1)^k;$$

clearly, $\nu(n) = n^{1/k}(1 + o(1)) \to \infty$ as $n \to \infty$. Since

$$\frac{H_n}{\ln n} \leqq \frac{H_{(\nu(n)+1)^k}}{\ln(\nu^k(n))} = \Big[\frac{H_{(\nu(n)+1)^k}}{\ln((\nu(n)+1)^k)}\Big]\Big(1 + O\Big(\frac{1}{\nu(n) \ln \nu(n)}\Big)\Big),$$

we have that

$$\limsup_n \frac{H_n}{\ln n} \leqq \alpha_2 \tag{2.48}$$

with probability one, too.

Similarly, with probability one

$$\liminf_n \frac{h_n}{\ln n} \geqq \alpha_1, \qquad \alpha_1 = \alpha(x_1). \tag{2.49}$$

A combination of (2.44), (2.48), (2.49) yields finally that with probability one

$$\liminf_n \frac{L_n}{\ln n} \geqq \alpha_1, \qquad \limsup_n \frac{L_n}{\ln n} \leqq \alpha_2.$$

Theorem 2 is proven.

The proof of Theorem 3 can be done essentially in the same way, and we omit it.

Acknowledgment. The authors are grateful to Dr. Bruce Weide for introducing one of them to the problem. Comments made by a referee helped us improve the presentation.

REFERENCES

[1] E. Horowitz and S. Sahni, *Fundamentals of Data Structures*, Computer Science Press, Potomac, MD, 1976.

[2] W. C. Lynch, *More combinatorial problems of certain trees*, Comput. J., 7 (1965), pp. 299–302.

[3] D. E. Knuth, *The Art of Computer Programming, Vol.* 3, Addison-Wesley, Reading, MA, 1973.

[4] L. Comtet, *Advanced Combinatorics; The Art of Finite and Infinite Expansions*, D. Reidel, Dordrecht, Boston, 1974.

[5] J. M. Robson, *The height of binary search trees*, Austral. Comput. J., 11 (1979), pp. 151–153.

[6] F. Ruskey, *On the average shape of binary trees*, this Journal, 1 (1980), pp. 43–50.

[7] A. Rényi and G. Szekeres, *On the height of trees*, J. Austral. Math. Soc., 7 (1967), pp. 497–507.

[8] V. E. Stepanov, *On the distribution of the number of vertices in strata of a random tree*, Theory Prob. Appl., 14 (1969), pp. 65–78.

[9] N. G. De Bruijn, D. E. Knuth and S. O. Rice, *The average height of planted plane trees*, in Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1972, pp. 15–22.

[10] J. W. Moon, *Counting labelled trees*, Canadian Mathematical Congress, 1970.

[11] P. Flajolet and A. Odlyzko, *Exploring binary trees and other simple trees*, J. Comput. System Sci., 25 (1982), pp. 171–213.

[12] B. Pittel, *On growing random binary trees*, J. Math. Anal. Appl., to appear.

[13] J. F. C. Kingman, *Subadditive ergodic theory*, Ann. Probab., 1 (1973), pp. 883–909.

[14] B. G. Pittel, *Limiting behaviour of a process of runs*, Ann. Probab., 9 (1981), pp. 119–129.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0011 $01.25/0

ENTROPY VERSUS SPEED IN ERGODIC MARKOV MAPS*

NATHAN FRIEDMAN† AND ABRAHAM BOYARSKY‡

Abstract. Let $f=(f_1, f_2, \cdots, f_n)$ be a sequence of positive numbers. We construct a class of piecewise linear, ergodic Markov maps $\mathscr{C}_f$ such that for each $\tau \in \mathscr{C}_f$, there exist an interval I and a partition $\{I_i\}_{i=1}^n$ of I with the property that $\mu(I_i)=f_i$, where μ is the unique measure invariant under τ. In $\mathscr{C}_f$ the trade-off between entropy (randomness) and the speed of numerically computing the orbit $\{\tau^j(x)\}_{j\geqq 0}$ can be assessed.

AMS(MOS) subject classification (1980). Primary 28D20, secondary 26A18.

1. Introduction.

There are various numerical procedures for generating sequences of numbers which appear random. In this paper we are concerned with procedures of the following type: given a transformation $\tau: I \to I$, where I is an interval, we choose an initial value $x \in I$ and then iterate. Ideally, the resulting sequence of numbers $\{\tau^j(x)\}_{j\geqq 0}$ should appear random and should not take long to compute.

Let $(f_1, f_2, \cdots, f_n)$ be a sequence of positive numbers, such that $\sum_{i=1}^n f_i = 1$. Denoting Lebesgue measure on I by λ, take consecutive intervals $I_1, I_2, \cdots, I_n$ such that $\lambda(I_i)=f_i$, $i=1,2,\cdots,n$, and set $I=\bigcup_{i=1}^n I_i$. Define the piecewise linear map $\bar{\tau}: I\to I$ by $\bar{\tau}(I_i)=I$ for all i. Then $\bar{\tau}$ is ergodic and λ itself is the unique absolutely continuous measure invariant under τ. τ is just the one-side Bernoulli shift based on the distribution $f=(f_1, f_2, \cdots, f_n)$. From the Birkhoff ergodic theorem it follows that $\{\bar{\tau}^j(x)\}_{j\geqq 0}$ "exhibits" the distribution f, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_{I_i}(\bar{\tau}^j(x)) \to f_i \quad \text{a.e. } \lambda.$$

We make the following heuristic remarks: (a) $\bar{\tau}$ has "relatively" large entropy (measure of randomness [1]), since each interval I_i is mapped onto all the other intervals with positive probability. (b) The orbit $\{\bar{\tau}^i(x)\}_{i\geqq 0}$ takes a "relatively" long time to be produced. This follows also from the fact that $\bar{\tau}(I_i)=I$ for all i, i.e., at each stage j the maximum number of computations (possibly as many as $1+\ln_2 n$) may have to be made in order to determine where $\bar{\tau}^j(x)$ is.

In view of (a) and (b) we see that for a specified distribution of probabilities $f=(f_1, f_2, \cdots, f_n)$, $\bar{\tau}=\bar{\tau}_f$ exhibits f very randomly since for any x, $\bar{\tau}(x)$ can range over all of I, but that the speed of computation.of the orbit $\{\bar{\tau}^j(x)\}_{j\geqq 0}$ is slow, since a possible maximum number of searches must be made at each stage of the iterative procedure to determine the subinterval containing $\bar{\tau}^j(x)$.

The purpose of this paper is to present a construction for a class of piecewise linear, ergodic Markov maps ($\bar{\tau}$ is in this class) such that all maps in this class exhibit a given, fixed distribution $f=(f_1, f_2, \cdots, f_n)$, but where the trade-off between entropy (randomness) and the speed of computation of the orbit can be assessed.

2. Preliminaries.

Let $\mathscr{B}$ denote the Lebesgue measurable subsets of I and let μ be a measure on $(I, \mathscr{B})$. μ is said to be invariant under the map $\tau: I\to I$ if, for all $A\in\mathscr{B}$, $\mu(A)=\mu(\tau^{-1}(A))$, where $\tau^{-1}(A)=\{x\in I: \tau(x)\in A\}$. μ is absolutely continuous

* Received by the editors November 16, 1981, and in revised form March 9, 1983.

† School of Computer Science, McGill University, Montreal, Quebec, Canada H3A 2K6. The research of this author was supported by the Natural Science and Engineering Research Council of Canada under grant A-4330.

‡ Department of Mathematics, Loyola Campus, Concordia University, Montreal, Quebec, Canada H4B 1R6. The research of this author was supported by the Natural Science and Engineering Research Council of Canada under grant A-9072.

if there exists an $f \in \mathscr{L}_1$, the space of integrable functions on I, $f(x) \geqq 0$, such that

$$\mu(A) = \int_A f(x)\,dx \quad \forall A \in \mathscr{B}.$$

We refer to f as a density invariant under τ. It is well known [2] that the densities invariant under τ (nonsingular) are the fixed points of the Frobenius–Perron operator P_τ: $\mathscr{L}_1 \rightarrow \mathscr{L}_1$, defined by

$$P_\tau f(x) = \frac{d}{dx} \int_{\tau^{-1}[0,x]} f(s)\,ds.$$

For τ piecewise C^2 and satisfying $\inf |\tau'(x)| > 1$, where the derivative exists, it is shown in [2] that τ admits an absolutely continuous invariant measure.

A piecewise continuous map τ: $I \rightarrow I$ is called Markov if there exist points in I: $a_0 < a_1 < \cdots < a_n$ such that for $j = 0, 1, \cdots, n-1$, $\tau|_{(a_j, a_{j+1})}$ is a homeomorphism onto some interval $(a_{k(j)}, a_{l(j)})$.

If τ is a piecewise linear, Markov map, then it is shown in [3], where there are unnecessary restrictions, that the Frobenius–Perron operator, when restricted to the space of piecewise constant functions on the partition $0 = a_0 < a_1 < \cdots < a_n = 1$ defined by τ, is a matrix $M = M_\tau = (m_{lj})$, where

$$(1) \qquad m_{lj} = |\tau_j'|^{-1} \delta_{lj},$$

τ_j' being the slope of τ on $I_j = (a_{j-1}, a_j)$, and $\delta_{lj} = 1$ if $I_l \subset \tau(I_j)$ and 0 otherwise. We shall refer to M as the matrix induced by τ, and to τ as the transformation associated with M. If the partition $\mathscr{I} = \{I_j\}_{j=1}^n$ is such that M is irreducible and not a permutation matrix, then $\mathscr{I}$ will be called a strong Markov partition. Then M has 1 as its eigenvalue of maximum modulus and the geometric and algebraic multiplicities of the eigenvalue 1 are also 1 [5]. Hence the system of equations $\pi M = \pi$ has a unique solution. Since all fixed points of P_τ are piecewise constant on $\mathscr{I}$ [4], it follows that π, the (left) fixed point of M, is the unique (up to constant multiples) density invariant under τ when viewed as a step function on the partition $\mathscr{I}$.

In order to compute the entropy of τ, we consider the Markov chain induced on $\mathscr{I}$ by τ. The *transition matrix* $T = (t_{lj})$ is defined by

$$t_{lj} \equiv \frac{\lambda(I_l \cap \tau^{-1}(I_j))}{\lambda(I_l)} = \frac{\lambda(I_j)}{\lambda(I_l)} m_{lj},$$

where t_{lj} denotes the proportion of the interval I_l which is mapped onto I_j. The left eigenvector of T is the distribution f. We remark that the $n \times n$ matrix T yields f while the $n \times n$ matrix M yields the step function $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$, i.e., the density invariant under τ. Thus $f_i = \pi_i \lambda(I_i)$. It is possible to employ either T or M to construct the desired transformations. We have chosen to work with the M matrices since it appears easier to concatenate transformations with them (see § 4). However, in the definition of the entropy of τ, T is the natural choice.

Let $\tau \in \mathscr{C}$, the space of piecewise linear ergodic Markov maps on I, and let $\mathscr{I} = \{I_i\}_{i=1}^n$ denote the partition with respect to which τ is Markov. Let T be the transition matrix induced by τ and f the resulting distribution on $\mathscr{I}$. Then the *entropy* of τ is given by the expression [1, p. 91].

$$H(\tau) = -\sum_{\substack{l=1\\j=1}}^{n} f_l t_{lj} \ln t_{lj}.$$

Before proceeding let us give examples of two different transformations in $\mathscr{C}$ which exhibit the same distribution f.

Example 1. τ_1: $[\frac{1}{4}, 1] \to [\frac{1}{4}, 1]$ is defined by

$$\tau_1(x) = \begin{cases} 2x, & \frac{1}{4} \leqq x \leqq \frac{1}{2}, \\ -\frac{3}{2}x + \frac{7}{4}, & \frac{1}{2} \leqq x \leqq 1. \end{cases}$$

Clearly $\tau_1 \in \mathscr{C}$, and $\mathscr{I}_1 = \{(\frac{1}{4}, \frac{1}{2}), (\frac{1}{2}, 1)\}$. It is easy to see that

$$M_1 = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{2}{3} & \frac{2}{3} \end{pmatrix}, \qquad T_1 = \begin{pmatrix} 0 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

and $\pi = (1, \frac{3}{2})$, i.e.,

$$\pi(x) = \begin{cases} 1, & \frac{1}{4} \leqq x \leqq \frac{1}{2}, \\ \frac{3}{2}, & \frac{1}{2} < x \leqq 1. \end{cases}$$

It follows from this, or by directly computing the left fixed point of T, that $f = (\frac{1}{4}, \frac{3}{4})$. Hence, with the convention $0 \log 0 = 0$, we get

$$H(\tau_1) = -\tfrac{3}{4}(\tfrac{1}{3} \ln \tfrac{1}{3} + \tfrac{2}{3} \ln \tfrac{2}{3}) = .477.$$

Example 2. Consider now $\tau_2 \in \mathscr{C}$ defined on $[0, 1]$ by $\tau_2(0) = 0$, $\tau_2(\frac{1}{4}) = 1$ and $\tau_2(1) = 0$. Then, with respect to $\mathscr{I}_2 = \{(0, \frac{1}{4}), (\frac{1}{4}, 1)\}$,

$$T_2 = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

and $fT_2 = f$, where $f = (\frac{1}{4}, \frac{3}{4})$. The entropy of τ_2 is

$$H(\tau_2) = -(\tfrac{1}{4} \ln \tfrac{1}{4} + \tfrac{3}{4} \ln \tfrac{3}{4}) = .562.$$

That $H(\tau_2)$ is greater than $H(\tau_1)$ is intuitively clear from the fact that the interval $(\frac{1}{4}, \frac{1}{2}) \in \mathscr{I}_1$ maps onto only one interval, namely $(\frac{1}{2}, 1)$, whereas $(0, \frac{1}{4}) \in \mathscr{I}_2$ maps onto two intervals.

However, as for speed of computation, the orbit $\{\tau_1^j(x)\}_{j \geqq 0}$ can be generated more quickly than $\{\tau_2^j(x)\}_{j \geqq 0}$ since once $\tau_1^j(x) \in (\frac{1}{4}, \frac{1}{2})$, it follows that $\tau_1^{j+1}(x) \in (\frac{1}{2}, 1)$, i.e., no search is required at one of the intervals. This, however, is not the case with τ_2, where $\tau(0, 1) = (0, 1)$ and $\tau(\frac{1}{4}, 1) = (0, 1)$.

3. Two special Markov maps. Let $a = x_0 < x_1 < \cdots < x_n = b$ be any partition of $J = [a, b]$. Let $J_i = (x_{i-1}, x_i)$ and define the piecewise linear, continuous map $\tau: J \to J$ by the conditions $\tau(x_{i-1}) = x_{i-1}$, $1 \leqq i < n$ and $\tau(x_n) = x_0$. Then:

(i) $\tau(J_i) = J_{i+1}$, for $1 \leqq i < n$,

(ii) $\tau(J_n) = \bigcup_{i=1}^n J_i$.

τ is clearly a Markov map. Hence its Frobenius–Perron operator P_τ, when restricted to piecewise constant functions on the partition, is represented by a matrix $M = M_\tau$. Let $\mathscr{C}$ denote the class of Markov maps satisfying (i) and (ii).

LEMMA 1. *Let $\tau \in \mathscr{C}$. Then it is ergodic, and its unique invariant density $\pi(x)$, which is piecewise constant, is given by*

$$\pi_i(x) = \pi(x)|_{J_i} = \frac{x_i - a}{x_i - x_{i-1}}.$$

Proof. By (1), the $n \times n$ matrix M is given by

$$m_{i,j} = \begin{cases} \dfrac{x_i - x_{i-1}}{x_{i+1} - x_i}, & j = i+1, \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leqq i < n$ and

$$m_{n,j} = \frac{x_n - x_{n-1}}{b-a}, \qquad 1 \leqq j \leqq n.$$

The matrix M has nonzero entries in the superdiagonal and in the nth row. Hence, it follows from [4, Corollary 2.1] that M is primitive, and therefore τ is ergodic. It remains only to show that $\pi M = \pi$. From the form of M, we get

$$\sum_{r=1}^{N} \pi_r m_{r,s} = \begin{cases} \pi_{s-1} m_{s-1,s} + \pi_n m_{n,s}, & s \neq 1, \\ \pi_n m_{n,s}, & s = 1. \end{cases}$$

For $s = 1$,

$$\pi_n m_{n,s} = \frac{x_n - a}{x_n - x_{n-1}} \cdot \frac{x_n - x_{n-1}}{x_n - a} = 1 = \pi_1.$$

For $s \neq 1$,

$$\pi_{s-1} m_{s-1,s} + \pi_n m_{n,s} = \frac{x_{s-1} - a}{x_{s-1} - x_{s-2}} \cdot \frac{x_{s-1} - x_{s-2}}{x_s - x_{s-1}} + 1 = \frac{x_s - a}{x_s - x_{s-1}} = \pi_s.$$

Hence $\pi M = \pi$. Since the support of $\pi(x)$ is all of J, it is the unique (up to constant multiples) invariant density under τ. Q.E.D.

Note that $f_i = x_i - a$, which increases with i.

Example 3. Let $x_i = i$, $0 \leqq i \leqq n$. Then $\pi_i = i$ and $\pi = (1, 2, \cdots, n)$. The induced matrix M is

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & & & \\ 0 & 0 & 0 & 1 & 0 & & & \\ & & & & \ddots & & & \\ & & & & & \ddots & & \\ 0 & 0 & 0 & 0 & & \cdots & 0 & 1 \\ \dfrac{1}{n} & \dfrac{1}{n} & \dfrac{1}{n} & & & \cdots & & \dfrac{1}{n} \end{bmatrix}.$$

If $x_i = g + ih$, $0 \leqq i \leqq n$, then the same vector $\pi = (1, 2, \cdots, n)$ is obtained.

Example 4. Let $x_i = c^i$, $0 \leqq i \leqq n$, $c > 1$. Then

$$\pi_i = \frac{c^i - 1}{c^i - c^{i-1}} = \sum_{j=0}^{i-1} c^{-j}.$$

For $c = 2$, $\pi_i = (2^i - 1)/2^{i-1}$, which is a slowly rising step function.

Example 5. Let $x_i = c^{n-i}$, $0 \leqq i \leqq n$, $c < 1$. Then

$$\pi_i = \frac{c^{n-i} - c^n}{c^{n-i} - c^{n-i-1}} = \frac{(1/c)^i - 1}{(i/c)^i - (1/c)^{i-1}} = \sum_{j=0}^{i-1} c^j.$$

For $c=\frac{1}{2}$, we obtain $\pi_i=2^i-1$, which is a very rapidly rising step function. For example, the first 9 values of this step function are (1, 3, 7, 15, 31, 63, 127, 255, 1023).

Let us now define a class of Markov maps, $\bar{\mathscr{C}}$, similar to the class $\mathscr{C}$, satisfying the conditions:

(i') $\tau(J_i)=J_{i-1}$, $1<i\leqq n$,

(ii') $\tau(J_1)=\bigcup_{i=1}^n J_i$.

LEMMA 2. *Let $\tau\in\bar{\mathscr{C}}$. Then it is ergodic, and the unique invariant density $\pi(x)$, which is piecewise constant is given by*

$$\pi_i(x)=\pi(x)|_{J_i}=\frac{b-x_{i-1}}{x_i-x_{i-1}}.$$

Proof. By (1), the $n\times n$ matrix M is given by

$$m_{1,j}=\frac{x_1-x_0}{b-a},\qquad 1\leqq j\leqq n,$$

and for $1<i\leqq n$,

$$m_{i,j}=\begin{cases}\dfrac{x_1-x_{i-1}}{x_i-x_{i-1}}, & j=i-1,\\ 0, & \text{otherwise.}\end{cases}$$

The remainder of the proof is similar to that of Lemma 1. Q.E.D.

Note that $f_i=b-x_{i-1}$, which decreases as i increases.

Remark. Suppose M is an $n\times n$ matrix associated with some $\tau\in\mathscr{C}$ and let π satisfy $\pi M=\pi$. Let P be the $n\times n$ skew diagonal matrix

$$\begin{pmatrix} & 0 & & & 1\\ & & & 1 & \\ & & \cdot^{\cdot^{\cdot}} & & \\ & 1 & & 0 & \\ 1 & & & & \end{pmatrix},$$

and let $B=PMP^T$. It is interesting to note that B is the Frobenius–Perron operator of a transformation $\tau\in\bar{\mathscr{C}}$ whose unique (up to constant multiples) invariant density function is obtained by reversing the entries of π. While the nonzero entries of M and B have the same values (in different positions), the transformation τ^* associated with B is defined on a different partition than τ. However, the partition on which τ^* is defined, say $y_0<y_1<\cdots<y_n$, must satisfy:

$$y_i-y_{i-1}=x_{n-i+1}-x_{n-i},\qquad 1\leqq i\leqq n.$$

This result can be viewed as a generalization of [4, Prop. 4.3] for this class of Markov maps, since in [4] all intervals are assumed to be of equal length.

4. Construction of ergodic Markov maps. Suppose the Markov maps $\tau^{(i)}\in\mathscr{C}\cup\bar{\mathscr{C}}$, $1\leqq i\leqq k$, where $\tau^{(i)}\colon I_i\to I_i$ each admit the invariant density $\pi^{(i)}=(\pi_1^{(i)},\pi_2^{(i)},\cdots,\pi_{n_i}^{(i)})$, as defined in Lemmas 1 and 2. Since the location of I_i on the real line is irrelevant, we can arrange the intervals $(I_1,I_2,\cdots,I_n)$ consecutively so that $I=\bigcup_{i=1}^n I_i$ is an interval. Now define $\tau\colon I\to I$ by letting $\tau|_{I_i}=\tau^{(i)}$. τ is well defined, since $\tau^{(i)}\colon I_i\to I_i$. Let $\pi=(\pi^{(1)},\pi^{(2)},\cdots,\pi^{(k)})$. Clearly π is a density function invariant under τ, but it is not ergodic. To see this, let μ be the measure induced by π, i.e.,

$$\mu(A)=\int_A\pi(x)\,dx.$$

Then $\mu(\tau^{-1}(I_i))=\mu(I_i)$ for any i, but $0<\mu(I_i)<1$. We next demonstrate the construction of a piecewise linear Markov map $\tau:I\to I$, such that π is an invariant density under τ and the measure μ is ergodic.

THEOREM 1. *Let $\tau^{(i)}$, $1\leqq i\leqq k$, be a collection of Markov maps, $\tau^{(i)}\in\mathscr{C}\cup\bar{\mathscr{C}}$, such that $\tau^{(i)}$ admits the invariant density $\pi^{(i)}=(\pi_1^{(i)},\pi_2^{(i)},\cdots,\pi_{n_i}^{(i)})$ defined in Lemmas* 1 *and* 2. *Then there exists an ergodic Markov map τ which admits the invariant density $\pi=(\pi^{(1)},\pi^{(2)},\cdots,\pi^{(k)})$.*

Proof. Let $x_0^{(i)}<x_1^{(i)}<\cdots<x_{n_i}^{(i)}$ be the partition used in defining $\tau^{(i)}$, and let $A^{(i)}$ be the $n_i\times n_i$ matrix induced by $\tau^{(i)}$. Let $\tau^{(i_1)},\cdots,\tau^{(i_h)}$ be the subsequence of $\tau^{(i)}$'s in $\mathscr{C}$ and $\tau^{(j_1)},\cdots,\tau^{(j_l)}$ those in $\bar{\mathscr{C}}$.

From Lemma 1 it follows that if $\tau^{(i)}\in\mathscr{C}$, then

$$\pi_{n_i}^{(i)}=\frac{x_{n_i}^{(i)}-x_0^{(i)}}{x_{n_i}^{(i)}-x_{n_i-1}^{(i)}},$$

and the last row of the induced matrix $M^{(i)}$ has all its entries equal to $1/\pi_{n_i}^{(i)}$. If $\tau^{(i)}\in\bar{\mathscr{C}}$, then by Lemma 2,

$$\pi_1^{(i)}=\frac{x_{n_i}^{(i)}-x^{(i)}}{x_1^{(i)}-x_0^{(i)}},$$

and the first row of $M^{(i)}$ is $1/\pi_1^{(i)}$.

Let $b_1=0$, and $b_i=\sum_{j=1}^{i-1}n_j$, $1<i\leqq k+1$. Let $n=b_{k+1}$. Now, define the partition $x_0<x_1<\cdots<x_n$ by

$$x_0=x_0^{(1)},$$

$$x_{b_i+j}=x_{b_i}+x_j^{(i)}-x_0^{(i)},\qquad 1\leqq i\leqq k,\quad 1\leqq j\leqq n_i.$$

Let $I_j^{(i)}=(x_{j-1}^{(i)},x_j^{(i)})$ and $I_j=(x_{j-1},x_j)$. We now define $\tau\colon[x_0,x_n]\to[x_0,x_n]$ as follows:

$$\text{if } \tau^{(i)}(I_j^{(i)})=I_k^{(i)},\qquad \text{then define } \tau(I_{b_i+j})=I_{b_i+k}$$

$$\text{if } \tau^{(i)}(I_j^{(i)})=\bigcup_{k=1}^{n_i}I_k^{(i)},\qquad \text{then define } \tau(I_{b_i+j})=\bigcup_{k=1}^{n}I_k.$$

Clearly τ is Markov. The matrix induced by τ is

$$M=\begin{pmatrix} M^{11} & M^{12} & \cdots & M^{1k}\\ M^{21} & M^{22} & \cdots & M^{2k}\\ \vdots & & & \vdots\\ M^{k1} & M^{k2} & & M^{kk}\end{pmatrix},$$

where $M^{ij}=(m_{r,s}^{(ij)})$ is an $n_i\times n_j$ matrix defined as follows:

(i) If $\tau^{(i)}\in\mathscr{C}$,

$$m_{r,s}^{(ii)}=m_{r,s}^{(i)}\qquad \forall r\neq n_i,$$

$$m_{n_i,s}^{(ii)}=\frac{x_{n_i}^{(i)}-x_{n_i-1}^{(i)}}{x_n-x_0}\qquad \forall 1\leqq s\leqq n_i,$$

and $\forall i\neq j$,

$$m_{r,s}^{(ij)}=0\qquad \forall r\neq n_i,$$

$$m_{n_i,s}^{(ij)}=m_{n_i,1}^{(ii)}\qquad \forall 1\leqq s\leqq n_j.$$

(ii) If $\tau^{(i)} \in \bar{\mathscr{E}}$,

$$m_{r,s}^{(ii)} = m_{r,s}^{(i)} \qquad \forall r \neq 1,$$

$$m_{1,s}^{(ii)} = \frac{x_1^{(i)} - x_0^{(i)}}{x_n - x_0} \qquad \forall 1 \leqq s \leqq n_i,$$

and $\forall i \neq j$,

$$m_{r,s}^{(ij)} = 0 \qquad \forall r \neq 1,$$

$$m_{1,s}^{(ij)} = m_{1,1}^{(ii)} \qquad \forall 1 \leqq s \leqq n_i.$$

The primitivity of M and hence the ergodicity of τ follows from [4, § 4]. It remains only to show that $\pi M = \pi$.

Suppose $b_i < s \leqq b_{i+1}$, and let $s' = s - b_i$. Then

$$\sum_{r=1}^{n} \pi_r m_{r,s} = \sum_{j=1}^{k} \sum_{r=1}^{n_j} \pi_r^{(j)} m_{r,s'}^{(ji)} = \sum_{r=1}^{n_i} {}^{\cdot}\pi_r^{(i)} m_{r,s}^{(ii)} + \sum_{\substack{t=1 \\ i_t \neq s'}}^{h} \pi_{n_{i_t}}^{(i_t)} m_{n_{i_t},s'}^{(i_t i)} + \sum_{\substack{t=1 \\ j_t \neq s'}}^{l} \pi_1^{(j_t)} m_{1,s'}^{(j_t i)}.$$

Let $h = n_i$ if $\tau^{(i)} \in \mathscr{E}$ and $h = 1$ if $\tau^{(i)} \in \bar{\mathscr{E}}$. Then

$$\begin{aligned} \sum_{r=1}^{n_i} \pi_r^{(i)} m_{r,s'}^{(ii)} &= \sum_{\substack{r=1 \\ r \neq m}}^{n_i} \pi_r^{(i)} m_{r,s'}^{(i)} + \pi_h^{(i)} m_{h,s'}^{(ii)} \\ &= \sum_{r=1}^{n_i} \pi_r^{(i)} m_{rs'}^{(i)} - \pi_h^{(i)} m_{h,s}^{(i)} + \pi_h^{(i)} m_{h,s'}^{(ii)} \\ &= \pi_{s'}^{(i)} - 1 + \pi_h^{(i)} m_{h,s}^{(ii)}. \end{aligned}$$

Therefore, it suffices to prove that

$$\sum_{t=1}^{h} \pi_{n_{i_t}}^{(i_t)} m_{n_{i_t},1}^{(i_t i)} + \sum_{t=1}^{l} \pi_1^{(j_t)} m_{1,1}^{(j_t 1)} = 1, \tag{2}$$

since $\pi_{n_i}^{(i)} m_{n_i,1}^{(i)} = 1$ if $\tau^{(i)} \in \mathscr{E}$ and $\pi_1^{(i)} m_{1,1}^{(i)} = 1$ if $\tau^{(i)} \in \bar{\mathscr{E}}$. Now,

$$\pi_{n_{i_t}}^{(i_t)} m_{n_{i_t},1}^{(i_t,i)} = \frac{x_{n_{i_t}}^{(i_t)} - x_0^{(i_t)}}{x_{n_{i_t}}^{(i_t)} - x_{n_{i_t-1}}^{(i_t)}} \cdot \frac{x_{n_{i_t}}^{(i_t)} - x_{n_{i_t}-1}^{(i_t)}}{x_n - x_0} = \frac{x_{b_{i_t+1}} - x_{b_{i_t}}}{x_n - x_0}.$$

Similarly,

$$\pi_1^{(j_t)} m_{1,1}^{(j_t 1)} = \frac{x_{b_{j_t+1}} - x_{b_{j_t}}}{x_n - x_0}.$$

The sum in (2) then equals

$$\frac{1}{x_n - x_0} \left[\sum_{t=1}^{m} (x_{b_{i_t+1}} - x_{b_{i_t}}) + \sum_{t=1}^{l} (x_{b_{j_t+1}} - x_{b_{j_t}}) \right] = \frac{1}{x_n - x_0} \sum_{j=1}^{k} (x_{b_{j+1}} - x_{b_j}), \tag{3}$$

since $\{i, \cdots, i_m\} \cup \{j_1, \cdots, j_l\} = \{1, 2, \cdots, k\}$. Finally, the sum in (3) telescopes to yield $x_n - x_0$. Hence the desired sum is equal to 1.

Example 6. Let J_1, J_2, J_3, J_4, J_5 be an equal partition of $[0, 5]$ and let the distribution f be given by $(\frac{1}{4}, \frac{3}{4}, \frac{7}{8}, \frac{3}{8}, \frac{1}{8})$, as shown in Fig. 1, where without loss of generality, we assume the length of each subinterval is 1. We want to construct an ergodic Markov map, whose invariant density has the distribution f.

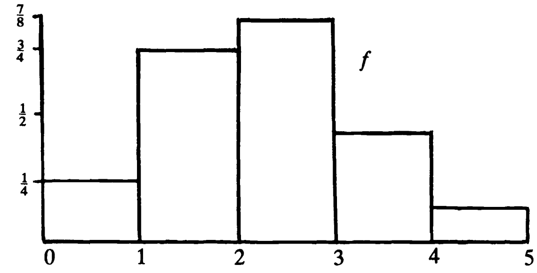


FIG. 1

First, we divide f into two parts, an increasing part $(\frac{1}{4}, \frac{3}{4})$ and a decreasing part $(\frac{7}{8}, \frac{3}{8}, \frac{1}{8})$. We now construct two Markov maps, which have, respectively, these distributions. As in the proof of Theorem 1, we define the set of numbers $x_0^{(1)}, x_1^{(1)}, x_2^{(1)}$ to satisfy:

$$x_1^{(1)}-x_0^{(1)}=\tfrac{1}{4} \quad \text{and} \quad x_2^{(1)}-x_1^{(1)}=\tfrac{3}{4}.$$

One such choice is $x_0^{(1)}=\frac{1}{4}, x_1^{(1)}=\frac{1}{2}, x_2^{(1)}=1$. From Lemma 1, we know that $\tau^{(1)}: [\frac{1}{4}, 1] \to [\frac{1}{4}, 1]$ satisfies

$$\tau^{(1)}(\tfrac{1}{4})=\tfrac{1}{2}, \qquad \tau^{(1)}(\tfrac{1}{2})=1$$

and $\tau^{(1)}(1)=\frac{1}{4}$, as shown in Fig. 2.

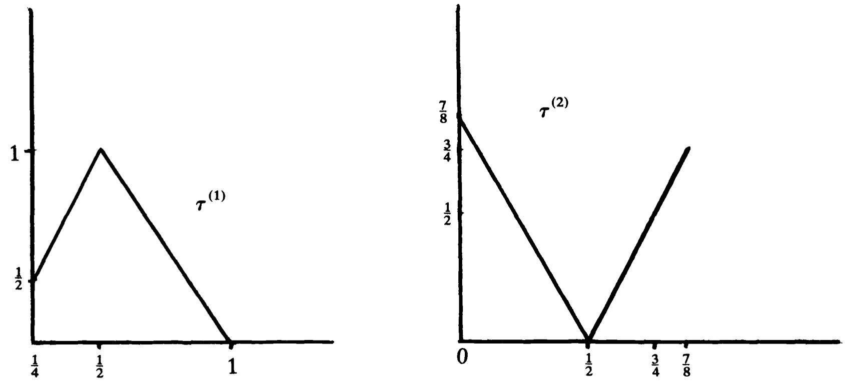


FIG. 2

Analogously, choose $x_0^{(2)}, x_1^{(2)}, x_2^{(2)}, x_3^{(2)}$ to satisfy

$$x_3^{(2)}-x_0^{(2)}=\tfrac{7}{8}, \quad x_3^{(2)}-x_1^{(2)}=\tfrac{3}{8}, \quad x_3^{(2)}-x_2^{(2)}=\tfrac{1}{8}.$$

One possible choice is $x_0^{(2)}=0, x_1^{(2)}=\frac{1}{2}, x_2^{(2)}=\frac{3}{4}, x_3^{(2)}=\frac{7}{8}$. By Lemma 2, we know that $\tau^{(2)}: [0, \frac{7}{8}] \to [0, \frac{7}{8}]$ satisfies

$$\tau^{(2)}(0)=\tfrac{7}{8}, \quad \tau^{(2)}(\tfrac{1}{2})=0, \quad \tau^{(2)}(\tfrac{3}{4})=\tfrac{1}{2}, \quad \tau^{(2)}(\tfrac{7}{8})=\tfrac{3}{4},$$

as shown in Fig. 2.

Furthermore, the matrices induced by $\tau^{(1)}$ and $\tau^{(2)}$ are, respectively,

$$M^{(1)}=\begin{pmatrix} 0 & \frac{1}{2} \\ \frac{2}{3} & \frac{2}{3} \end{pmatrix}, \qquad M^{(2)}=\begin{pmatrix} \frac{4}{7} & \frac{4}{7} & \frac{4}{7} \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix},$$

and

$$\pi^{(1)}=(1,\tfrac{3}{2}),\qquad \pi^{(2)}=(\tfrac{7}{4},\tfrac{3}{2},1).$$

Now let us construct the new partition as in the proof of Theorem 1 namely,

$$\tfrac{1}{4}<\tfrac{1}{2}<1<\tfrac{3}{2}<\tfrac{7}{4}<\tfrac{15}{8}. \tag{4}$$

The concatenated Markov map τ: $[\frac{1}{4},\frac{15}{8}]\to[\frac{1}{4},\frac{15}{8}]$ defined on this partition, as given in the proof of Theorem 1, is shown in Fig. 3. The map τ induces the matrix

$$A=\begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0\\ \frac{4}{13} & \frac{4}{13} & \frac{4}{13} & \frac{4}{13} & \frac{4}{13}\\ \frac{4}{13} & \frac{4}{13} & \frac{4}{13} & \frac{4}{13} & \frac{4}{13}\\ 0 & 0 & \frac{1}{2} & 0 & 0\\ 0 & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

One can readily verify that $(1,\frac{3}{2},\frac{7}{4},\frac{3}{2},1)M=(1,\frac{3}{2},\frac{7}{4},\frac{3}{2},1)$. Using the interval lengths of the partition (4), we obtain the desired distribution $(\frac{1}{4},\frac{3}{4},\frac{7}{8},\frac{3}{8},\frac{1}{8})$.

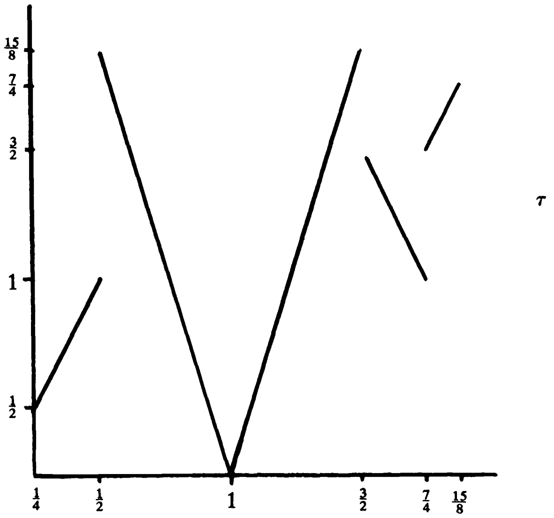


FIG. 3

5. Entropy. Let $f=(f_1,f_2,\cdots,f_n)$ be a sequence of positive numbers whose sum is one. It is possible to partition f into k strictly monotone (increasing or decreasing) subsequences $f=(f^{(1)},\cdots,f^{(k)})$, where $f^{(i)}=(f_1^{(i)},\cdots,f_{n_k}^{(i)})$.

In § 4 we presented a construction for a piecewise linear, ergodic Markov map τ, whose unique invariant measure μ satisfies $\mu(I_i)=f_i$, $i=1,2,\cdots,n$, where $I_1,I_2,\cdots,I_n$ are the consecutive intervals on which τ is defined. Let $I=\bigcup_{i=1}^{n}I_i$ and $I^{(i)}$ be the subsequence corresponding to $f^{(i)}$. Furthermore, τ induces an $n\times n$ matrix M whose left fixed point is the density of μ, i.e., the step function $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$, where $f_i=\pi_i\lambda(I_i)$, $i=1,2,\cdots,n$.

Now we can write M as the block matrix

$$\begin{pmatrix} M^{(1)}\\ \vdots\\ M^{(k)} \end{pmatrix},$$

where $M^{(i)}$ is an $n_i \times n$ matrix. If $f^{(i)}$ is increasing the last row of $M^{(i)}$ has all entries equal to $\lambda(I_{n_i}^{(i)})/\lambda(I)$ and each of the other rows has a single nonzero entry. If $f^{(i)}$ is decreasing, the first row of $M^{(i)}$ has entries $\lambda(I_1^{(i)})/\lambda(I)$ and all other rows have one nonzero entry. To simplify matters, let,

$$n_i^* = \begin{cases} n_i & \text{if } f^{(i)} \text{ is increasing,} \\ 1 & \text{if } f^{(i)} \text{ is decreasing.} \end{cases}$$

Then row n_i^* of $M^{(i)}$ has entries $\lambda(I_{n_i^*}^{(i)})/\lambda(I)$. The transition matrix T induced by τ can be partitioned as

$$\begin{pmatrix} T^{(1)} \\ \vdots \\ T^{(n)} \end{pmatrix},$$

where $T^{(i)}$ is an $n \times n$ matrix with entries

$$t_{lj}^{(i)} = \frac{\lambda(I_j)}{\lambda(I_l^{(i)})} m_{lj}^{(i)}.$$

It is not difficult to verify that all entries of $T^{(i)}$ other than those in row n_i^* are 0 (corresponding to the entries of $M^{(i)}$). Since $\ln 1 = 0$ and $0 \ln 0$ is defined as 0, these 0 and 1 terms do not contribute to the entropy of T. Moreover, we have

$$t_{n_i^*,j} = \frac{\lambda(I_j)}{\lambda(I_{n_i^*}^{(i)})} \cdot \frac{\lambda(I_{n_i^*}^{(i)})}{\lambda(I)} = \frac{\lambda(I_j)}{\lambda(I)}.$$

Hence,

$$H(\tau) = -\sum_{j=1}^{n} \sum_{i=1}^{k} f_{n_i^*} \frac{\lambda(I_j)}{\lambda(I)} \ln\left(\frac{\lambda(I_j)}{\lambda(I)}\right). \tag{5}$$

Note that if $f = (c, c, \cdots, c)$, then $\lambda(I_j) = c$, $j = 1, 2, \cdots, n$, and $\lambda(I) = nc$. Hence

$$H(\tau) = -\sum_{j=1}^{n} \sum_{i=1}^{n} \frac{c}{n} \ln \frac{1}{n} = cn \ln n.$$

The given sequence of numbers f can usually be decomposed into monotone sequences in many ways. Suppose we refine the previous decomposition by splitting one of the monotone subsequences (of length at least 2), say $f^{(p)}$, into two segments

$$f^{(p_1)} = (f_1^{(p)}, \cdots, f_m^{(p)}), \qquad f^{(p_2)} = (f_{m+1}^{(p)}, \cdots, f_{n_p}^{(p)}).$$

This new decomposition results in a new piecewise linear, ergodic Markov map τ', defined on the n intervals $I'_1, I'_2, \cdots, I'_n$. If $f^{(p)}$ is an increasing subsequence, the new partition has the property

$$\lambda(I'_i) = \lambda(I_i), \qquad 1 \leqq i \leqq n,$$

and

$$\lambda(I'_{m+1}) = f_{m+1}. \tag{6}$$

Note that in this case $\lambda(I'_{m+1}) = \lambda(I_m) + f_m$ and thus

$$\lambda(I') = (I) + f_m. \tag{7}$$

If $f^{(p)}$ is a decreasing subsequence, it is easy to verify that the new partition $I'_1, I'_2, \cdots, I'_n$ satisfies $\lambda(I') = \lambda(I) + f_{m+1}$.

We observe that the map τ' differs from τ in that it has one extra interval on which its range is the entire domain. One would therefore expect that $H(\tau')>H(\tau)$. We now prove this in the case $f^{(p)}$ is increasing. Similar results establish the result for $f^{(p)}$ decreasing.

THEOREM 2. *Let $f=(f_1, f_2, \cdots, f_n)$ be a sequence of positive numbers whose sum is 1 and let τ be a piecewise linear, ergodic, Markov map admitting the invariant measure μ, i.e., $\mu(I_i)=f_i$, where $\{I_i\}_{i=1}^n$ are the consecutive intervals on which τ is defined. Let us split one of the monotone subsequences in f and define the new Markov map τ' as above. Then $H(\tau')>H(\tau)$, although both τ and τ' admit the same invariant measure μ.*

Proof. From (5), we have

$$\begin{aligned}H(\tau)&=-\sum_{j=1}^{n}\sum_{i=1}^{k} f_{n_i^*}\frac{\lambda(I_j)}{\lambda(I)}\ln(\lambda(I_j))+\sum_{j=1}^{n}\sum_{i=1}^{k} f_{n_i^*}\frac{\lambda(I_j)}{\lambda(I)}\ln(\lambda(I))\\&=\sum_{i=1}^{k}\frac{f_{n_i^*}}{\lambda(I)}\ln(\lambda(I))\left(\sum_{j=1}^{n}\lambda(I_j)\right)-\sum_{i=1}^{k}\frac{f_{n_i^*}}{\lambda(I)}\sum_{j=1}^{n}\lambda(I_j)\ln(\lambda(I_j))\\&=\sum_{i=1}^{k} f_{n_i^*}\ln(\lambda(I))-\sum_{i=1}^{k}\frac{f_{n_i^*}}{\lambda(I)}\ln\left(\prod_{j=1}^{n}\lambda(I_j)^{\lambda(I_j)}\right).\end{aligned}$$

Let $F=\sum_{i=1}^{k} f_{n_i^*}$ and $A=\lambda(I)$. Thus

$$H(\tau)=F\left[\ln A-\frac{1}{A}\ln\left(\prod_{j=1}^{n}\lambda(I_j)^{\lambda(I_j)}\right)\right]=F\ln\left(\frac{A}{\prod_{j=1}^{n}\lambda(I_j)^{\lambda(I_j)/A}}\right).$$

Using (6) and (7) we get

$$\begin{aligned}H(\tau')&=(F+f_m)\ln\left(\frac{A+f_m}{\prod_{j=1}^{n}\lambda(I_j')^{\lambda(I_j')/(A+f_m)}}\right)\\&=\frac{(F+f_m)}{(A+f_m)}\ln\left(\frac{(A+f_m)^{A+f_m}}{(\prod_{j=1, j\neq m+1}^{n}\lambda(I_j)^{\lambda(I_j)}\cdot(\lambda(I_{m+1})+f_m)^{\lambda(I_{m+1})+f_m})}\right).\end{aligned}$$

From the construction of τ in § 4, we have $F=A$. Hence

$$H(\tau')=\ln\left(\frac{(A+f_m)^{A+f_m}}{B(\lambda(I_{m+1})+f_m)^{\lambda(I_{m+1})+f_m}}\right),$$

where $B=\prod_{j=1, j\neq m+1}^{n}\lambda(I_j)^{\lambda(i_j)}$. Let

$$G(f_m)=\frac{(A+f_m)^{A+f_m}}{B(\lambda(I_{m+1})+f_m)^{\lambda(I_{m+1})+f_m}}.$$

Since $H(\tau)=\ln G(0)$ and $H(\tau')=\ln G(f_m)$, to prove $H(\tau')>H(\tau)$, it suffices to show that $G(f_m)$ is an increasing function of f_m. This follows immediately from Lemma 3 in Appendix 1. Q.E.D.

Appendix 1.

LEMMA 3. *Let $a>0$, $B>0$, $c>0$ and $A>c$. Then*

$$G(a)=\frac{(A+a)^{A+a}}{B(c+a)^{c+a}}$$

is an increasing function of a on $[0, \infty)$.

Proof. Let $N(a)$, $D(a)$ denote the numerator and denominator of $G(a)$, respectively. Then

$$N'(a) = (A+a)^{A+a}(1+\ln(A+a)), \qquad D'(a) = B(a+c)^{c+a}(1+\ln(c+a)).$$

Hence

$$G'(a) = \frac{B(c+a)^{c+a}[(A+a)^{A+a}(1+\ln(A+a))] - (A+a)^{A+a}[B(c+a)^{c+a}(1+\ln(c+a))]}{D^2(a)}$$

$$= \frac{B}{D^2(a)}(c+a)^{c+a}(A+a)^{A+c}[\ln(A+a) - \ln(c+a)] > 0. \qquad \text{Q.E.D.}$$

REFERENCES

[1] P. Billingsley, *Ergodic Theory and Information*, John Wiley, New York, 1965.

[2] A. Lasota and J. A. Yorke, *On the existence of invariant measures for piecewise monotonic transformations*, Trans. Amer. Math. Soc., 186 (1973), pp. 481–488.

[3] A. Boyarsky and M. Scarowsky, *A class of transformations which have unique absolutely continuous invariant measures*, Trans. Amer. Math. Soc., 255 (1979), pp. 243–262.

[4] N. Friedman and A. Boyarsky, *Construction of ergodic transformations*, Advances in Mathematics, 45 (1982), pp. 213–254.

[5] ———, *Matrices and eigenfunctions induced by Markov maps*, Linear Algebra and Appl., 38 (1981), pp. 141–147.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0012 $01.25/0

ON A DISCRETE SEARCH PROBLEM ON THREE ARCS*

F. A. BOSTOCK†

Abstract. A discrete search game with an immobile hider is posed and solved. The game is a discrete analogue of a known continuous game where the search takes place on three arcs which connect two points. A solution to the continuous game is obtained as a limiting case.

In [1, p. 33] S. Gal describes a zero sum search game on three arcs, each of unit length and connecting two points A and B. The hider chooses any point on one of the arcs, which we will simply refer to as arc 1, arc 2 and arc 3. The searcher starts at the point B and moves continuously until he finds the hider, who is immobile. The payoff to the hider is the distance covered by the searcher before the discovery. This game comes within a category of search games for which Gal shows there is always a solution, and he presents optimal strategies in the special circumstances that the searcher has his strategies limited to a certain set.

In this paper we solve a discrete version of this search game on three arcs, and although we also limit the type of pure strategy that the searcher can choose, it is analogously less of a limitation than that imposed by Gal in the continuous case. A solution of a corresponding continuous game is then seen as the limit of the solution in the discrete version.

To obtain the discrete game we may divide each of the arcs n $(n=1,2,3)$ into $k+1$ equal intervals of length $1/(k+1)$ by a sequence of k points $A_{n1}, A_{n2}, \cdots, A_{nk}$ so that for each $n=1,2,3$, A_{n1} is the point nearest to the point A. The hider may choose to hide only at one of the $3k$ points A_{ni}, $n=1,2,3$, $i=1,2,\cdots,k$ or at A or at B. Starting at B, the searcher moves in a sequence of steps, each time to an adjacent point, and for our purposes is restricted to those sequences which do visit each of the $3k+2$ points. The payoff to the hider, who is immobile, is the number of steps taken by the searcher until discovery. This game we call the discrete search game Δ_k on three arcs. It is well known that such a game, where one of the players has only a finite number of strategies, will have a solution. We obtain a subgame Γ_k by limiting the searcher to pure strategies of the following types.

Type 1. S_{mni}. Here $m, n=1,2,3$ with $m \neq n$ and $i=0,1,2,\cdots,k$. Let $p \in \{1,2,3\} \backslash \{m,n\}$. The searcher goes along arc m to the point A. If $i \neq 0$ he goes along arc n to the point A_{ni}, then returns to the point A, goes along arc p to B, and then finally along arc n to the point $A_{n,i+1}$. If $i=0$ he goes along arc p to B, and then finally along arc n to A.

Type 2. T_{min}. Here $m, n=1,2,3$ with $m \neq n$ and $i=1,2,\cdots,k$. Let $p \in \{1,2,3\} \backslash \{m,n\}$. The searcher goes along arc m to the point $A_{m,k-i+1}$, then returns to B, and goes along arc n to A. He now goes along arc m to the point $A_{m,k-i}$, then returns to A, and finally goes along arc p to B.

Type 3. U_{minjm} *and* U_{minjn}. Here $m, n=1,2,3$ with $m \neq n$ and $i, j=1,2,\cdots,k$. Let $p \in \{1,2,3\} \backslash \{m,n\}$. The searcher goes along arc m to the point $A_{m,k-i+1}$, then returns to B, goes along arc n to the point $A_{n,k-j+1}$, and then returns to B again. He now goes along arc p to A. In the case of U_{minjm} he then goes along arc m to the point $A_{m,k-i}$, returns to A and finally goes along arc n to the point $A_{n,k-j}$. In the case of U_{minjn} he goes along arc n to the point $A_{n,k-j}$, returns to A and finally goes along arc m to the point $A_{m,k-i}$.

* Received by the editors August 11, 1981, and in revised form January 20, 1983.

† University of Southampton, Faculty of Mathematical Studies, Southampton SO9 5NH, England.

The game Γ_k is essentially a matrix game and we denote the value by v_k. Of course the value of Δ_k is not greater than v_k and the guess is that they are equal. Certainly it seems clear that there exists a finite subset of the set of pure searcher strategies of Δ_k such that any strategy not in this subset will be chosen with zero probability in any optimal searcher strategy. Let us finally note here that in [1] for the continuous game, Gal limits consideration to the pure searcher strategies which correspond to those of Type 1 in our discrete version.

We now proceed to solve the game Γ_k. The method will be first to solve the subgame Γ'_k, whereby pure searcher strategies are limited to those of Type 1, and then show that this solution is applicable when Types 2 and 3 are included. Shortly we will give a theorem which essentially presents optimal strategies for the game Γ'_k. But first for the benefit of those readers who may not be entirely familiar with the theory of matrix games, we give a few items of general information which it is hoped will be helpful in giving a clearer understanding of both the statement and proof of our theorem.

Suppose $C=(c_{ij})$ is an $m \times n$ matrix with real number entries. Players 1 and 2 respectively choose a row and a column (each choice being unknown to the other), and then Player 2 pays to Player 1 an amount equal to the corresponding entry in the matrix. In repeated play suppose P_1 chooses row i with probability x_i and P_2 chooses column j with probability y_j; then the players are said to be using mixed strategies $X=(x_1, x_2, \cdots, x_m)$, $Y=(y_1, y_2, \cdots, y_n)$ and we have the corresponding expectation function $E(X, Y)=XCY^T$. When a particular player chooses a row or column with probability 1 he is said to be using a pure strategy and that strategy is often denoted simply by the number of that row or column. It is understood of course that P_1 will be trying to maximize $E(X, Y)$ and P_2 trying to minimize $E(X, Y)$. The fundamental theorem for matrix games [2, p. 37] tells us that for any matrix there exist a real number v and so called optimal strategies X^* for P_1, Y^* for P_2 such that for all X, Y, $E(X, Y^*) \leqq v \leqq E(X^*, Y)$. The X^* and Y^* are not necessarily unique but it is not difficult to show that v (called the value of the game) is unique, and further $v=E(X^*, Y^*)$. In the theorem below it will be observed that although certain strategies are shown to be optimal there is no indication as to how these strategies were discovered. The author feels that no real purpose would be served in giving a detailed description of the method he used in this case, but a few remarks might assist the reader who wishes to study the matter further. The technique used was based on a theorem [2, p. 67] which gives candidates for optimal strategies in terms of submatrices. Let it be said that the successful use of this method requires some guesswork, particularly regarding the choice of which pure strategies were to be played with zero probability.

It must also be stressed that the method only yields mixed strategies which may be optimal and must then be checked for optimality. Fortunately in this respect there is a simple test [2, p. 39] to verify that a given pair of strategies is optimal. This says that X^*, Y^* will be optimal and v will be the value if for all $i=1, 2, \cdots, m$, $j=1, 2, \cdots, n$

$$E(i, Y^*) \leqq v \leqq E(X^*, j).$$

Note that $E(X^*, j)$ and $E(i, Y^*)$ are merely the dot products of X^* and Y^* with the jth column and ith row vectors respectively. This is precisely the criterion we will use later in the proof of our theorem.

We return now to Γ'_k which we note is a $(1+3k)\times 6(k+1)$ matrix game, and introduce some notation. The mixed hider strategy which chooses each of A_{1i}, A_{2i}, A_{3i}

with probability $x_i/3$ $(i = 1, \cdots, k)$, A with probability x_0 and B with probability zero will be denoted by the $(3k+1)$-vector

$$X' = \left(x_0, \frac{x_1}{3}, \frac{x_1}{3}, \frac{x_1}{3}, \cdots, \frac{x_k}{3}, \frac{x_k}{3}, \frac{x_k}{3}\right).$$

The mixed searcher strategy which chooses each of S_{mni} $(m, n = 1, 2, 3, m \neq n)$ with probability $y_i/6$ $(i = 0, 1, \cdots, k)$ will be denoted by the $6(k+1)$-vector

$$Y' = \left(\frac{y_0}{6}, \cdots, \frac{y_0}{6}, \frac{y_1}{6}, \cdots, \frac{y_{k-1}}{6}, \frac{y_k}{6}, \cdots, \frac{y_k}{6}\right).$$

Although X' and Y' are not completely general mixed strategies for Γ'_k we will see soon that they are of a sufficiently general form to enable us to find optimal strategies for Γ'_k. We also let X'_p denote the mixed hider strategy obtained from X' by putting $x_p = 1$ and $x_i = 0$ for $i \neq p$. The mixed searcher strategy Y'_p is defined similarly. For reasons that will later be apparent we will let X, Y, X_p, Y_p be the corresponding $(k+1)$-vectors given by $X = (x_0, x_1, \cdots, x_k)$ and $Y = (y_0, y_1, \cdots, y_k)$. Finally we remark that in this context the symbol A_{mi} (resp. A) will denote the pure strategy for which the hider chooses the point A_{mi} (resp. A), and likewise the symbol S_{mnj} concerning the searcher.

Now by symmetry it is clear that for each $i, j = 0, 1, \cdots, k$

$$E'(X'_i, Y'_j) = \begin{cases} E'(A_{mi}, Y'_j) & \text{for } m = 1, 2, 3, \\ E'(X'_i, S_{mnj}) & \text{for } m, n = 1, 2, 3, \quad m \neq n, \end{cases} \tag{1}$$

where E' is the expectation function for Γ'_k, and one may note that $E'(X'_0, S_{mnj}) = E'(A, S_{mnj})$. This suggests that we define a $(1+k)\times(1+k)$ matrix $W = (w_{ij})$, $i, j = 0, 1, \cdots, k$ by $w_{ij} = 3E'(X'_i, Y'_j)$ (the factor 3 is merely for convenience). In the obvious manner we think of the $(k+1)$-vectors X, Y, X_p, Y_p as being mixed strategies for the matrix game W. Now suppose we were able to find optimal strategies X^*, Y^* and the value v for W, then for all $i, j = 0, 1, \cdots, k$ we would have

$$E(i, Y^*) \leqq v \leqq E(X^*, j),$$

where E is the expectation function for W. It is not difficult to see that this implies

$$E'(X'_i, Y^{*\prime}) \leqq \frac{v}{3} \leqq E'(X^{*\prime}, Y'_j).$$

Then by virtue of (1) it follows from the previously mentioned criterion for optimal strategies, that $X^{*\prime}$ and $Y^{*\prime}$ will be optimal strategies for Γ'_k and $v/3$ will be the value. Thus it will be enough if we can solve the matrix game W. It may easily be verified that W is as below.

$$\begin{bmatrix}
3k+3 & 3k+3 & 3k+3 & 3k+3 & \cdots & 3k+3 & 3k+3 \\
5k+4 & 3k+6 & 3k+8 & 3k+10 & \cdots & 5k+2 & 5k+4 \\
5k+3 & 5k+7 & 3k+9 & 3k+11 & \cdots & 5k+3 & 5k+5 \\
5k+2 & 5k+6 & 5k+10 & 3k+12 & \cdots & 5k+4 & 5k+6 \\
5k+1 & 5k+5 & 5k+9 & 5k+13 & \cdots & 5k+5 & 5k+7 \\
 & & & \cdots & & & \\
4k+7 & 4k+11 & 4k+15 & 4k+19 & \cdots & 6k-1 & 6k+1 \\
4k+6 & 4k+10 & 4k+14 & 4k+18 & \cdots & 6k & 6k+2 \\
4k+5 & 4k+9 & 4k+13 & 4k+17 & \cdots & 8k+1 & 6k+3
\end{bmatrix}$$

The matrix W is defined exactly by the following equations:

$$w_{0j}=3k+3, \qquad j=0,1,2,\cdots,k,$$

$$w_{1j}=\begin{cases}5k+4, & j=0,\\ 3k+4+2j, & 0<j\leqq k,\end{cases}$$

$$w_{i0}=\begin{cases}3k+3, & i=0,\\ 5k+5-i, & 0<i\leqq k.\end{cases}$$

For $i=1,2,\cdots,k-1$,

$$w_{i+1,j}-w_{ij}=\begin{cases}-1, & 0\leqq j<i,\\ 2k+1, & j=i,\\ 1, & i<j\leqq k.\end{cases}$$

For $j=0,1,2,\cdots,k-1$,

$$w_{i,j+1}-w_{ij}=\begin{cases}0, & i=0,\\ 2, & 0<i\leqq j,\\ 2-2k, & i=j+1,\\ 4, & j+1<i\leqq k.\end{cases}$$

The matrix game W is solved in the following theorem.

THEOREM 1. *Let $f=1/g=k/(k+1)$, and let r be the least nonnegative integer such that $g^{k-r-1}<2$ (equivalently $2f^{k-r-1}-1>0$). Then for the matrix game W,*

$$X^*=\left(0,f,f^2,\cdots,f^{k-r-1},\frac{k}{2}(2f^{k-r-1}-1),0,\cdots,0\right)\bigg/\frac{k}{2}$$

is an optimal strategy for Player 1, and

$$Y^*=(k(1+g^{k-r-1}),1,g,g^2,\cdots,g^{k-r-2},0,\cdots,0)/2kg^{k-r-1}$$

is an optimal strategy for Player 2. The value v of the game is given by $v=2(k+1)f^{k-r}+4k+3-r$.

Proof. Let W_j and $W^j, j=0,1,\cdots,k$ respectively denote the row and column vectors of W. For $j=0,1,\cdots,k$ we denote the dot products $2kg^{k-r-1}Y^*\cdot W_j$ and $(k/2)X^*\cdot W^j$ by p_j and q_j respectively.

Elementary calculations will yield

$$\frac{p_1}{2kg^{k-r-1}}=\frac{2q_0}{k}=2(k+1)f^{k-r}+4k+3-r,$$

and the proof of the theorem can be concluded by proving that:

(2a) $$p_1=p_2=\cdots=p_{k-r},$$

(2b) $$p_i\leqq p_1 \quad \text{for } i\notin\{1,2,\cdots,k-r\},$$

and

(3a) $$q_0=q_1=\cdots=q_{k-r-1},$$

(3b) $$q_i\geqq q_0 \quad \text{for } i\notin\{0,1,\cdots,k-r-1\}.$$

The proofs of (2a), (2b) and (3a) are straightforward and do not in fact involve the use of our particular choice of r. The proof of (3b) is perhaps worth noting. Since $w_{ik}\geqq w_{i0}$ for $i=0, 1, \cdots, k$ we have $q_k\geqq q_0$. If now $r=0$ then (3b) follows. For $r\geqq 1$, it is easy to verify that, $q_{k-r}-q_{k-r-1}=k(k+1)(1-2f^{k-r})\geqq 0$ by the choice of r. For $k-r\leqq i\leqq k-2$, $q_{i+1}-q_i=k\geqq 0$ which implies (3b), and the proof of the theorem.

Because the choice of r does not appear in the proofs of (2a) and (2b) we might remark that player 2 could interpret the choice of r as minimizing the quantity $2(k+1)f^{k-r}+4k+3-r$.

By the theorem, X^* and Y^* are optimal strategies of the game Γ'_k whose value $v'_k=v/3$. We now go on to prove that they remain optimal in the game Γ_k. In doing so we find it convenient to prove a somewhat more general result. Let $X=(x_0, x_1, \cdots, x_k)$ be any hider strategy such that $x_0=0$ and $x_1\geqq x_2\geqq\cdots\geqq x_k$. It is easy to see that X^* is such a strategy since we may show that $(k/2)(2f^{k-r-1}-1)\leqq f^{k-r-1}$ as follows. For $k\geqq 2$ and $r\geqq 1$ we have $2f^{k-r}\leqq 1\leqq k^2/(k^2-1)$, whence $2f^{k-r-1}(k-1)\leqq k$ and hence the result. The remaining cases involving $k=1, 2, 3, 4$ and $r=0$ may be verified directly.

Since $E(X^*, S_{120})=v'_k$ it will be enough to prove that $E(X, S)\geqq E(X, S_{120})$ for any pure searcher strategy S of Type 2 or 3. We begin with Type 2, and because $E(X^*, S_{mno})$ is independent of m and n we need only consider strategies S of the kind T_{1i2}. Introducing for $p=0, 1, \cdots, k$ the notation $N_p=\sum_{t=0}^{p} x_t$ and $D_p=\sum_{t=0}^{p} tx_t$, it is straightforward to verify that

$$E(X, T_{1i2})-E(X, S_{120})=\tfrac{2}{3}((i-1)N_k+iN_{k-i}+D_{k-i}),$$

whence $E(X, T_{1i2})\geqq E(X, S_{120})$.

We pass finally to strategies of Type 3, and we will attain our objective by showing that against X strategies of the kind U_{1i2j1} and U_{1i2j2} are not better than T_{1i2}. These strategies only differ from T_{1i2} subsequent to arriving at the point $A_{2,k-j+1}$, and it will be convenient to compare expectancies counting this point as the starting point. Denoting these expectancies by $E_1(X, T_{1i2}), E_2(X, U_{1i2j1})$ and $E_3(X, U_{1i2j2})$ it is straightforward to verify that

$$E_2-E_1=\tfrac{2}{3}((j+i-k)N_k-D_k+jN_{k-i}+(k+j-i)N_{k-j}+D_{k-j})$$

and

$$E_3-E_1=\tfrac{2}{3}((j+i-k)N_k-D_k+kN_{k-i}+jN_{k-j}+D_{k-j}).$$

First let c_p denote the coefficient of x_p in E_2-E_1. When $i\leqq j$,

$$c_p=\begin{cases} 2j, & 1\leqq p\leqq k-j,\\ \tfrac{2}{3}(2j+i-k-p), & k-j<p\leqq k-i,\\ \tfrac{2}{3}(j+i-k-p), & k-i<p\leqq k,\end{cases}$$

and when $i\geqq j$,

$$c_p=\begin{cases} 2j, & 1\leqq p\leqq k-i,\\ \tfrac{4}{3}j, & k-i<p\leqq k-j,\\ \tfrac{2}{3}(j+i-k-p), & k-j<p\leqq k.\end{cases}$$

In both cases we observe that if $c_p\leqq 0$ for $p=t$ then $c_p\leqq 0$ for all $p\geqq t$. Since also $x_1, x_2, \cdots, x_k$ is a decreasing sequence, to establish that $E_2-E_1\geqq 0$ it will be enough to show that $\sum_{p=1}^{k} c_p\geqq 0$. This summation of the coefficients may be obtained directly or, perhaps more simply, by a difference in stepcounts (by a stepcount we mean, count each $x_p, p \neq 0$, as 1) of U_{1i2j1} and T_{1i2}. In both cases $\sum_{p=1}^{k} c_p = (j/3)(2k-j-1)$ and the result $E_2 \geqq E_1$ follows. In a precisely similar manner it can be shown that $E_3 \geqq E_1$.

We can now say that X^* and Y^* are optimal strategies for the game Γ_k with value $v_k = \frac{1}{3}(2(k+1)f^{k-r} + 4k + 3 - r)$. The reader will already be aware that the lengths of the arcs and the equality of the intervals of subdivision play no essential part in the discrete game Γ_k. Consider now the game $\bar{\Gamma}_k$, identical to Γ_k except that the payoff is $1/(k+1)$ times the number of steps taken by the searcher to discover the hider. Optimal strategies will be unchanged but the value $\bar{v}_k$ will be $(2(k+1)f^{k-r} + 4k + 3 - r)/3(k+1)$. From the choice of r as the least nonnegative integer such that $(1+1/k)^{k-r-1} < 2$, it easily follows that

$$\frac{k-1}{k+1} - \frac{\ln 2}{(k+1)\ln(1+(1/k))} < \frac{r}{k+1} \leqq \frac{k}{k+1} - \frac{\ln 2}{(k+1)\ln(1+(1/k))},$$

whence $r/(k+1) \to 1 - \ln 2$ as $k \to \infty$.

Also,

$$\bar{v}_k = \frac{1}{3}\left(2\left[\left(1+\frac{1}{k}\right)^{k+1}\right]^{r/(k+1)}\left(1+\frac{1}{k}\right)^{-k} + \frac{4k+3}{k+1} - \frac{r}{k+1}\right)$$

so that $\bar{v}_k \to \frac{1}{3}(4 + \ln 2)$ as $k \to \infty$.

Now consider the optimal strategy X^* of the theorem, and for a given m $(m = 1, 2, 3)$ let $p(N)$ denote the probability that the hider chooses a point A_{mi} such that $1 \leqq i \leqq N$.

Then,

$$p(N) = \begin{cases} (2/k)\sum_{i=1}^{N} f^i = 2(1-f^N), & 1 \leqq N \leqq k-r-1, \\ 1, & N > k-r-1. \end{cases}$$

Hence for any real number x such that $0 \leqq x < 1$, if $F_k(x)$ denotes the probability that on a given arc the hider chooses a point whose distance from A is at most x, then

$$\lim_{k\to\infty} F_k(x) = \lim_{k\to\infty} 2(1 - f^{(k+1)x}) = 2(1 - e^{-x}).$$

Turn now to the optimal strategy Y^*. Concerning the choice of an S_{mnj}; for given m and n let $p(N)$, $N = 0, 1, 2, \cdots, k$, denote the probability that the chosen integer j is such that $0 \leqq j \leqq N$.

Then,

$$p(N) = \begin{cases} k(1+g^{k-r-1}), & N = 0, \\ [k(1+g^{k-r-1}) + \sum_{i=0}^{N-1} g^i]/2kg^{k-r-1} = \frac{1}{2}(1+g^{N-k+r+1}), & 1 \leqq N \leqq k-r-1, \\ 1, & N > k-r-1. \end{cases}$$

Hence for any real number y such that $0 \leqq y < 1$, if $G_k(y)$ denotes the probability that for given m and n the point A_{nj} $(j \neq 0)$ or $A(j = 0)$ is at most distance y from A, then

$$\lim_{k\to\infty} G_k(y) = \lim_{k\to\infty} \tfrac{1}{2}(1 + g^{(k+1)y-k+r+1})$$

$$= \lim_{k\to\infty} \tfrac{1}{2}(1 + g^{(k+1)y-(k-1)}(g^{k+1})^{r/(k+1)}) = \tfrac{1}{2} + \tfrac{1}{4}e^y.$$

Thus we have found optimal strategies and the value for the continuous game corresponding to $\bar{\Gamma}_k$.

REFERENCES

[1] S. Gal, *Search Games*, Academic Press, London, 1980.
[2] J. C. C. McKinsey, *Introduction to the Theory of Games*, McGraw-Hill, New York, 1952.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, 1984

0196-5212/84/0501-0013 $01.25/0

SOME ASPECTS OF CLUSTERING FUNCTIONS*

GERHARD HERDEN†

Abstract. Janowitz's concept of hierarchical clustering which includes the concepts of hierarchical clustering due to Jardine and Sibson and Matula is extended—following the main stream of the theory of partially ordered sets—to describe all connections between hierarchies and isotone functions which measure the homogeneity or compactness of sets of data. In particular a very general description of those hierarchies which correspond bijectively to Hubert's k-clustering functions is presented. As a consequence an exact characterization of the discrepancies between the original concept of Jardine and Sibson—its generalization due to Janowitz—and Hubert's concept of hierarchical clustering is possible.

Introduction. One of the main concepts of general cluster analysis is the concept of a real valued function f which measures the homogeneity (compactness) of the subsets of a given finite set S of data (cf. Bock [2, § 8]).

A first systematic study of these functions was published in 1977 by L. Hubert [5]. Following Hubert's notation these functions will be called *clustering functions.* Hubert observed especially that a monotone clustering function induces canonically a *hierarchy* (H, h) on S. But he did not clarify this point completely. Clustering functions which are not monotone do not induce suitable hierarchies on S (cf. Bock [2, § 37] and § 1 of this paper).

In 1968 N. Jardine and R. Sibson (11) presented a very useful model of hierarchical clustering. This model was discussed extensively by both authors in their book on mathematical taxonomy [12]. On the other hand the model of Jardine and Sibson did not include many powerful clustering techniques. Especially it did not include the *complete linkage method* which had already been studied by Johnson [13] (cf. also Hubert [3] and [4]). In order to include these techniques Jardine and Sibson's model was generalized to a purely order theoretic model by M. F. Janowitz [6] which includes—at least in its most general form (cf. for example [8])—D. W. Matula's graph theoretic model [14] as was noticed by Janowitz (cf. [8]). Janowitz's idea was based upon the discovery that Jardine and Sibson's *numerically stratified clustering* functions are just the *residual mappings* from the nonnegative reals to the set of symmetric and reflexive relations on S. Hence clustering methods may be regarded as transformations on certain sets of residual mappings. This allowed Janowitz in particular to use the powerful tool of residual mappings in his study of hierarchical clustering methods (cf. his studies in [6, § 6] which he continued in [7]).

Numerically stratified clustering functions correspond bijectively to the set of dissimilarity coefficients on S (cf. [11] and [12]). In order to save this fundamental correspondence Janowitz followed in [6] the spirit of Jardine and Sibson's ideas.

Unfortunately there is no bijective correspondence between the set of dissimilarity coefficients and the set of monotone clustering functions on S. This was already observed by Hubert [5] and is clarified completely in the third section of this paper. In order to overcome this difficulty we show in the first section of this paper how to modify the original model of Janowitz so that Hubert's monotone clustering functions can be studied within a new model of "Janowitz type". Our model emphasizes the *maximal linked sets* or the hierarchies on S which one really wants to know.

* Received by the editors February 3, 1982, and in revised form March 8, 1983.

† Fachbereich für Mathematik und Didaktik der Mathematik, Universität/Gesamthochschule, Essen, West Germany.

In Jardine and Sibson's concept all clustering methods are *subdominant* methods. In contrast to this, we want to study within our model also dominant methods (cf. the complete linkage method). Especially we would like to describe—for a given monotone clustering function f—all *minimal dominating* monotone clustering functions f^d of f which satisfy some overlapping criterion in the sense of Hubert [5]. In order to approach this problem we thus clarify—in a first step—all connections between Hubert's monotone k-clustering functions and the hierarchies on S (cf. § 2). In a forthcoming paper we shall present our solution of this problem. The discrepancies between Hubert's monotone k-clustering functions and k-ultrametrics which were introduced by Jardine and Sibson (cf. for example [12]) are described completely in § 4.

Following the referee's suggestions an effort has been made to make our concept of hierarchical clustering as general as possible. In particular we develop our model within the mainstream of the theory of partially ordered sets. Furthermore no finiteness conditions are required. The reader should consult the account of T. S. Blyth and M. F. Janowitz [1] whenever necessary.

1. Clustering functions and hierarchies.

1.1. The basic situation. Let $(L, \leqq)$ and $(M, \leqq)$ be partially ordered sets with 0. In all "classical" concepts of hierarchical clustering $(L, \leqq)$ is the power set of a finite set S of data partially ordered by set inclusion and $(M, \leqq)$ is the set of nonnegative reals partially ordered in the usual manner.

Now we consider the complete lattice $(\bar{L}, \subset)$ of order ideals of L. For each nonempty set $K \subset L$ the order ideal generated by K is denoted by I_K. We shall see that the maximal generators of certain order ideals of L replace in our model the maximal linked sets of S which play an important role in many concepts of hierarchical clustering.

Following the notation of Janowitz (cf. [6]) $\mathrm{Res}^+(M, \bar{L})$ denotes the set of residual mappings $g: M \to \bar{L}$ and $\mathrm{Res}(\bar{L}, M)$ denotes the set of residuated mappings $\bar{g}: \bar{L} \to M$. The reader would do well to recall that a mapping $g: M \to \bar{L}$ is residual iff g is isotone and there exists a (necessarily unique) mapping $\bar{g}: \bar{L} \to M$ such that $\bar{g}g(m) \leqq m$ for all $m \in M$ and $g\bar{g}(I) \supset I$ for all $I \in \bar{L}$. The mapping $\bar{g}$ is called the *residuated mapping associated with* g. We assume the reader is familiar with the basic facts of residuation theory (cf. [1]).

Definition 1.1. An isotone mapping $f: L \to M$ is called a *pre-clustering function* iff there exists an isotone mapping $\bar{f}: \bar{L} \to M$ which satisfies the following conditions:

(P0) $\bar{f}(0) = 0$.

(P1) If $I \in \bar{L}$ and if $\bar{f}(I_a) \leqq m$ for all $a \in I$ then $\bar{f}(I) \leqq m$.

(P2) $\bar{f}(I_a) = f(a)$ for all $a \in L$.

These conditions imply immediately that $\bar{f}$ is uniquely determined. On the other hand it is easy to see that an isotone mapping $\bar{f}: \bar{L} \to M$ which satisfies the conditions (P0) and (P1) is residuated. Hence an isotone mapping $f: L \to M$ is a pre-clustering function iff there exists a residuated mapping $\bar{f}: \bar{L} \to M$ such that $f(a) = \bar{f}(I_a)$ for all $a \in L$. Let $P(L, M)$ be the set of pre-clustering functions $f: L \to M$. We summarize the afore-mentioned facts:

Proposition 1.1. *There are natural bijections between any pairs of the following sets*: $P(L, M)$, $\mathrm{Res}(\bar{L}, M)$ *and* $\mathrm{Res}^+(M, \bar{L})$.

Some remarks. 1. Let $\mathrm{Trip}(\bar{L}, M)$ be the set of all triples (A, B, h) where A is a closure map on $\bar{L}$, B is a dual closure map on M and h is an (order) isomorphism of the range of A onto the range of B, then a combination of Proposition 1.1 with a result of Blyth and Janowitz [1] implies that there are canonical bijections between any pairs of the following sets: $P(L,M)$, Res $(\bar{L},M)$, Res$^+$ $(M,\bar{L})$ and Trip $(\bar{L},M)$.

2. A pre-clustering function $f:L\rightarrow M$ is obviously bounded, i.e., there exists some $m\in M$ such that $f(a)\leqq m$ for all $a\in L$. Let $I(L,M)$ be the set of all bounded isotone mappings $f:L\rightarrow M$ with $f(0)=0$. In order to present a sufficient condition for $I(L,M)$ to coincide with $P(L,M)$ we assume that $(M,\leqq)$ is a complete meet semilattice. In this case for every $f\in I(L,M)$ the isotone mapping $\bar{f}:\bar{L}\rightarrow M$ defined by $\bar{f}(I):=\inf\{m\in M|f(a)\leqq m$ for all $a\in I\}$ clearly satisfies the conditions (P0), (P1) and (P2). Hence $I(L,M)\subset P(L,M)$ which implies the equation of $I(L,M)$ and $P(L,M)$. In particular we may conclude that every monotone clustering function in the sense of Hubert (cf. the introduction) is a pre-clustering function. We now consider the set $C(M,\bar{L})$ of mappings $g:M\rightarrow\bar{L}$ such that $L\in\operatorname{Im}(g)$ and arbitrary meets are preserved by g. There is a natural one-to-one correspondence between $I(L,M)$ and $C(M,\bar{L})$ defined by $g:I(L,M)\rightarrow C(M,\bar{L})$ with $g_f(m):=\{a\in L|f(a)\leqq m\}$ for all $m\in M$ and $f:C(M,\bar{L})\rightarrow I(L,M)$ with $f_g(a):=\inf\{m\in M|a\in g(m)\}$ for all $a\in L$.

We have just proved that $I(L,M)=P(L,M)$. This implies that $C(M,\bar{L})=$ Res$^+$ $(M,\bar{L})$. Thus we have verified the well-known fact that Res$^+$ $(M,\bar{L})=C(M,\bar{L})$ if $(M,\leqq)$ is a complete meet semilattice (cf. [1]).

3. The following remark is due to the referee. If L is finite and if $(M,\leqq)$ is the set of nonnegative reals partially ordered in the usual manner then every $g\in$ Res$^+$ $(M,\bar{L})=C(M,\bar{L})$ can be characterized by the following three properties (cf. also [6, Lemma 4.1]):

(i) g is isotone;

(ii) $g(h)=L$ for some $h\in M$;

(iii) for every $h\in M$ there exists some $\delta>0$ such that $g(h)=g(h+\delta)$.

The reader may note the obvious connection with Jardine and Sibson's numerically stratified clustering functions. To be more precise let S be a finite set and put $SS:=\{\{a,b\}|a,b\in S\}\cup\{\emptyset\}$.

If SS is partially ordered by set inclusion then $(\overline{SS},\subset)$ is obviously (order) isomorphic to the power set of $\{\{a,b\}|a,b\in S\}$.

Because of Remark 2 there exists a natural bijection between $I(SS,M)$ and $C(M,\overline{SS})$. The reader may easily verify that this bijection yields to the well-known one-to-one correspondence between dissimilarity coefficients on S and numerically stratified clustering functions (cf. [12]).

4. Let $(K,\leqq)$ be another partially ordered set with 0. Following the spirit of Janowitz (cf. for example [8]) we define a *pre-cluster method* to be a function $F:$ Res$^+$ $(M,\bar{L})\rightarrow$ Res$^+$ $(M,\bar{K})$. Because of Proposition 1.1 a pre-cluster method may also be thought of as a function $\bar{F}:\operatorname{Res}(\bar{L},M)\rightarrow\operatorname{Res}(\bar{K},M)$ or as a function $T:P(L,M)\rightarrow P(K,M)$.

If L and K are finite sets and if $(M,\leqq)$ is a join semilattice then it is easy to see that all isotone mappings $f_L:L\rightarrow M$ and $f_K:K\rightarrow M$ with $f_L(0)=f_K(0)=0$ are pre-clustering functions. This proves in particular that the following inclusions hold: $\operatorname{Res}(L,M)\subset P(L,M)$ and $\operatorname{Res}(K,M)\subset P(K,M)$ or equivalently Res$^+$ $(M,L)\subset$ Res$^+$ $(M,\bar{L})$ and Res$^+$ $(M,K)\subset$ Res$^+$ $(M,\bar{K})$.

In the finite case we have thus proved—as the reader may easily verify—that every L-cluster method in the sense of Janowitz is a pre-cluster method. On the other hand $(\bar{L},\leqq)$ and $(\bar{K},\leqq)$ are partially ordered sets with 1. Hence our model may be regarded as an abstract model of "Janowitz type" (cf. the introduction). We shall see soon that our model provides just the right tool to study Hubert's monotone clustering functions within a very general context.

The model of Janowitz includes Matula's graph theoretic model of cluster analysis (cf. [8]). Hence Matula's graph theoretic concept may be subsumed in particular under our concept of pre-clustering.

1.2. Maximal elements and hierarchies. Let $f: L \to M$ be a pre-clustering function. For every $m \in \operatorname{Im}(f)$ the "clusters" one is really interested in are those elements a of L which satisfy the following conditions:

(i) $f(a) \leqq m$;

(ii) if $a < b$ then $f(b) \not\leqq m$ for all $b \in L$.

The elements a of L which satisfy the conditions (i) and (ii) will be called *maximal elements* of f.

Examples. Let L be the power set of a finite set S of data and let M be the set of nonnegative reals.

1. The maximal elements of f coincide with the sets $P_{\max}(f, \varepsilon)$ which Hubert considered in (5).

2. If f is the diameter-function of some dissimilarity coefficient d on S then the maximal elements of f coincide with the maximal linked sets of d which Jardine and Sibson considered in their concept of hierarchical clustering (cf. [12]).

Let H be the set of maximal elements of f. In the case that L is not finite it may happen that H is empty. This means in particular that the maximal elements of f—the clusters—do not describe f.

On the other hand Johnson emphasized in his now "classical" paper on hierarchical clustering schemes [13] the one-to-one correspondence between ultrametrics and *hierarchies* (collections of maximal linked sets of ultrametrics) on a finite set S of data. This means in our terminology that Johnson emphasized the one-to-one correspondence between diameter-functions f of ultrametrics and hierarchies (collections of maximal elements of f) on S.

In order to study this fundamental property within our concept of hierarchical clustering we thus have to look for classes "C" of pre-clustering functions which satisfy the following conditions:

(i) every $f \in C$ is completely described by its maximal elements;

(ii) there exists a subset $\tilde{L} \subset \bar{L}$ and a bijection $g: C \to \operatorname{Res}^{+}(M, \tilde{L})$.

In [10] Janowitz studied the class C of isotone mappings $f: L \to M$ (with $f(0) = 0$) such that for every $m \in M$ the order ideal $\{a \in L \mid f(a) \leqq m\}$ is finitely generated. If we assume M to be a join semilattice—this would make direct contact with the model contained in Janowitz [6]—C satisfies the conditions (i) and (ii). In order to prove (ii) let $\tilde{L}$ be the set of all finitely generated order ideals of L and verify the existence of a canonical bijection $g: C \to \operatorname{Res}^{+}(M, \tilde{L})$.

Unfortunately the elements of C are not necessarily pre-clustering functions. Furthermore we do not want to hang onto any version of finiteness. Hence we shall look at a quite different and larger class of pre-clustering functions which satisfy the conditions (i) and (ii).

Henceforth we assume that *every linearly ordered subset of* $(L, \leqq)$ *has a least upper bound.* Now we set $\tilde{L} := \{I \in \bar{L} \mid \sup(K) \in I$ for every linearly ordered subset $K \subset I\}$. Clearly $\tilde{L}$ contains all principal ideals of L. Our assumption implies that for every order ideal $I \in \bar{L}$ there exists some uniquely determined smallest order ideal $I^c \in \tilde{L}$ such that $I \subset I^c$. Moreover the mapping $I \to I^c$ is a closure operator on $\bar{L}$.

We are now ready for the following

DEFINITION 1.2. An isotone mapping $f: L \to M$ is called a *clustering function* iff there exists an isotone mapping $\tilde{f}: \tilde{L} \to M$ which satisfies the following conditions:

(C0) $\tilde{f}(0) = 0$.

(C1) If $I \in \tilde{L}$ and if $\tilde{f}(I_a) \leqq m$ for all a of an arbitrary set of generators of I then $\tilde{f}(I) \leqq m$.

(C2) $\tilde{f}(I_a) = f(a)$ for all $a \in L$.

We denote by Clus (L, M) the set of clustering functions $f: L \to M$. The basic properties of clustering functions will be summarized in the following

LEMMA 1.2. (i) Clus $(L, M) \subset P(L, M)$.

(ii) *For every* $f \in$ Clus (L, M) *and every* $a \in L$ *there exists a maximal element* b *of* f *such that* $a \leqq b$ *and* $f(a) = f(b)$.

(iii) *There exist canonical bijections between any pairs of the following sets*: Clus (L, M), Res $(\tilde{L}, M)$ *and* Res$^+$ $(M, \tilde{L})$.

(Hence Clus (L, M) is a suitable class of pre-clustering functions we have looked for.)

Proof. (i) Condition C1 implies that the function $\bar{f}: \bar{L} \to M$ defined by $\bar{f}(I) := \tilde{f}(I^c)$ for all $I \in \bar{L}$ satisfies condition P1.

(ii) For every $a \in L$ we consider the set $A := \{b \in L | a \leqq b,\ f(a) = f(b)\}$. The inequality $\tilde{f}(I_A^c) \leqq f(a)$ implies that $f(b) \leqq f(a)$ for all $b \in I_A^c$. Hence every linearly ordered subset $K \subset A$ has an upper bound in A. Zorn's lemma leads now to the desired conclusion.

(iii) Clearly $\tilde{f}$ is uniquely determined for all $f \in$ Clus (L, M). On the other hand it is easy to see (cf. the first part of this section) that $f: L \to M$ is a clustering function iff there exists a residuated mapping $\tilde{f}: \tilde{L} \to M$ such that $f(a) = \tilde{f}(I_a)$ for all $a \in L$.

We now characterize those pre-clustering functions $f: L \to M$ which are clustering functions.

PROPOSITION 1.3. *A pre-clustering function* $f: L \to M$ *is a clustering function iff for every linearly ordered subset* K *of* L *and every* $m \in M$ *the following condition holds*:

(B) *If* $f(a) \leqq m$ *for all* $a \in K$ *then* $f(\sup(K)) \leqq m$.

Proof. $\Rightarrow$ Let K be a linearly ordered subset of L such that $f(a) \leqq m$ for all $a \in K$. The inequality $\tilde{f}(I_K^c) \leqq m$ implies immediately that condition (B) holds.

$\Leftarrow$. Let N be an arbitrary nonempty subset of L_c such that $f(a) \leqq m$ for all $a \in N$. We must show that $\bar{f}(I_N^c) \leqq m$. In order to prove this inequality we first construct I_N^c by transfinite induction.

1. $N_1 := I_N$.
2. If α is a limit ordinal then we set $N_\alpha := \bigcup_{\beta < \alpha} N_\beta$ and if α is not a limit ordinal then we set $N_\alpha := \{a \in L |$there exists a linearly ordered subset K of $N_{\alpha-1}$ such that $a \leqq \sup(K)\}$.

It is easy to see that every N_α is an order ideal of L. L is a set. Hence there exists some ordinal number γ such that $I_N^c = N_\gamma$. Thus it is sufficient to prove that $\bar{f}(N_\alpha) \leqq m$ for all ordinal numbers α. This will be done by transfinite induction.

1. If $\alpha = 1$ or if α is a limit ordinal such that $\bar{f}(N_\beta) \leqq m$ for all $\beta < \alpha$ then the desired inequality follows from condition (P1).
2. If α is not a limit ordinal and if $\bar{f}(N_{\alpha-1}) \leqq m$ then we consider for every $a \in N_\alpha$ a linearly ordered set $K_a \subset N_{\alpha-1}$ such that $a \leqq \sup(K_a)$. Condition (B) implies that $f(a) \leqq f(\sup(K_a)) \leqq m$. Hence $\bar{f}(N_\alpha) \leqq m$ because of condition (P1).

Remarks. 1. If L is finite then $P(L, M)$ and Clus (L, M) coincide. Hence Remarks 3 and 4 of § 1 remain valid for clustering functions.

2. If $(M, \leqq)$ is a complete meet semilattice we may conclude from Remark 2 of § 1 and from Proposition 1.3 that $f \in$ Clus (L, M) iff the following conditions hold:

(i) f is bounded.

(ii) f is isotone.

(iii) $f(0)=0$.

(iv) If $f(a)\leqq m$ for all a of a linearly ordered subset $K\subset L$ then $f(\sup(K))\leqq m$.

This demonstrates in particular that Hubert's monotone clustering functions appear in our formulation as very special clustering functions.

Before we now start to characterize those subsets of L which appear as a collection of maximal elements of some clustering function f two pieces of notation seem to be useful.

1. For every nonempty subset H of L and every $b\in L$ we set $H_b:=\{a\in H \mid b\leqq a\}$.
2. For every nonempty subset $H\subset L$ the order ideals of H will be denoted by I^H.

Definition 1.3. A pair (H, h) $(H\subset L, h: H\to M)$ is called a *hierarchy* iff there exists an isotone mapping $\bar h:\bar H\to M$ such that the following conditions are satisfied:

(H0) $\{a\in H \mid h(a)=0\}\neq\emptyset$.

(H1) $I_H=L$.

(H2) $h(a)=\bar h(I_a^H)$ for all $a\in H$.

(H3) $a<b\Rightarrow h(a)<h(b)$ for all $a, b\in H$.

(H4) For every $b\in L$ there exists some $c\in H_b$ such that $h(c)\leqq h(a)$ for all $a\in H_b$.

(H5) If $h(a)\leqq m$ for all a of some subset $K\subset H$ then there exists an order ideal $I^H\in\bar H$ such that $I_K^c\subset I_{I^H}$ and $\bar h(I^H)\leqq m$.

Examples. 1. Let $(L,\leqq)$ be a complete lattice and let every principal ideal of $(M,\leqq)$ be a complete lattice. We shall prove a bit later that a pair (H, h) $(H\subset L, h: H\to M)$ is a hierarchy iff the following conditions hold:

(H0) $\{a\in H \mid h(a)=0\}\neq\emptyset$.

(H$^+$1) $1\in H$.

(H3) $a<b\Rightarrow h(a)<h(b)$ for all $a, b\in H$.

(H$^+$4) For every nonempty subset $K\subset H$ there exists some $b\in H$ such that $\inf(K)\leqq b$ and $h(b)\leqq\inf(h(K))$.

(H$^+$5) For every linearly ordered set $(J,\leqq)$ and every family $\{a_j\}_{j\in J}$ of elements of H there exists some $b\in H$ such that $\sup_{k\in J}(\inf_{j\geqq k} a_j)\leqq b$ and $h(b)\leqq \sup_{j\in J} h(a_j)$.

2. Let $(L,\leqq)$ be a finite lattice (for example the power set of a finite set S of datas partially ordered by set inclusion) and let $(M,\leqq)$ be the set of nonnegative reals (partially ordered in the usual manner) then a pair (H, h) $(H\subset L, h: H\to M)$ is a hierarchy iff the following conditions hold:

(H0) $\{a\in H \mid h(a)=0\}\neq\emptyset$.

(H$^+$1) $1\in H$.

(H3) $a<b\Rightarrow h(a)<h(b)$ for all $a, b\in H$.

Let Hier (L) be the set of all hierarchies on L. The following theorem clarifies the connections between all concepts of hierarchical clustering which we have considered.

Theorem 1.4. *There are natural bijections between any pairs of the following sets*: Clus (L, M), Res (L, M), Res$^+$ (M, L) *and* Hier (L).

Proof. Because of Lemma 1.2(iii) it is sufficient to prove the existence of a natural bijection between Clus (L, M) and Hier (L).

1. For every $f\in$ Clus (L, M) we set $H_f:=\{a\in L \mid a$ is a maximal element of $f\}$ and $h_f:=f|_{H_f}$. In order to prove that (H_f, h_f) is a hierarchy we define an isotone mapping $\bar h_f: H\to M$ by $\bar h_f(I^H):=\tilde f(I_{I^H}^c)$ for all $I^H\in\bar H$. Lemma 1.2(ii) implies immediately that (H_f, h_f) satisfies the conditions (H0), (H1) and (H4). The conditions (H2) and (H3) follow immediately from the definitions of h_f respectively $\bar h_f$. In order to prove (H5) let K be an arbitrary subset of H such that $h(a)\leqq m$ for all $a\in K$. Condition (C1)

implies that $I_K^c \subset \{a \in L \mid f(a) \leqq m\}$. Hence we set $I^H := \{a \in H \mid h_f(a) \leqq m\}$. Because of $I_{I^H} = \{a \in L \mid f(a) \leqq m\} \in \tilde{L}$ the desired conclusion follows.

2. Let (H, h) be a hierarchy. For every $b \in L$ there exists some $c \in H_b$ such that $h(c) \leqq h(a)$ for all $a \in H_b$. This leads to an isotone mapping $f_h : L \to M$ with $f_h(0) = 0$ defined by $f_h(b) := h(c)$ for all $b \in L$.

In order to show that f_h is a clustering function we consider in a first step for each $b \in L$ the set $H_b^m := \{c \in H_b \mid h(c) \leqq h(a) \text{ for all } a \in H_b\}$. In the next step we consider for an arbitrary nonempty subset $N \subset L$ the set $K_N := \bigcup_{b \in N} H_b^m$. Now we define an isotone mapping $\bar{f}_h : \bar{L} \to M$ by $\bar{f}_h(I) := \bar{h}(I_{K_I}^H)$ for all $I \in \bar{L}$. We are now ready to verify in three steps that f_h is a pre-clustering function which satisfies Condition (B). Proposition 1.3 then implies that f_h is a clustering function.

(i) h satisfies Condition (P1). Let I^H be an order ideal of H such that $\bar{h}(I_a^H) \leqq m$ for all $a \in I^H$. Because of Condition (H5) there exists an order ideal J^H of H such that $I_{I^H}^c \subset I_{J^H}$ and $\bar{h}(J^H) \leqq m$. The inclusion $I^H \subset J^H$ now implies that $\bar{h}(I^H) \leqq m$.

(ii) f_h is a pre-clustering function. This follows immediately with the help of (1) from the definitions of f_h respectively $\bar{f}_h$.

(iii) f_h satisfies Condition (B). Let N be a linearly ordered subset of L such that $f_h(a) \leqq m$ for all $a \in N$. We consider the order ideal $I^H := I_{K_N}^H$. Because of Condition (H5) there exists some order ideal J^H of H such that $I_N^c \subset I_{I^H}^c \subset I_{J^H}$ and $\bar{h}(J^H) \leqq m$. f_h is a pre-clustering function. Hence we may conclude that $\bar{f}_h(I_N^C) \leqq \bar{f}_h(I_{I^H}^c) \leqq \bar{f}_h(I_{J^H}) \leqq m$. This proves Condition (B).

3. For $f \in \text{Clus}\,(L, M)$ the clustering function which is induced by (H_f, h_f) will be denoted by f^+ and for $(H, h) \in \text{Hier}\,(L)$ the hierarchy which is induced by f_h will be denoted by (H^+, h^+). In order to finish the proof we have to show that $f = f^+$ and that $(H, h) = (H^+, h^+)$.

(i) $f = f^+$. Let a be an arbitrary element of L. The definition of (H_f, h_f) implies the existence of some $b \in H_f$ such that $a \leqq b$ and $f(a) = h_f(b)$. Hence $f^+(a) \leqq f(a)$. On the other hand we have $f(a) \leqq h_f(b)$ for all $b \in (H_f)_a$ and we may conclude that $f(a) \leqq f^+(a)$.

The reader may easily verify that the equation of (H, h) and (H^+, h^+) follows if $H = H^+$.

(ii) $H \subset H^+$. Let a be an arbitrary element of H. We have to show that a is a maximal element of f_h. Because of Lemma 1.2(ii) there exists some maximal element b of f_h such that $a \leqq b$ and $f_h(a) = f_h(b)$. We now assume that $a < b$. The definition of f_h implies the existence of some $c \in H$ such that $b \leqq c$ and $f_h(b) = h(c)$. But since $h(a) = f_h(a) = f_h(b) = h(c)$ and $a < c$ this contradicts Condition (H3).

(iii) $H^+ \subset H$. Let a be an arbitrary element of H^+. Then a is a maximal element of f_h. On the other hand there exists some $b \in H$ such that $a \leqq b$ and $f_h(a) = h(b) = f_h(b)$. The maximality of a now implies that $a = b$.

We are now ready for a short proof of the first example. The reader may verify that the proof is sufficient.

1. Let $f : L \to M$ be a clustering function. We show that (H_f, h_f) satisfies the conditions (H^+1), (H^+4) and (H^+5). But these conditions follow immediately from Lemma 1.2(ii), the second remark of this section and Proposition 1.3.

2. Let (H, h) $(H \subset L, h : L \to M)$ satisfy the conditions (H0), (H^+1), (H3), (H^+4) and (H^+5). Condition (H^+4) implies that (H4) holds. Furthermore we may conclude from conditions (H1), (H3), (H0) and (H^+5) that f_h is a bounded isotone mapping with $f_h(0) = 0$ and that f_h satisfies condition (B). Hence $f_h \in \text{Clus}\,(L, M)$ because of Remark 1.6 of this section.

We end this subsection with three supplementary remarks:

1. In our concept of hierarchical clustering "clusters" appear in three equivalent versions: as maximal elements of some clustering function f, as maximal generators of the order ideals $\{a \in L \mid f(a) \leqq m\}$ for some clustering function f (cf. the description of the basic situation) or as the elements of a hierarchy (H, h).

2. Let $(M, \leqq)$ be the set of nonnegative reals partially ordered in the usual manner. We may define—as is easily verified—a canonical pre-cluster method $\mathrm{Clus}^-: P(L,M) \to P(L,M)$ by $\mathrm{Clus}^-(f) := \sup\{f^- \in \mathrm{Clus}(L,M) \mid f^- \leqq f\}$ for all $f \in P(L,M)$. $\mathrm{Clus}^-(f)$ is the uniquely determined maximal subdominating clustering function of f.

Furthermore for every $f \in P(L,M)$ and every $m \in \mathrm{Im}(f)$ we may define a clustering function $f_m^+ : L \to M$ by

$$f_m^+(a) := \begin{cases} 0 & \text{if } a = 0, \\ m & \text{if } 0 < a \text{ and } f(a) \leqq m, \\ \bar{f}(L) & \text{otherwise,} \end{cases} \quad \text{for all } a \in L$$

such that $f \leqq f_m^+$ for all $m \in M$. This implies that $\inf\{f^+ \in \mathrm{Clus}(L,M) \mid f \leqq f^+\} = f$. Hence there exists no uniquely determined minimal dominating clustering function of f.

On the other side one may verify that $\inf_{j \in J}(f_j)$ is a clustering function for every linearly ordered family $\{f_j\}_{j \in J}$ of clustering functions. Hence we may conclude from Zorn's lemma that for every clustering function $f^+ \geqq f$ there exists some minimal dominating clustering function f^d with $f \leqq f^d \leqq f^+$.

It would be interesting to describe pre-cluster methods which associate with every pre-clustering function f a minimal dominating clustering function f^d.

3. Let C be some "suitable" subset of $\mathrm{Clus}(L,M)$. Of particular interest are those *cluster methods* $T^s: \mathrm{Clus}(L,M) \to \mathrm{Clus}(L,M)$ which associate with every $f \in \mathrm{Clus}(L,M)$ some maximal subdominating clustering function $f^s \in C$ of f and those cluster methods $T^d: \mathrm{Clus}(L,M) \to \mathrm{Clus}(L,M)$ which associate with every $f \in \mathrm{Clus}(L,M)$ some minimal dominating clustering function $f^d \in C$ of (cf. Remark 2). These methods generalize in a natural way the *single* and *complete linkage methods*. Of course there are equivalent formulations of these methods using $\mathrm{Res}(\tilde{L}, M)$, $\mathrm{Res}^+(M, \tilde{L})$ or $\mathrm{Hier}(L)$ instead of $\mathrm{Clus}(L,M)$.

A suitable class C of clustering functions will be studied in the next section.

2. *K*-clustering functions and *K*-hierarchies. Let K be a nonempty subset of L. In order to study a generalized version of Hubert's monotone k-clustering functions within our model of hierarchical clustering we consider the order filter F_K which is generated by K.

DEFINITION 2.1. A pre-clustering function $f: L \to M$ is called a *K-pre-clustering function* if and only if it satisfies the following condition for all $a, b \in L$:

(0K) if $I_a \cap I_b \cap F_K \neq \varnothing$ then there exists some $c \in L$ such that $f(c) \leqq \bar{f}(I_a \cup I_b)$ and $I_a \cup I_b \subset I_c$.

Examples. 1. If $(L, \leqq)$ has a greatest element then for every $m \in M$ the mappings $f_m: L \to M$ defined by

$$f_m(a) := \begin{cases} 0 & \text{if } a = 0, \\ m & \text{else,} \end{cases} \quad \text{for all } a \in L,$$

are clearly K-pre-clustering functions for every nonempty subset $K \subset L$.

2. If $(L, \leqq)$ is a lattice and if $(M, \leqq)$ is a join semilattice then condition (0K) may be replaced by the following equivalent condition of "Janowitz type":

If $\inf\{a, b\} \in F_K$ then $f(\sup\{a, b\}) \leqq \sup\{f(a), f(b)\}$ for all $a, b \in L$.

Hence—if L is finite—every join homomorphism $f: L \rightarrow M$ is a K-pre-clustering function for every nonempty subset $K \subset L$.

3. Let L be the power set of a finite set S of data and let M be the set of nonnegative reals. For every natural number $k \geqq 1$ we set $K := \{A \in L \mid |A| = k\}$.

In this case condition (0K) is equivalent to the following condition which is due to Hubert (5):

(A_k) If $|A \cap B| \geqq k$ then $f(A \cup B) \leqq \max\{f(A), f(B)\}$ for all $A, B \in L$.

Hence Hubert's monotone k-clustering functions are special K-pre-clustering functions.

Moreover if f is especially the diameter function of some dissimilarity coefficient d on S then condition (A_k) and hence condition (0K) just means that d is a (weakly) k-ultrametric in the sense of Jardine and Sibson (12).

In order to formulate the basic facts of "K-clustering" we need the following notation and definitions:

1. $P_K(L, M)$ denotes the set of K-pre-clustering functions.

2. The elements of $\mathrm{Clus}_K\ (L, M) := \mathrm{Clus}\ (L, M) \cap P_K(L, M)$ are called K-clustering functions.

3. $\mathrm{Res}_K\ (\bar{L}, M)$ denotes the set of all residuated mappings $\bar{f}: \bar{L} \rightarrow M$ such that $\bar{f}_L \in P_K(L, M)$.

4. $\mathrm{Res}_K\ (\tilde{L}, M)$ denotes the set of all residuated mappings $\tilde{f}: \tilde{L} \rightarrow M$ such that $f_L \in \mathrm{Clus}_K\ (L, M)$.

5. $\mathrm{Res}_K^+\ (M, \bar{L})$ denotes the set of all residual mappings $g: M \rightarrow \bar{L}$ which satisfy for all $m \in M$ and all $a, b \in g(m)$ the following condition:

(0^+K) if $I_a \cap I_b \cap F_K \neq \varnothing$ then there exists some $c \in g(m)$ such that $I_a \cup I_b \subset I_c$.

6. $\mathrm{Res}_K^+(M, \tilde{L})$ denotes the set of all residual mappings $g: M \rightarrow \tilde{L}$ which satisfy condition (0^+K) for all $m \in M$ and all $a, b \in g(m)$.

7. For every $I \in \bar{L}$ we denote by $I_{\max}$ the set of all maximal elements of I. For each $f \in P(L, M)$ and each $m \in M$ the set of all maximal elements of $\{a \in L \mid f(a) \leqq m\}$ will be denoted especially by $L(f)^m_{\max}$.

We are now ready for the following:

Proposition 2.1. (i) *There are natural bijections between any pairs of the following sets*: $P_K(L, M)$, $\mathrm{Res}_K\ (\bar{L}, M)$ *and* $\mathrm{Res}_K^+(M, \bar{L})$.

(ii) *There are natural bijections between any pairs of the following sets*: $\mathrm{Clus}_K\ (L, M)$, $\mathrm{Res}_K\ (\tilde{L}, M)$ *and* $\mathrm{Res}_K^+(M, \tilde{L})$.

Proof. After having verified that a pre-clustering function $f: L \rightarrow M$ satisfies condition (0K) for all $a, b \in L$ iff the corresponding residual mapping $g_f: M \rightarrow \bar{L}$ satisfies condition (0^+K) for all $m \in M$ and all $a, b \in g_f(m)$, one may use Proposition 1.1 and Lemma 1.2(iii) respectively.

Proposition 2.2. (characterization of $\mathrm{Clus}_K\ (L, M)$ and $\mathrm{Res}_K^+(M, L)$).

(i) *The following conditions are equivalent for every* $f \in \mathrm{Clus}\ (L, M)$:

(a) $f \in \mathrm{Clus}_K\ (L, M)$;

(b) $a = b$ *for all* $m \in M$ *and all* $a, b \in L(f)^m_{\max}$ *such that* $I_a \cap I_b \cap F_K \neq \varnothing$.

(ii) *The following conditions are equivalent for every* $f \in \mathrm{Res}^+\ (M, \tilde{L})$:

(a^+) $g \in \mathrm{Res}_K^+(M, \tilde{L})$;

(b^+) $a = b$ *for all* $m \in M$ *and all* $a, b \in g(m)_{\max}$ *such that* $I_a \cap I_b \cap F_K \neq \varnothing$.

Proof. Because $L(f)^m_{\max} = g_f(m)_{\max}$ for every $f \in \mathrm{Clus}\ (L, M)$ it is sufficient to prove the equivalence of (a) and (b).

(a) $\Rightarrow$ (b). Trivial.

(b) $\Rightarrow$ (a). Set $m := \bar{f}(I_a \cup I_b)$. A slight modification of the proof of Lemma 1.2(ii) implies the existence of elements c_a and $c_b \in L(f)^m_{\max}$ such that $a \leqq c_a$ and $b \leqq c_b$. The application of condition (b) leads now immediately to the desired conclusion.

In order to study the connections between K-clustering functions and hierarchies we introduce two more "overlapping criteria".

DEFINITION 2.2. A hierarchy (H, h) is called a *K-hierarchy* iff it satisfies for all $a, b \in H$ the following overlapping criteria:

(HK1) If $I_a \cap I_b \cap F_K \neq \emptyset$ and if $h(a) \leqq h(b)$ then $a \leqq b$.

(HK2) If $I_a \cap I_b \cap F_K \neq \emptyset$ then there exists some $c \in H$ such that $h(c) = \bar{h}(I_a^H \cup I_b^H)$and $I_a \cap I_c \cap F_K \neq \emptyset$ or $I_b \cap I_c \cap F_K \neq \emptyset$.

The reader may notice that (HK2) is a relatively weak condition. For example (HK2) is always satisfied by every hierarchy (H, h) if $(M, \leqq)$ is linearly ordered. In this case the hierarchy (H, h) is a K-hierarchy iff $a \leqq b$ or $b \leqq a$ for all $a, b \in H$ such that $I_a \cap I_b \cap F_K \neq \emptyset$.

Let $\mathrm{Hier}_K\ (L)$ be the set of K-hierarchies on L and let $f: L \to M$ be a K-clustering function. An easy computation—using the modified version of Lemma 1.2(ii) of the proof of Proposition 2.2 and condition (b) of Proposition 2.2—proves that (H_f, h_f) satisfies condition (HK1). On the other hand (HK2) is an immediate consequence of condition (0K). Furthermore a careful analysis of the proof of Theorem 1.4 shows that $f_h \in \mathrm{Clus}_K\ (L, M)$ if $(H, h) \in \mathrm{Hier}_K\ (L)$.

Hence we have obtained the following:

THEOREM 2.3. *There are natural bijections between any pairs of the following sets*: $\mathrm{Clus}_K\ (L, M)$, $\mathrm{Res}_K\ (\tilde{L}, M)$, $\mathrm{Res}_K^+ (M, \tilde{L})$ *and* $\mathrm{Hier}_K\ (L)$.

Examples and remarks. 1. Set $L^- := L \backslash \{0\}$ and let $(L, \leqq)$ and $(M, \leqq)$ be complete meet semilattices. If $(M, \leqq)$ is linearly ordered then the reader may verify—recalling the ideas of the first paragraph—that a pair (H, h) $(H \subset L, h: H \to M)$ is a L^--hierarchy iff it satisfies the following conditions:

(H0) $\{a \in H \mid h(a) = 0\} \neq \emptyset$.

(H1) $I_H = L$.

(H$^-$2) h is bounded.

(H3) $a < b \Rightarrow h(a) < h(b)$ for all $a, b \in H$.

(HL$^-$) $a \leqq b$ or $b \leqq a$ for all $a, b \in H$ such that $\inf \{a, b\} \neq 0$.

(H$^-$4) For every nonempty linearly ordered subset $K \subset H$ there exists some $b \in H$ such that $\inf (K) \leqq b$ and $h(b) \leqq \inf (h(K))$.

(H$^-$5) If $h(a) \leqq m$ for all a of some linearly ordered subset $K \subset H$ then there exists some $b \in H$ such that $a \leqq b$ for all $a \in K$ and $h(b) \leqq m$.

2. Let L be the power set of a finite set S of data and let M be the set of nonnegative reals. For the moment we denote by C the class of all L^--clustering functions $f: L \to M$ with $f(\{a\}) = 0$ for all $a \in S$, by U the set of all ultrametrics on S and by $H(L)$ the set of all pairs (H, h) $(H \subset L, h: H \to M)$ which satisfy the following conditions:

(H$^+$0) $\bigcup \{A \in H \mid h(A) = 0\} = S$.

(H$^+$1) $S \in H$.

(H3) $A \subset B \Rightarrow h(A) < h(B)$ for all $A, B \in H$.

(HL$^-$) $A \subset B$ or $B \subset A$ for all $A, B \in H$ such that $A \cap B \neq \emptyset$.

A combination of our results with some well known classical result (cf. for example Bock [2, Satz 37.1] or Johnson [13]) leads to the following at least implicitly well known fact that *there are natural bijections between any pairs of the following sets*: C, U *and* $H(L)$.

3. Let $(L, \leqq)$ be a join semilattice and let every principal ideal of $(M, \leqq)$ be a complete lattice. If $\{f_j\}_{j \in J}$ is an arbitrary *bounded* family of K-clustering functions then it is easy to see that also the clustering function $\sup_{j \in J} (f_j)$ satisfies condition (0K). On the other hand the mapping $f_0: L \to M$ (cf. the first example of this section)

is a K-clustering function for every nonempty subset $K \subset L$. Hence we may define a (K)-cluster method $T_K^s : \text{Clus}\,(L, M) \to \text{Clus}\,(L, M)$ by $T_K^s(f) := \sup\{f^- \in \text{Clus}_K\,(L, M) | f^- \leqq f\}$ for all $f \in \text{Clus}\,(L, M)$. T_K^s associates every clustering function $f: L \to M$ with its uniquely determined maximal subdominating K-clustering function. T_K^s may be regarded as a generalized cluster method of "Jardine–Sibson type" (cf. their cluster methods (B_k) in [12]).

In a forthcoming paper we shall study (K)-cluster methods $T_K^d : \text{Clus}\,(L, M) \to \text{Clus}\,(L, M)$ which associate with every clustering function $f: L \to M$ some (not necessarily unique) minimal dominating K-clustering function f_K^d of f.

3. Dissimilarity coefficients, K-ultrametrics and d-hierarchies. For the remainder of this paper we assume that $(L, \leqq)$ is a complete, locally atomic and upper continuous lattice and that every principal ideal of $(M, \leqq)$ is a complete lattice.

Let $AL \subset L$ consist of all atoms and the least element of L. We set $ALV := \{a_{ij} \in L |$ there exist elements $a_i, a_j \in AL$ such that $a_{ij} = \sup\{a_i, a_j\}\}$ and $LV := ALV \backslash AL$.

Since AL, ALV and LV are subsets of L they are canonically partially ordered.

DEFINITION 3.1. A clustering function $d := ALV \to M$ is called a *dissimilarity coefficient* iff $d(a) = 0$ for all $a \in AL$.

Examples. 1. *Let* $f := L \to M$ be an arbitrary clustering function with $f(a) = 0$ for all $a \in AL$ then $d_f := f|_{ALV}$ is a dissimilarity coefficient.

2. Let L be the power set of a finite set S of data and let M be the set of nonnegative reals. The reader may easily verify that every dissimilarity coefficient in the sense of Jardine and Sibson [12] may be regarded as a dissimilarity coefficient in the above sense and that conversely every dissimilarity coefficient in the above sense may be regarded as a dissimilarity coefficient in the sense of Jardine and Sibson.

Let $\text{DC}\,(L)$ be the set of dissimilarity coefficients on ALV and let $P(LV)$ be the power set of LV. Clearly we may identify the complete lattices $(\overline{LV}, \subset)$, $(LV, \subset)$ and $(P(LV), \subset)$.

The results of the first paragraph imply immediately the following very general

PROPOSITION 3.1. *There are natural bijections between any pairs of the following sets*: $\text{DC}\,(L)$, $\text{Res}\,(P(LV), M)$ *and* $\text{Res}^+\,(M, P(LV))$.

Clearly Proposition 3.1 holds also under the weaker assumption that $(M, \leqq)$ is a partially ordered set with 0. Hence Proposition 3.1 includes the results of Jardine and Sibson [12] and Janowitz [6] which establish a natural one-one correspondence between dissimilarity coefficients and numerically stratified clustering functions or respectively between dissimilarity coefficients and L-stratified clustering functions (cf. also Remark 3 of § 1.1).

LEMMA 3.2. *There exists a canonical order monomorphism* $\text{diam} : \text{DC}\,(L) \to \text{Clus}\,(L, M)$.

Proof. Because of $d(b) \leqq \tilde{d}(ALV)$ for all $b \in ALV$ we may define a mapping $\text{diam} : \text{DC}\,(L) \to \text{Clus}\,(L, M)$ by $\text{diam}_d\,(a) := \sup\{d(b) | b \leqq a, b \in ALV\}$ for all $d \in \text{DC}\,(L)$ and all $a \in L$.

The definition of diam_d implies immediately that diam_d is a bounded isotone mapping with $\text{diam}_d\,(0) = 0$. Hence diam_d is a pre-clustering function. On the other hand diam_d satisfies Condition (B) since $(L, \leqq)$ is upper continuous and we may conclude that diam_d is actually a clustering function. It is now very easy to see that $\text{diam} : \text{DC}\,(L) \to \text{Clus}\,(L, M)$ is an order monomorphism.

The image of $\text{diam}_d : \text{DC}\,(L) \to \text{Clus}\,(L, M)$ will henceforth be denoted by $\text{Clus}^d\,(L, M)$.

Remarks. 1. diam_d generalizes the usual definition of a diameter-function for a dissimilarity coefficient d in a natural way.

2. This interesting remark is due to the referee: One can define an equivalence relation on $\operatorname{Clus}(L, M)$ by the rule $f_1 \sim f_2$ in case $f_1(a) = f_2(a)$ for all $a \in LV$. Each equivalence class has a least element, namely, the one of the form diam_d. Furthermore $\operatorname{Clus}(L, M)_{/\sim}$ is canonically partially ordered by $\overline{f_1} \leqq \overline{f_2} \Leftrightarrow f_1(a) \leqq f_2(a)$ for all $a \in LV$ and it is easy to see that there exists a commutative diagram with canonical order homomorphisms:

$$\begin{tikzcd}
\operatorname{Clus}(L, M) \arrow[r] & \operatorname{Clus}(L, M)_{/\sim} \\
\operatorname{DC}(L) \arrow[u] \arrow[ur, equal] &
\end{tikzcd}$$

3. The reader may notice that the well-known "Ward algorithm" (cf. Ward [16]) may be regarded as a cluster-method $T_W : \operatorname{Clus}(L, M) \to \operatorname{Clus}^d(L, M) \subset \operatorname{Clus}(L, M)$.

Let a be an arbitrary element of L. In order to clarify the connections between dissimilarity coefficients and hierarchies a subset $K \subset L$ is called *a-complete* iff for every $b \leqq a$ with $b \in ALV$ there exists some $c \in K$ such that $b \leqq c$.

DEFINITION 3.2. A pair (H, h) $(H \subset L, h : H \to M)$ is called a *d-hierarchy* iff the following conditions hold:

(H^+0) $AL \subset I_{\{a \in H \mid h(a) = 0\}}$.

(H^+1) $1 \in H$.

(H3) $a < b \Rightarrow h(a) < h(b)$ for all $a, b \in H$.

($\mathrm{H}^{\oplus}4$) For every $a \in L$ there exists some $b \in H$ such that $a \leqq b$ and $h(b) \leqq \sup(h(K))$ for every a-complete subset $K \subset H$.

Let $\operatorname{Hier}^d(L)$ be the set of all d-hierarchies on L. The following theorem implies especially that $\operatorname{Hier}^d(L)$ is a subset of $\operatorname{Hier}(L)$.

THEOREM 3.3. *There are natural bijections between any pairs of the following sets*: $\operatorname{DC}(L)$, $\operatorname{Clus}^d(L, M)$, $\operatorname{Res}(P(LV), M)$, $\operatorname{Res}^+(M, P(LV))$ *and* $\operatorname{Hier}^d(L)$.

Proof. It is sufficient to establish a natural bijection between $\operatorname{Clus}^d(L, M)$ and $\operatorname{Hier}^d(L)$.

1. Let f be an arbitrary element of $\operatorname{Clus}^d(L, M)$. With the help of Lemma 1.2(ii) the properties of f imply that $(H_{f1} h_f)$ is a d-hierarchy.

2. Let (H, h) be an arbitrary d-hierarchy. One may first note that condition (H4) (cf. Definition 1.3) is an immediate consequence of condition ($\mathrm{H}^{\oplus}4$). Condition (H^+1) implies that f_h is bounded. Now a careful analysis of the proof of Theorem 1.4 allows us to conclude with the help of condition ($\mathrm{H}^{\oplus}4$) that $f_h(a) = \sup\{f_h(b) \mid b \leqq a, b \in ALV\}$ for all $a \in L$ and we may conclude from the proof of Lemma 2.2 that $f_h \in \operatorname{Clus}^d(L, M)$.

3. Because of the proof of Theorem 1.4 we have thus established the desired bijective correspondence between $\operatorname{Clus}^d(L, M)$ and $\operatorname{Hier}^d(L)$.

Remarks. 1. If L is the power set of a finite set S of data and M is the set of nonnegative reals one may replace condition ($\mathrm{H}^{\oplus}4$) by the weaker condition

($\mathrm{H}^0 4$) For every $A \in L$ and every A-complete subset $K \subset H$ there exists some $B \in H$ such that $A \subset B$ and $h(B) \leqq \max(h(K))$.

2. One can define an equivalence relation on $\operatorname{Hier}(L)$ by the rule $(H_1, h_1) \sim (H_2, h_2)$ iff for all $a \in H_1$ there exists an a-complete subset $K_a \subset H_2$ and for all $b \in H_2$ there exists a b-complete subset $K_b \subset H_1$ such that $h_2(c) \leqq h_1(a)$ for all $c \in K_a$ and $h_1(p) \leqq h_2(b)$ for all $p \in K_b$.

The reader may verify (cf. the second remark of this paragraph) that $(H_1, h_1) \sim (H_2, h_2)$ iff $f_{h_1} \sim f_{h_2}$. In particular each equivalence class has a least element which is a d-hierarchy.

We now consider the order filter F_K which we introduced in the second paragraph. In order to study dissimilarity coefficients which satisfy overlapping criteria we need two more definitions:

DEFINITION 3.3 (cf. Jardine and Sibson [12]). 1. A subset $U \subset LV$ is called (*weakly*) *K-transitive* iff for all $a_{ij} \in LV$ and all $a \in K$ the following condition holds: if a_{ik}, a_{kt} and $a_{tj} \in U \cup AL$ for all $a_{kt} \leqq a$ then $a_{ij} \in U$.

2. A dissimilarity coefficient $d: ALV \to M$ is called a (*weakly*) *K-ultrametric* iff for all $a_{ij} \in LV$ and all $a \in K$ the following inequality holds:

$$d(a_{ij}) \leqq \sup \{\sup \{d(a_{ik}), d(a_{kt}), d(a_{tj})\} | a_{kt} \leqq a, a_{kt} \in ALV\}.$$

Let U_K be the set of all (weakly) K-transitive subsets of LV and let DC_K (L) be the set of all (weakly) K-ultrametrics on L. We set: $\mathrm{Clus}_K^d (L, M) := \mathrm{Clus}_K (L, M) \cap \mathrm{Clus}^d (L, M)$ and $\mathrm{Hier}_K^d (L) := \mathrm{Hier}_K (L) \cap \mathrm{Hier}^d (L)$ and we are now ready to prove the following:

THEOREM 3.4 (cf. Hubert [5, Prop. 10]). *There are natural bijections between any pairs of the following sets*: DC_K (L), $\mathrm{Clus}_K^d (L, M)$, $\mathrm{Res}\,(U_K, M)$, $\mathrm{Res}^+ (M, U_K)$ *and* $\mathrm{Hier}_K^d (L)$.

Proof. Let $d: ALV \to M$ be an arbitrary dissimilarity coefficient with corresponding residual mapping $g_d: M \to P(LV)$. The reader may recall that g_d is defined by $g_d(m) := \{a \in LV | d(a) \leqq m\}$ for all $m \in M$. Definition 3.3 implies immediately that d is a (weakly) K-ultrametric iff $g_d(m) \in U_K$ for all $m \in M$. Hence we have already established the natural bijection between DC_K (L) and $\mathrm{Res}^+ (M, U_K)$. In order to complete the proof of the theorem it is thus sufficient to prove that the image of $\mathrm{diam}: \mathrm{DC}_K (L) \to \mathrm{Clus}\,(L, M)$ is $\mathrm{Clus}_K^d(L, M)$. But it is an immediate consequence of Definition 3.3(2) that $d \in \mathrm{DC}_K$ (L) iff diam_d satisfies condition (0K)—one may use here the version of the second example of the preceding paragraph—and nothing remains to prove.

Remarks. 1. The elements of $\mathrm{Res}^+ (M, U_K)$ may be regarded as natural generalizations of Jardine and Sibson's (fine) k-dendrograms.

2. There are three reasons why we excluded the study of strongly K-ultrametrics or even u-diametric dissimilarity coefficients (cf. Jardine and Sibson [12]):

1. They are excluded for the sake of brevity.

2. (Weakly) K-ultrametrics play a more important role in the models of Jardine and Sibson and Hubert. For example they lead to the well-known *flat* cluster methods B_k.

3. (Weakly) K-ultrametrics enabled us to make direct contact with clustering functions which satisfy overlapping criteria (cf. also Hubert [5]).

4. Characterization of overlapping criteria. In Jardine and Sibson's concept of hierarchical clustering [12] and also in the original concept of Janowitz [6] the "clusters" are precisely the maximal linked sets (ML-sets) of some dissimilarity coefficient d. Hence all their hierarchies may be constructed within Hubert's more general concept by considering the clustering functions diam_d. In particular Hubert's (monotone) k-clustering functions are natural generalizations of Jardine and Sibson's (weakly) k-ultrametrics. In order to clarify the connections between both concepts completely let $\mathrm{Clus}_K^0 (L, M)$ be the set of all clustering functions $f \in \mathrm{Clus}_K (L, M)$ such that $f(a) = 0$ for all $a \in AL$. We want to solve the following natural problem:

Determine all order filters F_K such that every $f \in \mathrm{Clus}_K^0 (L, M)$ is of the form diam_d *for some dissimilarity coefficient $d \in \mathrm{DC}_K$ (L), or equivalently*

determine all order filters F_K such that $\mathrm{Clus}_K^0 (L, M) = \mathrm{Clus}_K^d (L, M)$!

Clearly $\mathrm{Clus}_K^0(L, M) = \mathrm{Clus}_K^d(L, M)$ for all nonempty subsets $K \subset L$ if $M = \{0\}$. Thus we may assume without loss generality that M contains at least two elements $0 < m$.

The following theorem solves our problem:

THEOREM 4.1. *The following conditions are equivalent*:

(i) $\mathrm{diam}: \mathrm{DC}_K(L) \to \mathrm{Clus}_K^0(L, M)$ *is an* (*order*) *isomorphism.*

(ii) $\mathrm{Clus}_K^0(L, M) = \mathrm{Clus}_K^d(L, M)$.

(iii) $|AL \backslash F_K| \leqq 3$.

Proof. It is sufficient to prove the equivalence of the conditions (ii) and (iii).

(ii) $\Rightarrow$ (iii). If $|AL \backslash F_K| > 3$ then there exist atoms $a_1, a_2, a_3 \in L$ such that $a_i \notin F_K$ for all $1 \leqq i \leqq 3$. We set $T := \{a_1, a_2, a_3\}$ and $TV := \{b \in L |$ there exist elements $a_i \neq a_j \in T$ such that $b = \sup\{a_i, a_j\}\}$. Now we define a mapping $f: L \to M$ by

$$f(a) := \begin{cases} 0 & \text{if } a \in AL \cup TV, \\ m & \text{else,} \end{cases} \qquad \text{for all } a \in L.$$

The reader may verify immediately that f is actually a clustering function with $f(a) = 0$ for all $a \in AL$. We now consider two different elements $a, b \in L$ such that $\inf\{a, b\} > 0$. The definition of f implies that $f(\sup\{a, b\}) = m > \sup\{f(a), f(b)\}$ iff $a, b \in TV$. But in this case our assumption on F_K implies that $\inf\{a, b\} \notin F_K$. Hence we may conclude that $f \in \mathrm{Clus}_K^0(L, M)$.

Because of $f(\sup(T)) = m > \sup\{f(a) | a \leqq \sup(T),\ a \in ALV\} = 0$ there exists no $d \in \mathrm{DC}_K(L)$ such that $\mathrm{diam}_d = f$.

(iii) $\Rightarrow$ (ii). We have to verify the inclusion $\mathrm{Clus}_K^0(L, M) \subset \mathrm{Clus}_K^d(L, M)$. Hence we consider an arbitrary clustering function $f: L \to M$ with $f(a) = 0$ for all $a \in AL$. Let c be an arbitrary element of L. It is sufficient to prove that $f(c) \leqq \sup\{f(a) | a \leqq c,\ a \in ALV\}$. Therefore we consider the set A_c of all atoms b of L such that $b \leqq c$. We may assume without loss of generality that $|A_c| \geqq 3$. Because of $|AL \backslash F_K| \leqq 3$ there are at most two atoms $p, q \in AL \backslash F_K$. We reconstruct c in a first step by transfinite induction in the following way:

1. set $a_1 := b^1 := p$ if $p \leqq c$
   otherwise set $a_1 := b^1$ for some arbitrary $b^1 \in A_c \backslash \{q\}$
2. if α is a limit ordinal then we set $a_\alpha := \sup_{\beta < \alpha} a_\beta$ if α is not a limit ordinal then we consider the set $A_c^{\alpha-1}$ of all atoms $b \in L$ such that $b \leqq a_{\alpha-1}$ and set

$$a_\alpha := \left\{ \begin{matrix} a_{\alpha-1} & \text{if } A_c \backslash (A_c^{\alpha-1} \cup \{q\}) = \varnothing \\ \sup\{a_{\alpha-1}, b^\alpha\} & \text{for some arbitrary } b^\alpha \in A_c \backslash (A_c^{\alpha-1} \cup \{q\}) \text{ else} \end{matrix} \right\}.$$

AL is a set. Hence there exists an ordinal number γ such that $a_\alpha = a_\gamma$ for all $\alpha \geqq \gamma$. On the other hand L is locally atomic. This implies that $a_\gamma = c$ or that $a_\gamma^+ := \sup\{a_\gamma, q\} = c$. Because of $a_\gamma^+ = \sup\{a_\gamma, \sup\{b^2, q\}\}$ and $\inf\{a_\gamma, \sup\{b^2, q\}\} = b^2 \in F_K$ we may conclude that $f(a_\gamma^+) \leqq \sup\{f(a_\gamma), f(\sup\{b^2, q\})\}$. Thus it is sufficient to prove (by transfinite induction) that $f(a_\alpha) \leqq \sup\{f(a) | a \leqq a_\alpha,\ a \in ALV\}$ for all ordinal numbers α.

1. $f(a_1) = 0 \leqq \sup\{f(a) | a \leqq a_1,\ a \in ALV\}$.

2. If α iş a limit ordinal then we may conclude with the help of condition (B) that $f(a_\alpha) \leqq \sup_{\beta<\alpha} f(a_\beta) \leqq \sup_{\beta<\alpha}(\sup\{f(a) | a \leqq a_\beta, a \in ALV\}) \leqq \sup\{f(a) | a \leqq a_\alpha, a \in ALV\}$ if α is not a limit ordinal then we may assume without loss of generality that $\alpha \geqq 3$ and that $A_c \backslash (A_c^{\alpha-1} \cup \{q\}) \neq \varnothing$. Now the relation $\inf\{a_{\alpha-1}, \sup\{b^2, b^\alpha\}\} = b^2 \in F_K$ implies that $f(a_\alpha) \leqq \sup\{f(a_{\alpha-1}), f(\sup\{b^2, b^\alpha\})\} \leqq \sup\{f(a) | a \leqq a_\alpha,\ a \in ALV\}$. This completes the proof of the theorem.

Some supplementary remarks. 1. If L is a finite lattice then Theorem 4.1 remains valid if we assume $(M, \leqq)$ to be an arbitrary join semilattice. Hence Theorem 4.1 clarifies also the connections between the original concept of Janowitz [6] and Hubert's concept [5]. Let $\mathrm{Hier}_K^0(L)$ be the set of all K-hierarchies which satisfy condition (H0$^+$). We may conclude in particular that $\mathrm{Hier}_K^0(L)\backslash\mathrm{Hier}_K^d(L) \neq \varnothing$ iff $|AL\backslash F_K|>3$. No hierarchy $(H, h) \in \mathrm{Hier}_K^0(L)\backslash\mathrm{Hier}_K^d(L)$ can be constructed within Jardine and Sibson's or the original concept of Janowitz of hierarchical clustering. Hence Theorem 4.1 demonstrates in particular how Hubert's concept of hierarchical clustering extends the concept of Jardine and Sibson and in some sense also the more general concept of Janowitz.

2. We want to emphasize the following remark: The more interesting part of a hierarchy (H, h) is H because it contains the "clusters" one is really interested in. Hence we consider clustering functions $f_1, f_2: L \to M$ and cluster methods $T_1, T_2: \mathrm{Clus}\,(L, M) \to \mathrm{Clus}\,(L, M)$ and define:

$$f_1 \sim_f f_2 \Leftrightarrow H_{f_1} = H_{f_2},$$

$$T_1 \sim_T T_2 \Leftrightarrow T_1(f) \sim_f T_2(f) \text{ for all } f \in \mathrm{Clus}\,(L, M).$$

Generalizing the definitions of Janowitz [9] or Sibson [15] two clustering functions $f_1, f_2: L \to M$ are called *globally order equivalent* iff $f_1(a) \leqq f_1(b)$ is equivalent to $f_2(a) \leqq f_2(b)$ for all $a, b \in L$. Global order equivalence is clearly an equivalence relation on $\mathrm{Clus}\,(L, M)$ and will be denoted by writing $f_1 \sim f_2$.

Furthermore—following Janowitz [9]—two cluster methods $T_1, T_2: \mathrm{Clus}\,(L, M) \to \mathrm{Clus}\,(L, M)$ are said to be *order similar* iff $T_1(f) \sim T_2(f)$ for all $f \in \mathrm{Clus}\,(L, M)$.

It is very easy to see that the following implications hold for arbitrary clustering functions $f_1, f_2, f: L \to M$ and arbitrary cluster methods $T_1, T_2: \mathrm{Clus}\,(L, M) \to \mathrm{Clus}\,(L, M)$:

$$f_1 \sim f_2 \Rightarrow f_1 \sim_f f_2,$$

$$T_1(f) \sim T_2(f) \Rightarrow T_1(f) \sim_T T_2(f).$$

In order to show that the converse implications are generally false let L be the power set of $S := \{0_1, 0_2, 0_3\}$ and let M be the set of nonnegative reals. We define clustering functions $f_1, f_2: L \to M$ by

$$f_1(A) := \begin{cases} 0 & \text{if } |A| \leqq 1, \\ 1 & \text{if } A \in \{\{0_1, 0_2\}, \{0_2, 0_3\}\}, \quad \text{for all } A \in L, \\ 2 & \text{else}, \end{cases}$$

$$f_2(A) := \begin{cases} 0 & \text{if } |A| \leqq 1, \\ 1 & \text{if } A = \{0_1, 0_2\}, \\ 2 & \text{if } A = \{0_2, 0_3\}, \\ 3 & \text{else}, \end{cases} \quad \text{for all } A \in L;$$

and clustering methods $T_1, T_2: \mathrm{Clus}\,(L, M) \to \mathrm{Clus}\,(L, M)$ by $T_1 := \mathrm{id}_{\mathrm{Clus}\,(L,M)}$ and

$$T_2(f)(A) := \begin{cases} f(A) & \text{if } A \neq \{0_2, 0_3\}, \\ f(\{0_1, 0_2\}) & \text{if } A = \{0_2, 0_3\} \text{ and if } f(\{0_1, 0_2\}) \leqq f(\{0_2, 0_3\}) < f(S), \\ f(\{0_2, 0_3\}) & \text{else}, \end{cases}$$

for all $f \in \mathrm{Clus}\,(L, M)$ and all $A \in L$.

The reader may immediately verify that $f_1 \sim_f f_2$ and that $T_1 \sim_T T_2$ but that neither f_1 and f_2 are globally order equivalent nor T_1 and T_2 are order similar.

Thus "$\sim_f$"-equivalent clustering functions and "$\sim_T$"-equivalent cluster methods generalize naturally globally order equivalent clustering functions and order similar cluster methods respectively. Within a concept of hierarchical clustering which emphasizes the "clusters" or "hierarchies" in which one is really interested they seem to be the more natural equivalence relations on Clus (L, M) and on the set of cluster methods T: Clus $(L, M) \to$ Clus (L, M) respectively.

Generalizations of *weakly order equivalent* clustering functions or *weakly order similar* cluster methods (cf. Janowitz [9]) may be obtained by defining:

$$f_1 \sim_w f_2 \Leftrightarrow H_{f_1} \backslash \{a \in H_{f_1} | h_{f_1}(a) = 0\} = H_{f_2} \backslash \{a \in H_{f_2} | h_{f_2}(a) = 0\},$$

$$T_1 \sim_W T_2 \Leftrightarrow T_1(f) \sim_w T_2(f) \text{ for all } f \in \text{Clus } (L, M).$$

3. In [5] Hubert studied also "clustering functions" which are not globally monotone on L but which are monotone on some order filter F_K of L. In this case we replace L by $F_K \cup \{0\}$ and call two "clustering functions" $f_1, f_2 : L \to M K$-equivalent iff $f_1|_{F_K} = f_2|_{F_K}$. Let Clus $(L, M)_{/\sim K}$ be the set of all equivalence classes of K-equivalent "clustering functions" then it is easy to see that there exists a canonical order isomorphism: Clus $(L, M)_{/\sim K} \to$ Clus $(F_K \cup \{0\}, M)$.

Hence we also studied within our model of hierarchical clustering classes of K-equivalent k-clustering functions in the sense of Hubert.

REFERENCES

[1] T. S. Blyth and M. F. Janowitz, *Residuation Theory*, Pergamon Press, London, 1972.

[2] H. Bock, *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Göttingen, 1974.

[3] L. Hubert, *Some extensions of Johnson's hierarchical clustering algorithms*, Psychometrika, 37 (1972), pp. 262–274.

[4] ———, *Some applications of graph theory to clustering*, Psychometrika, 39 (1974), pp. 283–309.

[5] ———, *A set theoretical approach to the problem of hierarchical clustering*, J. Math. Psych., 15 (1977), pp. 70–88.

[6] M. F. Janowitz, *An order theoretic model for cluster analysis*, SIAM J. Appl. Math., 34 (1978), pp. 55–72.

[7] ———, *Semiflat L-cluster methods*, Discr. Math., 21 (1978), pp. 47–60.

[8] ———, *Monotone equivariant cluster methods*, SIAM J. Appl. Math., 37 (1979), pp. 148–165.

[9] ———, *Preservation of order equivalence*, J. Math. Psych., 20 (1979), pp. 78–88.

[10] ———, *Applications of the theory of partially ordered sets to cluster analysis*, Universal Algebra and Applications, Banach Center Publ., 9 (1980), pp. 305–319.

[11] N. Jardine and R. Sibson, *A model for taxonomy*, Math. Biosci., 2 (1968), pp. 465–482.

[12] ———, *Mathematical Taxonomy*, John Wiley, New York, 1971.

[13] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrika, 32 (1967), pp. 241–254.

[14] D. W. Matula, *Graph theoretic techniques for cluster analysis algorithms*, Classification and Clustering, Van Ryzin, ed., Academic Press, New York, pp. 95–129.

[15] R. Sibson, *Order invariant methods for data analysis*, J. Royal Stat. Soc. Ser., B, 34 (1972), pp. 311–349.

[16] J. H. Ward, *Hierarchical grouping to optimize an objective function*, J. Amer. Stat. Assoc., 58 (1963), pp. 236–244.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0014 $01.25/0

A NOTE ON OPTIMAL AND SUBOPTIMAL DIGRAPH REALIZATIONS OF QUASIDISTANCE MATRICES*

J. M. S. SIMÕES-PEREIRA†

Abstract. Dropping the symmetry axiom from the definition of a distance, we obtain the definition of a quasidistance. Matrices whose entries are quasidistances among a finite set of points are called quasidistance matrices. A digraph with valued arcs realizes a quasidistance matrix when, for a certain subset $\{1, \cdots, n\}$ of its vertices, the shortest path from an arbitrary vertex i to an arbitrary vertex j has a length equal to the entry d_{ij} of the matrix.

We give the optimal and suboptimal digraph realizations of quasidistance matrices of order 2 and 3 (except for a special class); these are realizations whose total length is as small as possible. We characterize quasidistance matrices which have optimal realizations. We prove several results relating quasidistance matrices to their principal submatrices and to matrices obtained by decreasing all nondiagonal entries in a column or in a row. We investigate and characterize a class of quasidistance matrices, to be called shrinkable, which by their definition and properties may be considered the analogues of tree-realizable (symmetric) distance matrices.

1. Introduction. Quasidistances are "nonsymmetric distances". In operations research, for instance, the following are frequently used quasidistances: mileages on a road network with one-way streets; travel times, in particular flight times which may be shorter in one direction than in the opposite one; transportation costs which may be lower downhill; transmission costs through a communications network; international telephone or postal rates; airline fares and so on.

A quasidistance n-matrix is a nonnegative, square matrix D of order n with entries d_{ij} such that, for $i, j, k \in \{1, \cdots, n\}$, $d_{ii} = 0$ and $d_{ij} \leqq d_{ik} + d_{kj}$. If $d_{ij} = d_{ik} + d_{kj}$ for some k distinct from i and j, d_{ij} is called composite; otherwise it is called basic.

Let $G = (W, E)$ be a digraph, W and E or, more explicitly, $W(G)$ and $E(G)$, its vertex and arc sets, respectively; let $V \subseteq W$ and $|V| = n$. Consider a function $f: E \rightarrow R^+$, where R^+ is the set of positive real numbers. This function assigns a length or weight to each arc of G. For any $i, j \in W$, let $d(i, j)$ be the minimum value among sums $\sum f(e)$ taken over any directed path $P(i, j)$ from i to j. The digraph G is a realization of the matrix D if, for some V, $d(i, j) = d_{ij}$ for $i, j \in \{1, \cdots, n\}$; the vertices in V are called external, those in $W - V$ internal. Trivially, internal vertices can be required not to be sinks nor sources and not to have degree two. Denoting by $T(G)$ the sum $\sum f(e)$ over $E(G)$, a realization G is optimal if $T(G)$ is minimal among all realizations of D.

The analogues of these concepts for (symmetric) distance matrices, and their applications, have been investigated (see [1]–[20]); not so for quasidistances. Results for the symmetric and nonsymmetric cases are not always similar: for example, whereas a distance matrix always has an optimal graph realization [11], a quasidistance matrix only exceptionally has one. In § 2, we characterize these exceptional quasidistance matrices.

A first pioneering effort on digraph realizations of quasidistance matrices with integer entries is due to Imrich [8]. Another pioneering effort is due to Patrinos and

* Received by the editors November 9, 1982, and in revised form March 8, 1983. This research was supported by the Scientific Affairs Division of the North Atlantic Treaty Organization under grant RG 167.81 and by a grant from the City University of New York PSC/CUNY Research Award Program. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26–29, 1982.

† Department of Computer Science, Hunter College and The Graduate School of The City University of New York, New York, New York, 10021.

Hakimi [14] who allowed quasidistances to be negative; this assumption, which makes proofs easier, is, however, somewhat unnatural in many applications. For example, to realize the matrix

$$\begin{bmatrix} 0 & 4 & 4 & 5 \\ 4 & 0 & 4 & 5 \\ 4 & 4 & 0 & 5 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

their algorithm starts with the leading principal submatrix of order 3 whose realization is naturally chosen as the digraph in Fig. 1.1. Adding vertex 4 requires an arc $(4, u)$ of length 3 and an arc $(u, 4)$ of length -1.

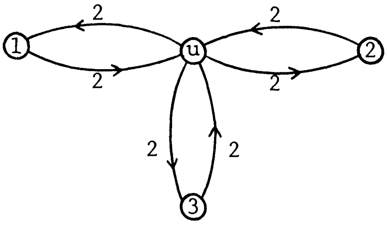


FIG. 1.1

In § 3, we generalize a compactification and reduction technique introduced by Zaretzkii [20] for integer (symmetric) distances; this technique allows us to avoid nonpositive lengths, as long as the originally given matrix is nonnegative.

In the absence of optimal realizations, suboptimal ones become important. Let τ be the infimum of the set of the total weight values for all realizations of D; we call G a suboptimal realization of D if, given $\varepsilon \in R^+$, the arcs of G may be reassigned (positive) lengths such that $\tau < T(G) < \tau + \varepsilon$. For obvious reasons, we will then say that $T(G)$ does not significantly differ from τ. More generally, we say that two variables do not significantly differ when their difference is arbitrarily small. For instance, if, given $\varepsilon \in R^+$, the arcs of G' and the arcs of G'' may be reassigned (positive) lengths such that $\tau < T(G') < \tau + \varepsilon$ and $\tau < T(G'') < \tau + \varepsilon$, then we say that $T(G')$ and $T(G'')$ do not significantly differ. Total weights of suboptimal realizations of the same matrix are not significantly different; neither are they significantly different from the value of their corresponding τ.

We will refer to optimal and suboptimal realizations (as the case may be) as best realizations. Best realizations of 3-matrices, except for one special class, will be given in § 4.

In § 5, we study some properties of submatrices of quasidistance n-matrices and in § 6, we characterize shrinkable quasidistance n-matrices, a result which is the nonnegative analogue of the Patrinos–Hakimi characterization of hypertree-realizable quasidistance matrices.

2. Optimal and suboptimal realizations.

First we prove:

THEOREM 2.1. *A quasidistance 2-matrix D has a suboptimal but no optimal realization.*

Proof. Let

$$D = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix},$$

where $a \leqq b$, without loss of generality. The digraph of Fig. 2.3 realizes D with a total weight $T = b + 2\varepsilon$. This realization is suboptimal. In fact, since $d_{21} = b$, we have $T \geqq b$ in any realization of D. A path $P(1, 2)$ must exist, hence one arc leaving vertex 1 and another (or the same) entering vertex 2 must exist. These arcs can not be in the path which realizes d_{21}. Since lengths are positive, we have $T > b$, hence no optimal realization can exist. This completes the proof.

Loosely speaking, the realization in Fig. 2.1 yields the realization in Fig. 2.3 by pasting or attaching most of $P(1, 2)$ into $P(2, 1)$, as visualized in Fig. 2.2.

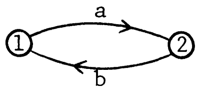


FIG. 2.1

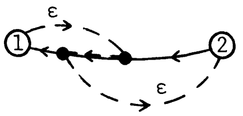


FIG. 2.2

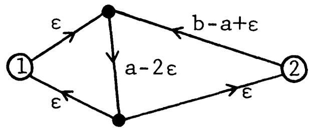


FIG. 2.3

LEMMA 2.2. *A quasidistance n-matrix D has at least n basic entries.*

Proof. Let i be fixed. If d_{ip} is not basic, then there exists q such that $d_{ip} = d_{iq} + d_{qp}$; if d_{iq} is not basic, repeat this argument with q in the role of p. It follows that, for every i, there is at least one j such that d_{ij} is basic, which proves the lemma.

For $h \geqq 3$, we have:

THEOREM 2.3. *A quasidistance n-matrix D has an optimal realization if and only if D can be realized by a simple, directed cycle, or, equivalently, D has n basic entries.*

Proof. Let G be an optimal realization of D. Since all distances are defined, each external vertex of G has both indegree and outdegree at least one. The digraph G cannot have an internal or external vertex of indegree greater than one; in fact, if L_1 and L_2 are the lengths of two arcs incident to the vertex v as in Fig. 2.4, then, after replacing them by the configuration in Fig. 2.5, the total length of G decreases by ε, where $\varepsilon < \min\{L_1, L_2\}$, and D is still realized. Similarly, G cannot have a vertex of outdegree greater than one. Hence G is a simple, directed cycle.

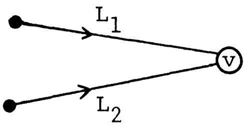


FIG. 2.4

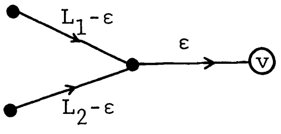


FIG. 2.5

Conversely, let G be a simple, directed cycle which realizes a matrix D. Without loss of generality, let $(1, 2)$, $(2, 3)$, $\cdots$, $(n-1, n)$, $(n, 1)$ be the arcs of the cycle. Obviously, $T(G) = d_{1n} + d_{n1}$. Now, any realization of D needs a path $P(1, n)$ of length d_{1n} formed by pairwise disjoint subpaths $P(i, i+1)$ of lengths $d_{i,i+1}$ for $i = 1, \cdots, n-1$. This is proved by induction: $P(1, 2)$ of length d_{12} is (obviously) needed and $P(2, 3)$ can not intersect $P(1, 2)$ without contradicting the equality $d_{13} = d_{12} + d_{23}$ which holds because we assume that G realizes D. Now, if $P(1, k)$ is needed, then $P(k, k+1)$ can not intersect $P(1, k)$ without contradicting the equality $d_{1,k+1} = d_{1k} + d_{k,k+1}$. Similarly, a path $P(n, 1)$ of length d_{n1} must exist and will be disjoint from $P(1, n)$, otherwise we contradict at least one of the equalities $d_{21} = d_{2n} + d_{n1}$ or $d_{n2} = d_{n1} + d_{12}$. The total length is thus at least the length of G which completes the proof of the theorem.

Another useful result is the following one:

THEOREM 2.4. *Let D and D' be quasidistance matrices, with D obtained from D' by adding a constant K to all off-diagonal entries in column i (or in row i) of D'. If*

G' is a best realization of D', then $T(G)$ can not be significantly smaller than $T(G')$ for any realization G of D.

Proof. The arguments being similar, we only consider the case where K is added to the entries of a column.

With Q a positive (not arbitrarily small) constant, suppose that $T(G) < T(G') - Q$. We will show how to obtain from G a realization of D' with total weight less than $T(G') - Q + \varepsilon$, a contradiction of the fact that G' is best.

Consider all paths coming to vertex i in G. Trivially, we may suppose that all these paths share the last arc, this means that the indegree of i is one. If the length of this last arc is at least $K + \varepsilon$, then, by decreasing it to ε, we obtain a realization of D' whose total weight $T(G) - K$ is significantly smaller than $T(G')$ and the statement of the theorem holds. If, now, the length of this last arc is at most K, then we will use a technique which modifies G without increasing $T(G)$ by significantly more than K and which assigns to the aforementioned last arc a length of at least $K + \varepsilon$. For this purpose, let $d_{ai} = \min_{j \neq i} \{d_{ji}\}$ and let $P(a, i)$ be a path of length d_{ai}, which therefore may only contain internal vertices (Fig. 2.6). Clearly, $d_{ai} \geqq K + \varepsilon$.

It is useful to distinguish three types of internal vertices on a path: when the arcs of the path are removed, vertices of type 1 become sources, those of type 2 transmitters, those of type 3 sinks (Fig. 2.6). Trivially, we may suppose these sources, sinks and transmitters have halfdegrees at most one (Fig. 2.7).

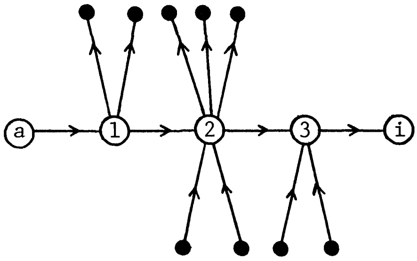


FIG. 2.6

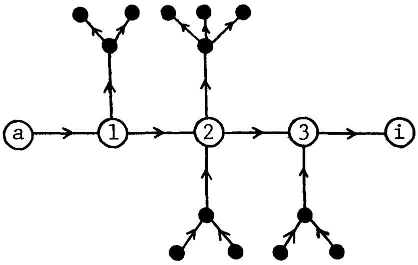


FIG. 2.7

Loosely speaking, we now modify G by rolling the internal vertices on $P(a, i)$ along $P(a, i)$ and away from i so that the length of the last arc of $P(a, i)$ becomes at least $K + \varepsilon$. To visualize the operation, think of attaching or pasting into $P(a, i)$ the arcs coming to it and detaching those coming from it, as in Figs. 2.8–2.10. We can

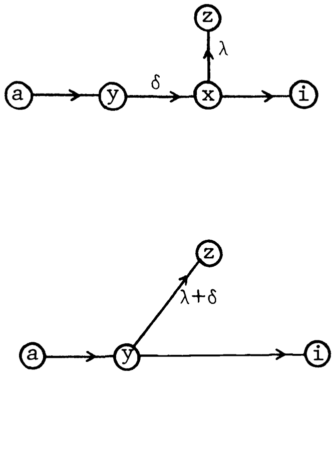


FIG. 2.8

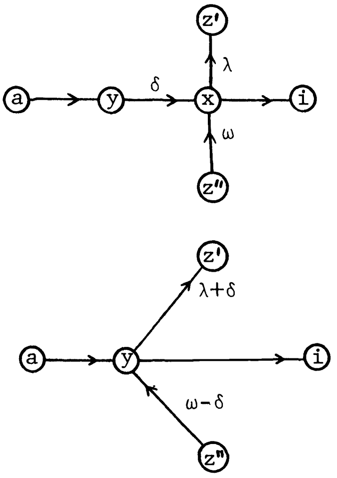


FIG. 2.9

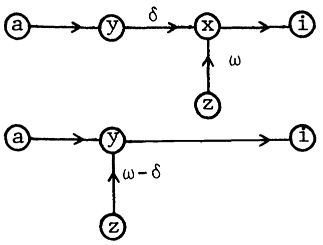


FIG. 2.10

achieve a last arc in $P(a, i)$ of length at least $K+\varepsilon$ without increasing $T(G)$ by more than K. In fact, for a type 1 vertex, rolling it back by δ means detaching an arc from $P(a, i)$, which increases $T(G)$ by δ (Fig. 2.8). For a vertex of type 2, rolling it back by δ means detaching an arc from and attaching an arc to $P(a, i)$, which leaves $T(G)$ unchanged (Fig. 2.9). For a vertex of type 3, rolling it back by δ means attaching an arc to $P(a, i)$, which decreases $T(G)$ by δ (Fig. 2.10).

Note that we may roll back a type 1 vertex until it coincides with the vertex preceding it, which retains its own type if it is of type 1 or 2, and becomes of type 2 if it is of type 3; a type 2 or 3 vertex may not become coincident with the vertex preceding it but may become arbitrarily close to it. Therefore, if nothing precludes these operations, then they allow us to increase the length of the last arc of $P(a, i)$ to $K+\varepsilon$ without increasing $T(G)$ by significantly more than the amount K.

Now, insufficient length of an arc coming to $P(a, i)$, say (x', x), is the only factor which may preclude the immediate execution of these operations (Fig. 2.11). Note, however, that by the minimality of d_{ai}, we have $d(b, x) \geqq d(a, x)$ for any b distinct from a and i. If $P(b, x)$ had no vertices, then x could be rolled back to a position arbitrarily close to a; otherwise, we have to roll back the internal vertices on $P(b, x)$ to positions as far from x on $P(b, x)$ as we need. This is always possible, the only nontrivial case being a type 1 vertex, say x', on $P(b, x)$. In this case, depicted in Fig. 2.12, attaching $P(b, x)$ to $P(a, x)$ for, say, a length $\delta'+\delta$, where δ is the length of (x', x), requires a previous rolling back of x' on $P(b, x)$ by the amount $\delta'+\varepsilon$. This operation increases $T(G)$ by only ε, which is not a significant increase.

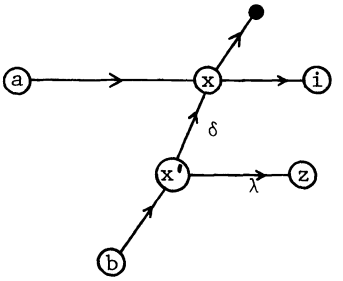


FIG. 2.11

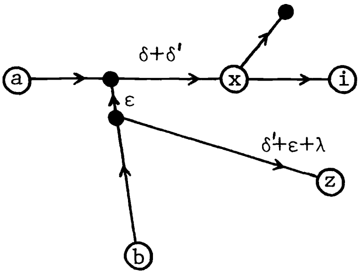


FIG. 2.12

The argument can be repeated with all vertices $a, b, \cdots$ ordered according to their increasing distances to vertex i. The proof of the theorem is thus completed.

3. Compactifications, reductions and shrinkability. Among (symmetric) distance matrices, those which are tree-realizable have been the most investigated. The compactification of a distance matrix has been defined, for instance, in [17], where we see that the compactification of a pendant index leads to a new matrix with a pair of equal rows (and, by symmetry, a pair of equal columns). By deleting one of these equal rows and one of the equal columns, we obtain a new distance matrix whose order is one unit lower. Obviously, we may define tree-realizable distance matrices as those which yield the 1-matrix [0] by some sequence of pendant index compactifications plus deletion of equal rows and columns.

We introduce here a similar compactification for quasidistance matrices and we show that those quasidistance matrices which can be compactified, in a way to be made precise, to the 1-matrix [0], have several properties similar to well known properties of tree-realizable matrices. We call such quasidistance matrices shrinkable.

Given a quasidistance n-matrix D, let $r_i = \min\{(d_{ai}+d_{ib}-d_{ab})\}$ with $a, b \in \{1, \cdots, n\}$, a and b not necessarily distinct but at least one of them distinct from i. Let $0 \leqq r \leqq r_i$ and let f and t be nonnegative reals chosen so that $f+t=r$, $f \leqq \min_{a\neq i}\{d_{ia}\}$, $t \leqq \min_{a \neq i}\{d_{ai}\}$. Denote by D' a quasidistance matrix obtained from D by subtracting f from all nondiagonal entries in row i and t from all nondiagonal entries in column i. A realization G of D can be obtained from a realization G' of D' by adding one vertex i to G' and two arcs: (i', i) of length t and (i, i') of length f. The operation which leads from D to D' or from G to G' is called a compactification with respect to i by the amount r or by the amounts f, t.

THEOREM 3.1. *In a quasidistance matrix D, if $r_i = d_{ij}+d_{ji}$ for some j, then, for each $p \neq j$ we have $d_{pi} = d_{pj}+d_{ji}$ and $d_{ip} = d_{ij}+d_{jp}$. Moreover, setting $r = r_i$ implies setting $f = d_{ij}$ and $t = d_{ji}$.*

Proof. By definition of r_i we have, for each p, $d_{ij}+d_{ji} \leqq d_{pi}+d_{ij}-d_{pj}$, hence $d_{pi} = d_{pj}+d_{ji}$. Similarly, $d_{ij}+d_{ji} \leqq d_{ji}+d_{ip}-d_{jp}$ yields $d_{ip} = d_{ij}+d_{jp}$. As a consequence, the definitions of f and t make the truth of the second statement obvious too.

When the condition of this theorem holds, we say that i is a *pendant index* (from j) and a compactification by r_i with respect to i yields a matrix where rows i and j are equal and columns i and j are equal too; removing row i and column i, we obtain a matrix of lower order which is called a *reduction* of D.

A quasidistance n-matrix D is called shrinkable when a sequence of matrices $D \equiv D_1, \cdots, D_z \equiv [0]$ with $z \geqq n$ exists, each one being either a compactification or a reduction of the preceding one. Such a sequence is called a shrinking sequence. By convention we also say that an identically zero n-matrix is shrinkable.

Remark 3.2. Shrinking sequences are not unique. To obtain one, we must have, in each successive matrix (except for the last one), for at least one i, $r_i > 0$. Now, if a sequence of choices of the index i leads to a shrinking sequence, then any other possible choice of i at each stage leads also to a shrinking sequence. This follows from the definitions: in successive matrices, say D' and D'', the values r''_q and r'_q are equal and given by homologous entries except for $q = i$, where $r''_i = r'_i - r'_i = 0$. Moreover, those indices q which become pendant in D'', i.e., for which $r''_q = d''_{iq}+d''_{qi}$ and $r'_q = d'_{aq}+d'_{qb}-d'_{ab}$ $(a \neq b)$, will be removed and for them, due to Theorem 3.1, r_q will remain unchanged since they become pendant until their removal.

Remark 3.3. If D is shrinkable, then all matrices in any one of its shrinking sequences are shrinkable. If D yields a shrinkable matrix by compactification or reduction, then D is shrinkable too.

Remark 3.4. Quasidistance 2-matrices are shrinkable.

4. Quasidistance 3-matrices. By Theorem 2.3, a nonshrinkable 3-matrix with three basic entries has an optimal realization whose total length is the sum of the basic entries (Fig. 4.1). To find realizations of other 3-matrices, the following lemma is useful.

LEMMA 4.1. *A quasidistance 3-matrix D is shrinkable if and only if*

$$(1) \qquad d_{ij}+d_{jk}+d_{ki} = d_{ik}+d_{kj}+d_{ji}.$$

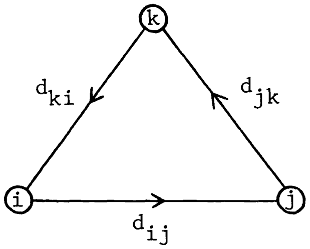


FIG. 4.1

Proof. Suppose equality (1) holds. We distinguish three cases:

Case 1. For one of the indices, say i,

$$0 \leqq r_i = d_{ji} + d_{ij} = d_{ki} + d_{ik}. \tag{2}$$

By Theorem 3.1, (2) implies $d_{jk} = d_{kj} = 0$, $d_{ik} = d_{ij}$, $d_{ki} = d_{ji}$. As a consequence, $r_i = 0$ means that D is identically zero; if $r_i > 0$, then D has a reduction of order 2. Hence D is shrinkable.

Case 2. Condition (2) does not occur but, for one of the indices, say i,

$$0 \leqq r_i = d_{ji} + d_{ij} < d_{ki} + d_{ik}. \tag{3}$$

First, $r_i = 0$ means that $d_{ji} = d_{ij} = 0$ and, by Theorem 3.1, (2) would occur for r_k. Consider $r_i > 0$. Again using Theorem 3.1, we reduce D to a 2-matrix, hence, by Remarks 3.3 and 3.4, D is shrinkable.

Case 3. Conditions (2) and (3) do not occur but, for one of the indices, say i,

$$0 \leqq r_i = d_{ji} + d_{ik} - d_{jk} = d_{ki} + d_{ij} - d_{kj} < \min\{d_{ji} + d_{ij}, d_{ki} + d_{ik}\}. \tag{4}$$

First, $r_i = 0$ means that $d_{jk} = d_{ji} + d_{ik}$ and $d_{kj} = d_{ki} + d_{ij}$ which implies that (3) occurs for r_j and r_k. Now, since $r_i > 0$, let ε be positive and arbitrarily small and choose $t = \min\{d_{ji}, d_{ki}, r_i\} - \varepsilon, f = r_i - t$. By (4) we have

$$t + f < \min\{d_{ji} + d_{ij}, d_{ki} + d_{ik}\}. \tag{5}$$

Since d_{jk} and d_{kj} are positive (otherwise $r_k = r_j = 0$ and (3) occurs), we have, by (4) again,

$$t + f < \min\{d_{ji} + d_{ik}, d_{ki} + d_{ij}\}. \tag{6}$$

The strict inequalities (5) and (6) and the positivity of the nondiagonal entries of D imply that, for ε sufficiently small, $f \leqq \min\{d_{ij}, d_{ik}, r_i\}$. (Obviously, we can also choose $f = \min\{d_{ij}, d_{ik}, r_i\} - \varepsilon$ and show that $t = r_i - f \leqq \min\{d_{ji}, d_{ki}, r_i\}$.)

The compactification of D by the amounts f, t yields a matrix D' with $r_i' = 0$ and, thus, in D', (3) occurs for j and k. D is therefore shrinkable.

Suppose now that (1) does not hold. Without loss of generality, let

$$d_{ij} + d_{jk} + d_{ki} < d_{ik} + d_{kj} + d_{ji}. \tag{7}$$

Equivalently, $d_{ki} + d_{ij} - d_{kj} < d_{ji} + d_{ik} - d_{jk}$. Moreover, if, say, $r_i = d_{ki} + d_{ik}$ then, by Theorem 3.1, $d_{ij} = d_{ik} + d_{kj}$ and $d_{ji} = d_{jk} + d_{ki}$, which contradicts (7). Similarly, if $r_i = d_{ji} + d_{ij}$. It follows that

$$r_i = d_{ki} + d_{ij} - d_{kj} < \min\{d_{ji} + d_{ik} - d_{jk}, d_{ki} + d_{ik}, d_{ji} + d_{ij}\}.$$

A similar argument shows that

$$r_j = d_{ij} + d_{jk} - d_{ik} < \min\{d_{kj} + d_{ji} - d_{ki},\ d_{ij} + d_{ji},\ d_{kj} + d_{jk}\},$$

$$r_k = d_{jk} + d_{ki} - d_{ji} < \min\{d_{ik} + d_{kj} - d_{ij},\ d_{ik} + d_{ki},\ d_{jk} + d_{kj}\}.$$

If $r_i = r_j = r_k = 0$, then D is clearly nonshrinkable. Otherwise, since compactifications do not destroy strict inequality in (7), at most three successive compactifications of D yield a new 3-matrix with $r_i = r_j = r_k = 0$. By Remark 3.3, D is nonshrinkable. This completes the proof of the lemma.

THEOREM 4.2. *A shrinkable quasidistance* 3*-matrix D has a suboptimal realization of total length $T = \max\{d_{ij} : i, j = 1, 2, 3\} + \varepsilon$.*

Proof. We exhibit the realization for the distinct cases considered in Lemma 4.1. Since we exclude zero lengths, Case 1 is realized by the digraph in Fig. 4.2, where one vertex is i and the other represents both j and k.

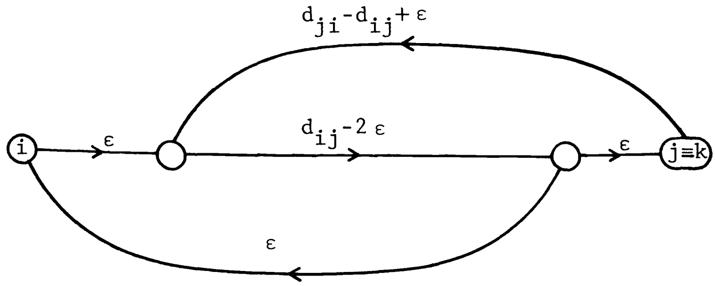


FIG. 4.2

Case 2 implies $d_{ik} = d_{ij} + d_{jk}$ and $d_{ki} = d_{kj} + d_{ji}$. Without loss of generality, let $d_{ik} \geqq d_{ki}$, this means, $d_{ij} + d_{jk} \geqq d_{kj} + d_{ji}$. We distinguish three subcases:

Subcase A. We have $d_{ji} - d_{jk} = \lambda > 0$ which implies $\gamma = d_{ij} - d_{kj} \geqq \lambda$. The digraph whose arcs are the solid lines of Fig. 4.3 is the suboptimal realization.

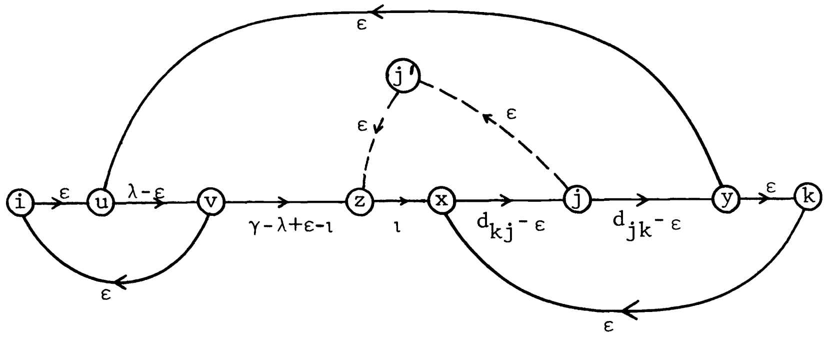


FIG. 4.3

Subcase B. We have $d_{kj} - d_{ij} = \lambda > 0$ which implies $\gamma = d_{jk} - d_{ji} \geqq \lambda$. The digraph whose arcs are the solid lines of Fig. 4.4 is the suboptimal realization.

Subcase C. We have $d_{ji} \leqq d_{jk}$ and $d_{kj} \leqq d_{ij}$. The digraph whose arcs are the solid lines of Fig. 4.5 is the suboptimal realization.

In Case 3, as seen in the proof of Lemma 4.1, we may suppose that r_i, r_j, r_k are all positive. Without loss of generality, let d_{ik} be the largest entry of D. Let D^* be a compactification of D with respect to j where $r_j^* = 0$. We distinguish three subcases:

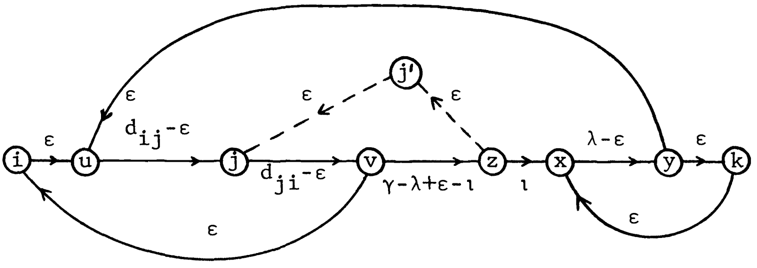


FIG. 4.4

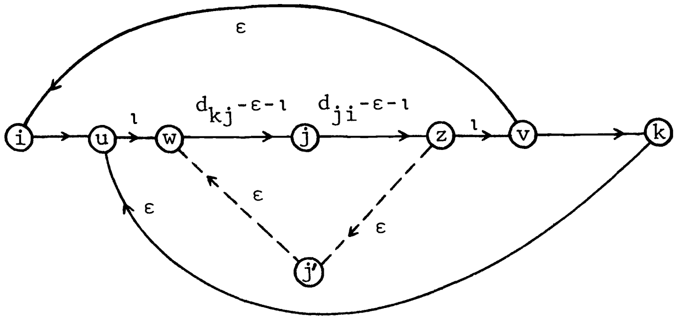


FIG. 4.5

Subcase A′. Let $r_j \leqq \min\{d_{ji}, d_{jk}\}$. Choose $f = r_j - \varepsilon$, $t = \varepsilon$. The digraph of Subcase A realizes D^*. To realize D, add the dashed lines of Fig. 4.3, this means, let z be a new vertex on the path $P(i, j)$ such that $d(z, j) = f - \varepsilon$ and add a new vertex j' and two arcs of length ε, (j, j') and (j', z). Since d_{ik} is the largest entry of D, f is such that z exists in $P(i, j)$, otherwise $d_{jk} > d_{ik}$; in fact, v precedes z in $P(i, j)$, otherwise $d_{ji} \geqq \varepsilon + d(v, j) + d(j, y) + \varepsilon + \lambda - \varepsilon + \varepsilon \equiv \varepsilon + d_{ik}$.

Subcase B′. Let $r_j \leqq \min\{d_{ij}, d_{kj}\}$. Choose $t = r_j - \varepsilon$, $f = \varepsilon$. The digraph of Subcase B realizes D^*. To realize D, add the dashed lines of Fig. 4.4, this means, let z be a new vertex on the path $P(j, k)$ such that $d(j, z) = t - \varepsilon$; add a new vertex j' and two arcs of length ε, (j', j) and (z, j'). The maximality of d_{ik} implies here too, in a similar way, that z exists and precedes x in $P(j, k)$.

Subcase C′. Let $r_j > \min\{d_{ij}, d_{kj}\}$ and $r_j > \min\{d_{ji}, d_{jk}\}$. It follows that $\min\{d_{ij}, d_{kj}\} = d_{kj}$ and $\min\{d_{ji}, d_{jk}\} = d_{ji}$. In fact, $r_j = d_{ij} + d_{jk} - d_{ik} > \min\{d_{ij}, d_{kj}\} = d_{ij}$ implies $d_{jk} - d_{ik} > 0$, which contradicts the maximality of d_{ik}. Similarly, $r_j = d_{ij} + d_{jk} - d_{ik} > \min\{d_{ji}, d_{jk}\} = d_{jk}$ implies $d_{ij} - d_{ik} > 0$, the same contradiction. Hence $d_{kj} < d_{ij}$ and $d_{ji} < d_{jk}$; after compactification, $d^*_{kj} < d^*_{ij}$ and $d^*_{ji} < d^*_{jk}$. The digraph of Subcase C realizes D^*. To realize D, add the dashed lines of Fig. 4.5, this means, let w and z be two new vertices on $P(i, j)$ and $P(j, k)$, respectively, such that $d(w, j) = f - \varepsilon$ and $d(j, z) = t - \varepsilon$; add a new vertex j' and two arcs of length ε, (z, j') and (j', w). Clearly, since $r_j < d_{kj} + d_{ji}$, we may choose $f \leqq d_{ji}$ and $t \leqq d_{kj}$ so that $f + t = r_j$.

Since the realizations of D in these three subcases are obviously suboptimal, the proof of the theorem is completed.

Now let D be nonshrinkable and D' obtained from D by three compactifications by the amounts r_i, r_j and r_k, respectively. The digraph of Fig. 4.6 realizes D, provided that the subdigraph spanned by the vertices i', j', k' realizes D'. Such a realization of D will be called a *straightforward realization* and this subdigraph will henceforth be called the *core cycle* of the realization. Let $\Delta = d(i', j') + d(j', k') + d(k', i')$ be the

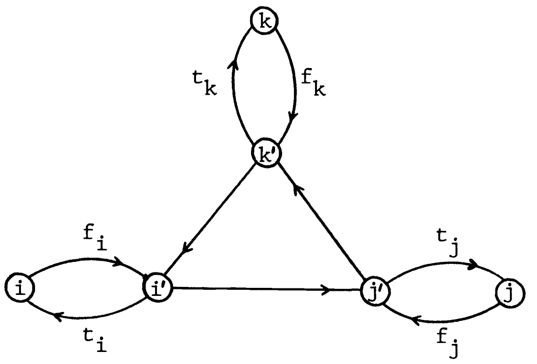


FIG. 4.6

length of the core cycle. It follows easily from the definitions that, since different choices for the values t_i, t_j, t_k, f_i, f_j, f_k are possible, the lengths $d(i', j')$, $d(j', k')$, $d(k', i')$ may change; their sum Δ is however invariant.

To obtain a better realization of D, replace each 2-cycle as suggested by Figs. 2.1, 2.2 and 2.3; i.e. "paste" the arcs of the 2-cycles into the arcs of the core cycle. If $t_j + f_k \leqq d(j', k')$, $t_k + f_i \leqq d(k', i')$, $t_i + f_j \leqq d(i', j')$, then there is a suboptimal realization with total weight $T = \Delta + \varepsilon$. Figure 4.7 depicts this realization with dashed lines showing how the arcs were "pasted".

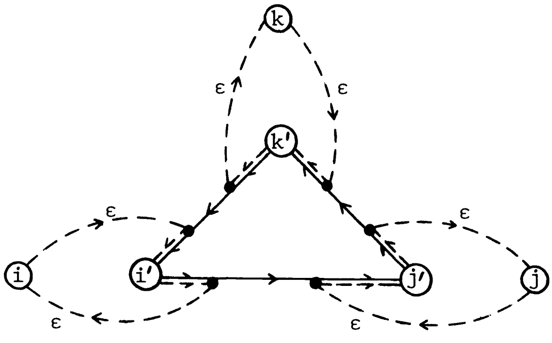


FIG. 4.7

These remarks and Theorem 2.4 immediately yield the following result:

THEOREM 4.3. *If* $\max\{d_{ij}: i, j = 1, 2, 3\} < \Delta$, *then there is a suboptimal realization* G *of* D *with* $T(G) = \Delta + \varepsilon$.

We also have:

THEOREM 4.4. *If* $\max\{d_{ij}: i, j = 1, 2, 3\} > \Delta$ *and* D *has one or two composite entries, then there is a suboptimal realization* G *of* D *with* $T(G) = \max\{d_{ij}: i, j = 1, 2, 3\} + \varepsilon$.

Proof. Little more than an outline is needed; the figures are self-explanatory and the reader may fill in the calculations.

First suppose that D has two composite entries, or, equivalently, $r_j = r_k = 0$. Without loss of generality, let d_{ik} be the largest entry of D. Figures 4.9 and 4.10 visualize how to obtain from a straightforward realization (Fig. 4.8) a suboptimal one. (Note that, if $d(y, i) = \varepsilon$, the last step is redundant.)

Now suppose D has one composite entry or, equivalently, $r_j = 0$. Figure 4.11 depicts a straightforward realization. Distinct cases arise when the largest entry of D is d_{kj} (Case 1), d_{ki} (Case 2), d_{ik} (Case 3). The case where the largest entry of D is d_{ji} is similar to Case 1.

Case 1. Adjust lengths of (j, k') and (k', i') so that f_k is as small as possible (Figs. 4.12 and 4.13). Since $d_{kj} > d_{ki}$, (i', i) fits into (i, j). Since $d_{kj} > d_{ik}$, $f_i' + t_k' < f_k'$. We can therefore attach (i, x) and (y, k) to (k, z) as in Fig. 4.14.

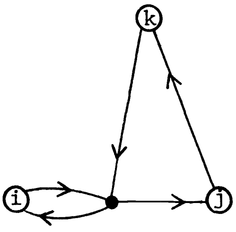


FIG. 4.8

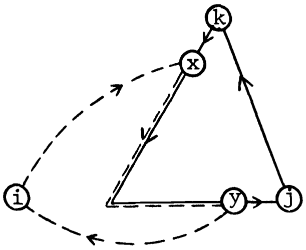


FIG. 4.9

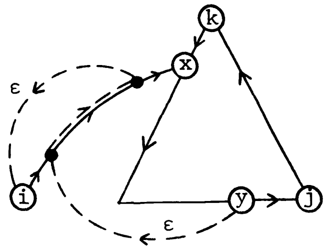


FIG. 4.10

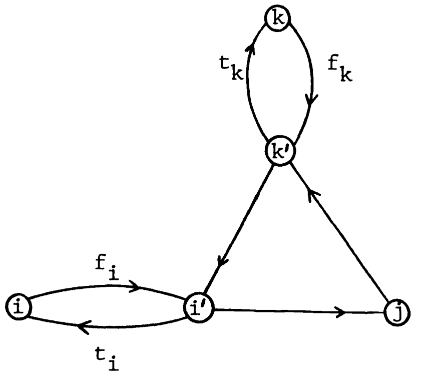


FIG. 4.11

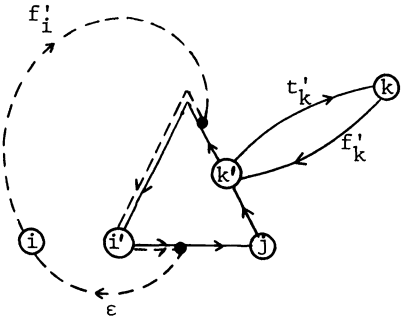


FIG. 4.12

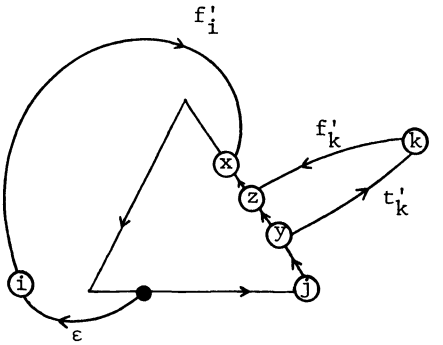


FIG. 4.13

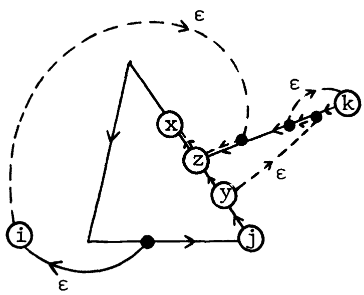


FIG. 4.14

Case 2. First obtain the digraph of Fig. 4.15. By the hypothesis of this case, $d(j, y) = \varepsilon$ and $d(x, j) = \varepsilon$. Then, if $f_i + t_k > d(k', i')$, we obtain the digraph in Fig. 4.16.

Case 3. First obtain the digraph of Fig. 4.15. By the hypothesis of this case, $d(u, v) = \varepsilon$. Since $d_{ki} < d_{ik}$, $t_i' + f_k' < f_i' + t_k'$. Loosely speaking, we may displace u and v along the core cycle, while keeping $d(u, v) = \varepsilon$, as long as t_k' and f_i' remain positive.

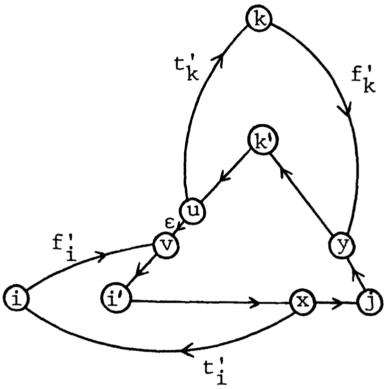


FIG. 4.15

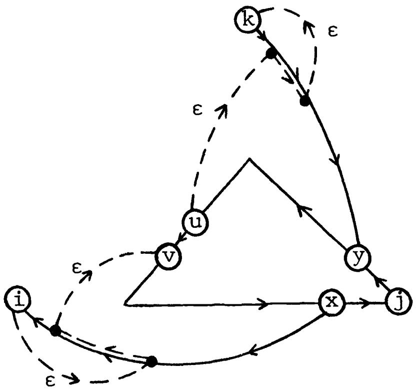


FIG. 4.16

Now, if $t_i' \leqq f_i'$ and $f_k' \leqq t_k'$, we obtain the digraph in Fig. 4.17. Otherwise, if even with $d(u, j) = 2\varepsilon$ we have $t_i' > f_i'$, then we obtain the digraph in Fig. 4.18. This completes the proof of Theorem 4.4.

Only for a nonshrinkable 3-matrix D which has no composite entry and where Δ is smaller than the largest entry has no realization been proved to be best.

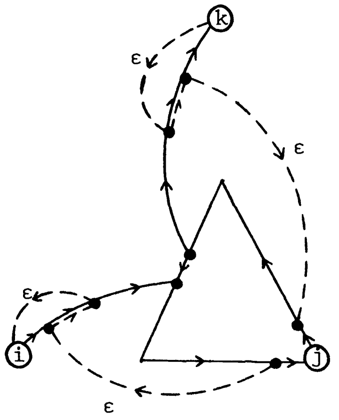


FIG. 4.17

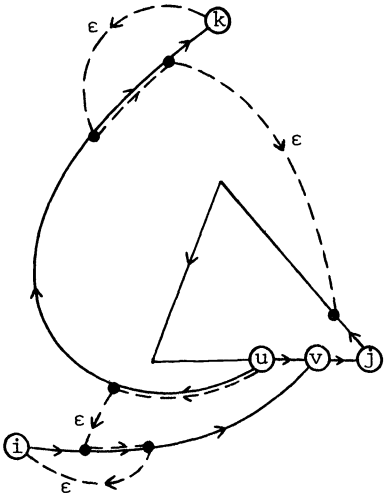


FIG. 4.18

5. Quasidistance *n*-matrices. In what follows, an n-matrix D will also be denoted $\langle\{i_1, \cdots, i_n\}\rangle$. If D is a principal n-submatrix of another matrix, say $\langle\{1, \cdots, m\}\rangle$, where $m \geqq n$, then the entries of D are those in the rows and columns $i_1, \cdots, i_n$ of $\langle\{1, \cdots, m\}\rangle$ and we say that $i_1, \cdots, i_n$ are the indices of D. All submatrices considered henceforth are principal.

The 3-submatrices of quasidistance n-matrices play a role similar to that of 4-submatrices of (symmetric) distance matrices. Results in this section are the analogues of results proved in [17] for the symmetric case.

LEMMA 5.1. *Submatrices of a shrinkable matrix are shrinkable.*

Proof. Let $\tilde{D} = \langle\{i_1, \cdots, i_n\}\rangle$ be a submatrix of $D = \langle\{1, \cdots, m\}\rangle$. If $\{D_i : 1 \leqq i \leqq z\}$ is a shrinking sequence of D, we form a shrinking sequence $\{\tilde{D}_i : 1 \leqq i \leqq z\}$ of $\tilde{D}$ as follows: if D_i is a compactification of D_{i-1} by r_q with respect to q, then let $\tilde{D}_i = \tilde{D}_{i-1}$

when q is not an index of $\tilde{D}_i$ and let $\tilde{D}_i$ be a compactification of $\tilde{D}_{i-1}$ by r_q with respect to q otherwise. If D_i is a reduction of D_{i-1}, where q was pendant from p, then we distinguish four subcases: 1) if neither q nor p are indices of $\tilde{D}_{i-1}$, set $\tilde{D}_i = \tilde{D}_{i-1}$; 2) if both q and p are indices of $\tilde{D}_{i-1}$, remove row and column q from $\tilde{D}_{i-1}$ to get $\tilde{D}_i$; 3) if q is an index of $\tilde{D}_{i-1}$ but p is not, note that $\tilde{D}_{i-1}$ is a submatrix of both D_{i-1} and D_i and set $\tilde{D}_i = \tilde{D}_{i-1}$; 4) if p is an index of $\tilde{D}_{i-1}$ but q is not, set $\tilde{D}_i = \tilde{D}_{i-1}$.

THEOREM 5.2. *If a quasidistance 4-matrix D contains a nonshrinkable 3-submatrix, then it contains at least two of them.*

Proof. With no loss of generality, let $\langle\{1, 2, 3\}\rangle$ be the only nonshrinkable 3-submatrix of D. By Lemma 4.1, $d_{12}+d_{24}+d_{41}=d_{14}+d_{42}+d_{21}$, $d_{31}+d_{43}+d_{14}=d_{13}+d_{34}+d_{41}$ and $d_{23}+d_{34}+d_{42}=d_{24}+d_{43}+d_{32}$. These equalities imply $d_{12}+d_{23}+d_{31}\neq d_{13}+d_{32}+d_{21}$, a contradiction which proves the theorem.

THEOREM 5.3. *If a quasidistance n-matrix $D=\langle\{1, \cdots, n\}\rangle$ contains a nonshrinkable 3-submatrix, then, for any index j of D, there is a nonshrinkable 3-submatrix which contains j.*

Proof. Suppose $\langle\{1, 2, 3\}\rangle$ is nonshrinkable. Denote by $a_1, \cdots, a_{n-3}$ the other indices of D. By Lemma 5.1, for each a_i, $\langle\{1, 2, 3, a_i\}\rangle$ is nonshrinkable and, by Theorem 5.2, it contains at least one nonshrinkable 3-submatrix other than $\langle\{1, 2, 3\}\rangle$. This clearly implies the truth of the theorem.

As a consequence, the existence of nonshrinkable 3-submatrices can be checked by an algorithm which is quadratic in n: only the $(n-1)(n-2)/2$ 3-submatrices containing a fixed index have to be checked, using for each of them, by Lemma 4.1, 4 sums and 1 comparison.

For $i=3, \cdots, n$, let $Q_i(D)$ denote the number of i-submatrices which contain a nonshrinkable 3-submatrix.

THEOREM 5.4. *If D contains a nonshrinkable 3-submatrix, then*

$$Q_i(D) \geqq \binom{n-2}{i-2} \quad \textit{for } i=3, \cdots, n.$$

Proof. Without loss of generality, let $\langle\{1, 2, 3\}\rangle$ be nonshrinkable. Denote by $a_1, \cdots, a_{n-3}$ the other indices of D. While proving Theorem 5.3, we showed that, besides $\langle\{1, 2, 3\}\rangle$, there is, for each a_j, a nonshrinkable 3-submatrix $\langle\{b_1, b_2, a_j\}\rangle$, where $\{b_1, b_2\}\subset\{1, 2, 3\}$. This proves the statement for $i=3$. For $i>3$, we may obtain a nonshrinkable i-submatrix either by joining an $(i-3)$-subset of $\{a_1, \cdots, a_{n-3}\}$ to $\{1, 2, 3\}$ or by joining an $(i-3)$-subset of $\{a_{j+1}, \cdots, a_{n-3}\}$ to $\{b_1, b_2, a_j\}$. The submatrices obtained are all distinct and there are at least

$$\binom{n-3}{i-3}+\binom{n-3}{i-2}=\binom{n-2}{i-2},$$

which completes the proof.

THEOREM 5.5. *Let D be an n-matrix, $n\geqq 4$. If D contains a nonshrinkable 3-submatrix, then it contains at least $n-2$ $(n-1)$-submatrices which are nonshrinkable.*

Proof. The statement is equivalent to saying that, if D has three $(n-1)$-submatrices each with no nonshrinkable 3-submatrix, then D has no nonshrinkable 3-submatrix. Let $S=\{1, \cdots, n\}$, $D=\langle S\rangle$ and let $\langle S-\{a\}\rangle$, $\langle S-\{b\}\rangle$, $\langle S-\{c\}\rangle$ each have no nonshrinkable 3-submatrix. It follows that $\langle\{a, b, c\}\rangle$ is the only 3-submatrix of D which may be nonshrinkable. This contradicts Theorem 5.4 and thus proves Theorem 5.5.

Since any $(i-1)$-submatrix of an n-matrix D is contained in $n-i+1$ i-submatrices of D, Theorem 5.5 immediately yields the following result:

THEOREM 5.6. *For $i\geqq 4$, $(n-i+1)\cdot Q_{i-1}(D)\geqq(i-2)\cdot Q_i(D)$.*

6. Patrinos–Hakimi hypertrees revisited. A result formally similar to the characterization of hypertree-realizable quasidistance matrices due to Patrinos and Hakimi [14] may be given for strictly nonnegative quasidistance matrices. For this purpose, let Γ be the sum of D and its transpose. The matrix Γ is a distance matrix and we call it the symmetrization of D. For any submatrix $\langle\{i, j, k, l\}\rangle$ of D, let $S_1 = d_{ij} + d_{ji} + d_{kl} + d_{lk} = \gamma_{ij} + \gamma_{kl}$, $S_2 = d_{ik} + d_{ki} + d_{jl} + d_{lj} = \gamma_{ik} + \gamma_{jl}$ and $S_3 = d_{il} + d_{li} + d_{jk} + d_{kj} = \gamma_{il} + \gamma_{jk}$.

THEOREM 6.1. *Let D yield D' by a compactification or a reduction and let Γ and Γ' be their respective symmetrizations. Γ' has at least one nontree-realizable 4-submatrix if and only if the same is true for Γ.*

Proof. Trivially, compactifications and reductions of a pendant index p not in $\{i, j, k, l\}$ do not alter the relations among the sums S_1, S_2, S_3 of the submatrix $\langle\{i, j, k, l\}\rangle$ of D. Let now i be pendant from q. If $q \notin \{i, j, k, l\}$, then, by Theorem 3.1, when reducing for i, we have the same relations among the sums S'_1, S'_2, S'_3 of the submatrix $\langle\{q, j, k, l\}\rangle$ of D' that we have among the sums S_1, S_2, S_3 of the submatrix $\langle\{i, j, k, l\}\rangle$ of D. If $q \in \{i, j, k, l\}$ this conclusion does not hold but this case never happens when the submatrix $\langle\{i, j, k, l\}\rangle$ of Γ is nontree-realizable. In fact, suppose, without loss of generality, that $q = k$ and that the submatrix $\langle\{i, j, k, l\}\rangle$ of Γ is nontree-realizable, which means that one of the sums S_1, S_2, S_3 is strictly greater than the other two. Using Theorem 3.1, we obtain $S_1 = S_3$, hence $S_2 > S_1$, which implies $d_{jl} + d_{lj} > d_{kj} + d_{jk} + d_{kl} + d_{lk}$, a contradiction to the fact that $d_{jl} \leqq d_{jk} + d_{kl}$ and $d_{lj} \leqq d_{lk} + d_{kj}$. This shows that when $q = k$ the submatrix $\langle\{i, j, k, l\}\rangle$ of Γ is tree-realizable, which completes the proof.

THEOREM 6.2. *Let D be a 4-matrix and let its symmetrization Γ be nontree-realizable. Then D is nonshrinkable.*

Proof. As we saw in the preceding proof, Γ of order 4 and nontree-realizable implies that no index of D may be pendant. Moreover, since compactification does not alter relations among the sums S_1, S_2, S_3 of D, we may (if necessary) compactify D and obtain a 4-matrix where, for each $i, r_i = 0$. By Remark 3.2 this matrix is nonshrinkable and so is D.

THEOREM 6.3. *Let D be an n-matrix and let its symmetrization Γ be nontree-realizable. Then D is nonshrinkable.*

Proof. By the hypothesis, Γ has a 4-submatrix Γ' which is nontree-realizable. By Theorem 6.2, the corresponding D' is nonshrinkable. By Lemma 5.1, D is also nonshrinkable.

THEOREM 6.4. *Let all 3-submatrices of D be shrinkable and the symmetrization Γ of D tree-realizable. Then D is shrinkable.*

Proof. Suppose D is nonshrinkable. By the definitions, Lemma 4.1 and Theorem 6.1, we may suppose that, for every index i in D, $r_i = 0$. By the definition of r_i and the hypothesis on the 3-submatrices, there is, for each i, at least one pair a, b such that $d_{ai} + d_{ib} = d_{ab}$ and $d_{bi} + d_{ia} = d_{ba}$. We can obviously choose a and b such that d_{ai} and d_{ib} are basic; by the hypothesis on the 3-submatrices, d_{ai} and d_{ib} basic imply d_{ia} and d_{bi} basic too. Our reasoning proceeds as follows.

Now let b play the role of i; there is at least one pair u, v such that $d_{ub} + d_{bv} = d_{uv}$ and $d_{vb} + d_{bu} = d_{vu}$ with d_{ub}, d_{bu}, d_{bv} and d_{vb} basic. We distinguish two cases:

Case 1. There is no pair u, v where i is one of u or v. This means that $d_{ub} + d_{bv} = d_{uv}$, $d_{vb} + d_{bu} = d_{vu}$, $d_{ib} + d_{bu} - d_{iu} = d_{ub} + d_{bi} - d_{ui} > 0$, $d_{ib} + d_{bv} - d_{iv} = d_{vb} + d_{bi} - d_{vi} > 0$. An easy calculation shows that, for the 4-submatrix $\langle\{i, b, u, v\}\rangle$ of Γ, the sum S_1 is strictly greater than S_2 and S_3 which means that this submatrix and, consequently, Γ, are not tree-realizable, a contradiction.

Case 2. There is v such that $d_{ib}+d_{bv}=d_{iv}$ and $d_{vb}+d_{bi}=d_{vi}$.

In this case we repeat the reasoning and, unless Case 1 occurs, we form a finite sequence of indices denoted, without loss of generality, by $1, 2, \cdots, n, 1$ and such that, for all i and with the usual convention that $n+1$ is 1,

$$d_{i-1,i}+d_{i,i+1}=d_{i-1,i+1} \text{ and } d_{i+1,i}+d_{i,i-1}=d_{i+1,i-1}.$$

Starting with 1, let k be the last index in the sequence such that $d_{1k}=d_{12}+d_{23}+\cdots+d_{k-1,k}$ and $d_{k1}=d_{k,k-1}+\cdots+d_{32}+d_{21}$. Such a k exists (because we cannot have $d_{1k}=d_{1,k-1}+d_{k-1,k}$ and $d_{k1}<d_{k,k-1}+d_{k-1,1}$) and it is $3\leqq k\leqq n-1$ (otherwise we would not have $d_{n-1,1}=d_{n-1,n}+d_{n1}$ and $d_{1,n-1}=d_{1n}+d_{n,n-1}$). Let $s=k-1$ and $j=k+1$; j is therefore the first index such that $d_{1j}<d_{1k}+d_{kj}$ and $d_{j1}<d_{jk}+d_{k1}$. Using these inequalities, an easy calculation shows that, for the submatrix $\langle\{1, s, k, j\}\rangle$ of Γ, the sum S_2 is strictly greater than S_1 and S_3, which means that this submatrix and, consequently, Γ are not tree-realizable, a contradiction. This completes the proof of the theorem.

Rephrasing Theorems 6.1, 6.3 and 6.4, we obtain:

THEOREM 6.5. *A quasidistance n-matrix is shrinkable if and only if its* (*principal*) 3-*submatrices are shrinkable and its symmetrization is tree-realizable.*

Acknowledgments. Thanks are due to Professor Eduardo M. de Sá, from the University of Aveiro and the Center for Mathematics of the University of Coimbra and to Professor C. M. Zamfirescu, from Hunter College for many stimulating discussions.

Note added in proof. We came across the paper by James A. Cunningham, *Free trees and bidirectional trees as representations of psychological distance*, J. Math. Psych., 17 (1978), pp. 165–188, where very interesting applications of distance and quasidistance matrices in psychological research are given. Moreover, a nonnegative version of the Patrinos–Hakimi theorem is also stated and used. However, for the sufficiency part of the proof (here Theorem 6.4), the author says only that it can be achieved by induction; since the details are not trivial, we decided to keep them here.

REFERENCES

[1] F. T. BOESCH, *Properties of the distance matrix of a tree*, Quart. Appl. Math., 26 (1968–69), pp. 607–609.

[2] P. BUNEMAN, *A note on the metric properties of trees*, J. Combin. Theory Ser. B, 17 (1974), pp. 48–50.

[3] S. CHAIKEN, A. K. DEWDNEY AND P. J. SLATER, *An optimal diagonal tree code*, this Journal, 4 (1983), pp. 42–49.

[4] A. K. DEWDNEY, *Diagonal tree codes*, Inform. and Control, 40 (1979), pp. 234–239.

[5] A. DRESS, *A characterization of tree like metric spaces or how to construct an evolutionary tree*, unpublished manuscript.

[6] M. EIGEN, W. GARDINER, P. SHUSTER AND R. WINKLER-OSWATITSCH, *The origin of genetic information*, Scientific American (April 1981), pp. 88–118.

[7] S. L. HAKIMI AND S. S. YAU, *Distance matrix of a graph and its realizability*, Quart. Appl. Math., 22 (1964–65), pp. 305–317.

[8] W. IMRICH, *Realisierung von Metriken in Graphen*, Sitzungsberichten des Osterreichischen Akademie der Wissenschaften, Abteilung II, 178 (1969), pp. 19–24.

[9] ———, *On metric properties of tree like spaces*, in Contributions to Graph Theory and Its Applications (Intern. Colloq. Oberhof 1977), Technische Hochschule Ilmenau, Ilmenau, 1977, pp. 129–156.

[10] W. IMRICH AND G. SCHWARZ, *Trees and length functions in groups*, Ann. Discr. Math., to appear.

[11] W. IMRICH, J. M. S. SIMÕES-PEREIRA AND C. M. ZAMFIRESCU, *On optimal embeddings of metrics in graphs*, J. Combin. Theory Ser. B., to appear.

[12] W. IMRICH AND E. STOTZKII, *The optimal embeddings of metrics into graphs*, Siberian Math. J., 13 (1972), pp. 558–565. (In Russian.)

[13] J. LAWRENCE, C. R. JOHNSON AND E. HOWE, *The structure of distances in networks*, to appear.

[14] A. N. PATRINOS AND S. L. HAKIMI, *The distance matrix of a graph and its tree realization*, Quart. Appl. Math., 30 (1972–73), pp. 255–269.

[15] P. H. SELLERS, *The theory and computation of evolutionary distances: pattern recognition*, J. Algorithms, 1 (1980), pp. 359–373.

[16] J. M. S. SIMÕES-PEREIRA, *A note on the tree realizability of a distance matrix*, J. Combin. Theory, 6 (1969), pp. 303–310.

[17] J. M. S. SIMÕES-PEREIRA AND C. M. ZAMFIRESCU, *Submatrices of non-tree-realizable distance matrices*, Linear Algebra and Its Applications, 44 (1982), pp. 1–17.

[18] E. A. SMOLENSKII, *A method for the linear recording of graphs*, Zh. Vichisl. Mat. i Mat. Fiz., 2 (1962); pp. 371–372; USSR Comput. Math. and Math. Phys., 2 (1963), pp. 396–397.

[19] B. C. TANSEL, R. L. FRANCIS, T. J. LOWE AND M. L. CHEN, *Duality and distance constraints for the non-linear p-center problem and covering problem on a tree network*, Operations Research, 30 (1982), pp. 725–744.

[20] K. A. ZARETZKII, *Constructing a tree on the basis of a set of distances between the hanging vertices*, Uspekhi Mat. Nauk, 20-6 (1965), pp. 90–92. (In Russian.)

SIAM J. ALG. DISC. METH.
Vol. 5, No. 1, March 1984

0196-5212/84/0501-0015 $01.25/0

CONVERGENT REGULAR SPLITTINGS FOR SINGULAR *M*-MATRICES*

DONALD J. ROSE†

Abstract. A classic theorem of numerical linear algebra (Varga's Theorem) says that any regular splitting of an M-matrix, A, is convergent. However when A is a singular M-matrix, such as might arise in queuing networks, this general theorem needs some modification. We define a class of block splittings, called R-regular splittings, and show that any R-regular splitting is convergent. Furthermore, given a block splitting like block Gauss-Seidel but not quite as special, we show how to tinker slightly to make the splitting convergent. We show also that some natural splittings require no tinkering at all. In an algorithmic context, our results indicate how to choose and order the blocks in a block iterative method for solving a singular M-matrix system of linear equations so as to insure convergence of the method.

1. Introduction. Consider an $n \times n$ matrix $Q = (q_{ij})$ with $q_{ij} \leqslant 0$ for $i \neq j$ and $\sum_{i=1}^{n} q_{ij} = 0$ for each $1 \leqslant j \leqslant n$. We call such matrices *Q-matrices* since they arise in the analysis of queuing network. Q-matrices are a subclass of the class of singular M-matrices (defined below). It is clear that a Q-matrix is singular since all columns sum to zero.

An application of the main results presented here says that block Gauss-Seidel iteration can always be made to work, perhaps after some tinkering, for solving an irreducible singular M-matrix system of linear equations. Gauss-Seidel iteration, and its overrelaxed generalization SOR, have been used by Kaufman [KGW], [KSM] to solve significant problems involving Q-matrices. In these problems n is large, the Q-matrices are sparse, and iteration is often the only feasible approach.

Let $A = (a_{ij})$ be an $n \times n$ matrix with $a_{ij} \leqslant 0$ for $i \neq j$. If A is nonsingular and each entry of A^{-1} is nonnegative, written as $A^{-1} \geqslant 0$, then A is said to be a *nonsingular M-matrix*. These matrices arise in numerous applications (see [BP] and [V]) ranging from the numerical solution of partial differential equations to mathematical economics, and they have motivated research in pure and applied linear algebra for nearly a century. When confronted with a nonsingular M-matrix life is, at least in theory, quite pleasant.

Any M-matrix A has nonnegative diagonal entries (all $a_{ii} \geqslant 0$); a nonsingular M-matrix has a positive diagonal. In fact, A is a nonsingular M-matrix if and only if all principal minors of A are positive. For any $n \times n$ matrix C, let $\rho(C) = \max|\lambda|$, where the maximum is taken over the set of eigenvalues of C. Another characterization of nonsingular M-matrices, and one that allows a convenient generalization to the singular case, says that A is a nonsingular M-matrix iff

$$A = sI - B \tag{1.1a}$$

where

$$s > \rho(B) \tag{1.1b}$$

and

$$B \geqslant 0\,. \tag{1.1c}$$

*Received by the editors December 7, 1982 and in final revised form February 22, 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 June 6, 1983.

†Bell Laboratories, Murray Hill, New Jersey 07974.

For a *nonnegative matrix* B as in (1.1c), it is classic that $\rho(B)$ is an eigenvalue of B. See [BP, Chapts. 2, 6]. Many authors use (1.1) as the definition of a nonsingular M-matrix and define a *singular* M-matrix (sM-matrix) as in (1.1) with (11.b) replaced by

$$(1.1\text{d}) \qquad s = \rho(B) .$$

The iterative methods we will consider are induced by regular splittings of A. These iterations take the form

$$(1.2) \qquad Mx_{n+1} = Nx_n + b$$

to attempt to solve $Ax = b$. Here A is *split* as

$$(1.3) \qquad A = M - N, \qquad M^{-1} \geqslant 0, \qquad N \geqslant 0 .$$

Equation (1.3) defines a *regular splitting.* Any regular splitting of a nonsingular M-matrix is convergent; that is, $\lim x_i = x$ where $Ax = b$ (Varga's theorem).

In contrast, a regular splitting of an sM-matrix is not always convergent. Indeed the well known Gauss-Seidel method, where M is the lower triangle of A, may fail. We show how to remedy this situation. Our main result is that any R-regular splitting of an irreducible sM-matrix is convergent. An R-regular splitting, defined in § 3, is a block splitting, like block Gauss-Seidel, but more general. Given a choice of diagonal blocks, there always exist block orderings, that is, block permutation matrices P, such that PAP^T has an R-regular splitting with the same diagonal blocks.

In § 2 we define terms and present some known results about sM-matrices we will use later. Section 3 discusses block splittings in general and R-regular splittings in particular; our main results are also given here. Section 4 contains most of the details; the discussion there is mainly graph-theoretic.

We conclude this section with a folk-theorem; that is, a result surely known but not easily found (see [G] for pieces of the result). It seems worth the telling, however, since it relates sM-matrices, Q-matrices, and *column stochastic matrices,* $T = (t_{ij})$, such that

$$\sum_{i=1}^{n} t_{ij} = 1, \qquad 1 \leqslant j \leqslant n .$$

Let us call a diagonal matrix $D = (d_{ij})$ (with $d_{ij} = 0$ for $i \neq j$) *diagonal positive* if each $d_{ii} > 0$.

THEOREM 0. *The following are equivalent:*

(i) *A is an irreducible sM-matrix;*

(iia) *$A = D_1Q$, D_1 diagonal positive and Q an irreducible Q-matrix, and*

(iib) *for any such Q, $QD_2 = I-T$, D_2 diagonal positive and T irreducible column stochastic;*

(iii) *$A = D_1(I-T)D_2$, D_1,D_2,T as in* (ii).

Proof. An $n \times n$ matrix is irreducible iff its graph, defined later, is strongly connected.

(i) implies (ii). Let $A = sI-B$, $s = \rho(A)$, $B \geqslant 0$ and irreducible. Hence B^T has a positive eigenvector, $x > 0$, such that $B^Tx = sx$ ([V, p. 28], [BP, p. 27]). Let D_1 be diagonal positive with $d_{ii} = x_i$. Then $x = De$, $e = (1,1,...,1)^T$, and $B^TDe = sDe$ so $e^T(D^TA) = 0$; i.e., $D^TA = Q$, an irreducible Q-matrix.

Any irreducible Q matrix has positive diagonal entries, say q_{ii}. Let D_2 be diagonal positive with $d_{ii} = q_{ii}^{-1}$. Then $QD_2 = I-T$, $T \geqslant 0$ irreducible and column stochastic since $O = e^T QD_2 = e^T(I-T)$.

(ii) implies (iii). Immediate.

(iii) implies (i). An irreducible column stochastic matrix, T has $\rho(T) = 1$ and hence a positive eigenvector $Tx = x$. Thus $[D_1(I-T)D_2](D_2^{-1}x) = 0$; i.e., there exists a $y > 0$ such that $Ay = 0$. So A is singular. Also $A = D_1(I-T)D_2 = sI-B$, $B \geqslant 0$ and irreducible for $s = \max d_{ii}$ where $D = (d_{ij}) = D_1D_2$. But $By = sy$ so $s = \rho(B)$; see [BP, p. 28]. □

2. Preliminaries. Given the definition of an M-matrix as in (1.1), it is not surprising that our results will depend upon the theory of nonnegative matrices. We will borrow as needed from this theory as well as from the theory of sM-matrices given in [BP, §§ 6.4, 7.6]. Some graph theory will also be necessary.

Let $A = (a_{ij})$ be an $n \times n$ matrix. The *graph* of A, $G(A) = (V,E)$, will be a *directed* graph with n vertices $V = \{v_i\}$ and directed edges $(v_i,v_j) \in E$ iff $a_{ij} \neq 0$. For an edge $e = (v_i,v_j)$, v_j is *adjacent from* v_i while v_i is *adjacent to* v_j. The subscript i on v_i implicitly assigns an ordering to the vertices in V. Each such ordering is a bijection $V \leftrightarrow \{1,2,...,n\}$ and represents a matrix in the class PAP^T, P a permutation matrix. The unordered graph represents the whole class PAP^T. Sometimes we refer to a vertex $v \in V$ as $i \in \{1,2,...,n\}$.

An *x,y path* in G of *length* $l > 0$ is an ordered set $p = [v_1,v_2,...,v_{l+1}]$ such that v_i is adjacent to v_{i+1}, $1 \leqslant i \leqslant l$. If $v_{l+1} = v_1$, p is called a *cycle.* A *loop* is a cycle of length $l = 1$, i.e., $p = [v_1,v_1]$. A *trivial graph* consists of a single vertex and no edge (no loop). If $S \subseteq V$, the *subgraph induced* by S is

$$G[S] = (S,E[S]) \quad \text{where} \quad E[S] = \{(x,y) \in E \,|\, x,y \in S\} .$$

Usually our matrices will be irreducible. A matrix is *irreducible* if $G(A)$ is strongly connected; that is, iff for any $x,y \in V$ there exists an x,y path. We will also discuss $p \times p$ block matrices $A = (A_{ij})$, $1 \leqslant i,j \leqslant p$, with square diagonal blocks. The *block graph* $\mathbf{G}(A) = (\mathbf{V},\mathbf{E})$ will be a directed graph with p vertices $\mathbf{V} = \{V_i\}$ and directed edges (V_i,V_j) iff $A_{ij} \neq 0$. Each V_i can be considered as the vertex set in the induced subgraph $G[V_i] = G(A_{ii})$. Each edge $(V_i,V_j) \in \mathbf{E}$ can be considered to be the set of directed edges from V_i to V_j; i.e., $\{(v_k,v_l) \in E \,|\, v_k \in V_i, v_l \in V_j\}$.

Recall that any reducible matrix A can be permuted into *reduced triangular block form*; i.e., there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_{11} & & & \\ A_{21} & A_{22} & & \\ \vdots & & \ddots & \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix}$$

where each A_{ii} is square and irreducible or a 1×1 null matrix (see [BP, pp. 39,261] or [V, p. 46]). Graph theoretically the $G(A_{ii})$ are the *strong components* of $G(A)$; they are trivial or strongly connected.

Let A be an $n \times n$ irreducible sM-matrix with *regular splitting* $A = M-N$; that is, M nonsingular with $M^{-1} \geqslant 0$ and $N \geqslant 0$. The matrix $H = M^{-1}N$ is called the *iteration matrix* of the splitting. The iterative method induced by the splitting $A = M-N$ is *convergent* iff $\lim_{k\to\infty} H^k$ exists (see [BP, pp. 197-198]; we have used "convergent" rather than "semiconvergent"). In this case we call the splitting a C-*regular splitting.* To determine whether a splitting is C-regular we must examine the structure of the spectrum of H.

PROPOSITION 1. *Let A be an irreducible sM-matrix and $A = M-N$ be a regular splitting. Then there exists a nonsingular matrix S (reducing H to Jordan form, say) such that*

$$H = M^{-1}N = S^{-1}\begin{bmatrix} e^{i\theta} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & K \end{bmatrix} S$$

where $\rho(K) < 1$ and the possibly empty upper triangular matrix $e^{i\theta}$ has diagonal entries $\omega \neq 1$ with $|\omega| = 1$. Furthermore, H is C-regular if and only if $e^{i\theta}$ does not exist; i.e.,

$$H = S^{-1}\begin{bmatrix} I & 0 \\ 0 & K \end{bmatrix} S$$

with $\rho(K) < 1$.

Proof. This result is known but is not exactly stated as such in [BP]. The first part uses irreducibility and follows from [BP, Thms. 6.4.12 (F_{13}), 6.4.16 (3); see also pp. 152-153, 197]. The second statement follows directly from [BP, Lemma 7.6.9]. □

So we must get rid of $e^{i\theta}$; this is done combinatorially. We will examine the graph $G(H)$ of the nonnegative matrix H. Usually $G(H)$ will not be strongly connected. It is never strongly connected for any point or block Gauss-Seidel method.

Let G be nontrivial and strongly connected and define the *cycle index* $\mu(G)$ as the greatest common divisor (gcd) of the length of all cycles of G. If $\mu(G) = 1$ we say G is *primitive;* for general G, we say G is *strongly primitive* if each strong component is primitive or trivial. We apply the theory of nonnegative matrices (V, Thm. 2.3] or [BP, pp. 32-35]) to obtain

PROPOSITION 2. *For A and H as in Proposition* 1, *H is C-regular if $G(H)$ is strongly primitive. If there exist a nontrivial strong component of G and no such component is primitive, H is not C-regular.*

Proof. As discussed above, H can be permuted (PHP^T) into block triangular form (either upper or lower) where the diagonal blocks correspond to the strong components of $G(H)$. Hence the eigenvalues of H are the eigenvalues of the nonnegative diagonal blocks. They have graphs G_i with cycle indices $\mu(G_i)$ if they are nontrivial. Any trivial G_i corresponds to a eigenvalue $\lambda_i = 0$. If $G(H)$ is strongly primitive, each irreducible diagonal block H_k with nontrivial graph G_k has a simple positive eigenvalue $\rho(H_k)$. Furthermore $\rho(H_k) \leqslant 1$ is the unique eigenvalue of this maximum modulus. So $e^{i\theta}$ does not exist. On the other hand if no $G(H_K)$ is primitive (and one is nontrivial) the H_k with $\rho(H_k) = 1$ has other eigenvalues on the unit circle. See the references for more detail. □

3. ***R*-regular splittings**. To motivate the notion of an R-regular splitting, consider the block 2×2 matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{3.1}$$

The corresponding block Gauss-Seidel splitting is

$$A = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix} - \begin{bmatrix} 0 & N_{12} \\ 0 & 0 \end{bmatrix} = M-N \ . \tag{3.2}$$

If A is irreducible, $M_{21} \neq 0$ and $N_{12} \neq 0$. Since the iterative method will involve solving linear systems with M_{11} and M_{22}, it is reasonable to demand that M_{11} be irreducible. If not there exists a permutation matrix P such that $P_1 M_{11} P_1^T$ is block triangular. Hence we can permute A into a form like (3.1) and (3.2) with M_{11} irreducible but smaller than the original M_{11} of (3.2). Notice that $M_{22} = A_{22}$ and that we can continue the process just described (on A_{22}) until A has p, say, irreducible diagonal blocks M_{ii}, $1 \leqslant i \leqslant p$, and corresponding conformable off-diagonal blocks, M_{ij}. Computationally, it does not matter where the off-diagonal blocks go; they represent matrix multiplications. The diagonal blocks represent linear equations solving, and they stay on the (block) diagonal under appropriate (block) permutations PAP^T.

We formalize the preceding discussion as follows. Let $A = (A_{ij})$ be an irreducible block $p \times p$ matrix $(1 \leqslant i,j \leqslant p)$ of order n. A *block splitting* of A is a splitting $A = M-N$ where $M = (M_{ij})$ and $N = (N_{ij})$ are block $p \times p$ matrices conformable with the block structure of A and $p > 1$. For example, when $M = D-L$ and $N = U$ where D, $-L$, and $-U$ are the block diagonal, block lower triangle, and block upper triangle of A, the splitting, $A = M-N$, is the *block Gauss-Seidel splitting*. A special case is the *point Gauss-Seidel splitting* where $p = n$ and D is diagonal.

Any block Gauss-Seidel splitting (with irreducible diagonal blocks) arises from a partitioning of V in $G(A) = (V,E)$ into disjoint subsets M_i, $1 \leqslant i \leqslant p$ (user chosen, say) such that the *induced subgraphs,*

$$G_i = (M_i,E_i), \ \ E_i = \{e = (v,w) \in E \,|\, v, w \in M_i\} \ , \tag{3.3}$$

are strongly connected. If the M_i chosen are not strongly connected, find the strong components of each [AHU, § 5.5] and start over with a larger p. These M_i are the vertices of a graph $\mathbf{G} = (\mathbf{M},\mathbf{E})$ with p vertices and edges (M_i,M_j) iff $(v,w) \in E$ for some $v \in M_i$, $w \in M_j$. Now any (re)ordering of the vertex set $\mathbf{M}$ induces an ordering on the vertices in the M_i by ordering them consecutively. This in turn, induces an ordering on all of V. The final ordered graph $G = (V,E)$ is $G(PAP^T)$ for some permutation matrix P. Furthermore the $G_i = (M_i,E_i)$ are $G(D_{ii})$, where D_{ii} are the diagonal blocks of the induced Gauss-Seidel splitting. For nonsingular M-matrices any Gauss-Seidel splitting is convergent; for sM-matrices we must order the vertices in $\mathbf{M}$ appropriately.

The splittings we will consider are somewhat more general than block Gauss-Seidel splittings. We will require $M = D-L$, D block diagonal and $L \geqslant 0$ strictly (block) lower triangular. $N = (N_{ij})$ will be nonnegative such that

$$A_0 = D-L-U(N) \tag{3.4}$$

is irreducible where

$$N \equiv L(N) + D(N) + U(N) \tag{3.5}$$

and $L(N)$, $D(N)$, and $U(N)$ are strictly block lower triangular, block diagonal, and strictly block upper triangular, respectively. In short some of the lower triangular part of $A = M-N$ may be put into $L(N)$ or $D(N)$. While there may be little gain in taking $L(N) \neq 0$, $D(N) \neq O$ could result from clever sparse LU approaches to factoring the diagonal blocks. Note that, in relation to (3.4),

$$A = (D-D(N)) - (L+L(N)) - U(N) . \tag{3.6}$$

If the A_{ii} of $A = (A_{ij})$ are irreducible, then A_0 irreducible implies A is irreducible since the edges of $\mathbf{G}(A)$ contain the edges of $\mathbf{G}(A_0)$.

Let $A = M-N$ be a block splitting of an $n \times n$ matrix $A = (A_{ij})$, $1 \leqslant i,j \leqslant p$. We define an *R-regular splitting* to be a block splitting such that:

(ia) M is block lower triangular with $M = D-L$, D block diagonal and $L \geqslant 0$ strictly block lower triangular;

(ib) $N \equiv L(N) + D(N) + U(N) \geqslant 0$, $L(N)$, $D(N)$, and $U(N)$ as in (3.5);

(ii) $D_{ii}^{-1} > 0$, $1 \leqslant i \leqslant p$;

(iii) $A_0 = D-L-U(N)$ is irreducible;

(iv) $\mathbf{G}(A_0)$ has a monotone cycle (a cycle $c = [i_1,i_2,\ldots,i_l,i_1]$ with $i_l \neq i_1$, and $i_j \geqslant i_{j+1}$, $1 \leqslant j \leqslant l-1$).

While conditions (i)-(iv) may appear to be special we claim that they are a natural generalization of a block Gauss-Seidel regular splitting. Conditions (ia) and (ii) insure that $M^{-1} \geqslant 0$; together with $N \geqslant 0$ of (ib) this implies $M-N$ is a regular splitting of A. If the D_{ii} are nonsingular M-matrices, as in our setting, $D_{ii}^{-1} > 0$ iff D_{ii} is irreducible. Usually the D_{ii} will be irreducible since the A_{ii} will be irreducible; as noted above, (iii) then implies that A itself is irreducible. Note that L of $M = D-L$ need not be the strictly lower triangular part of A; $L \neq 0$, however, since A_0 is irreducible. The monotone cycle condition (iv) may fail for an arbitrary block splitting. However, since $\mathbf{G}(A_0) = (\mathbf{V},\mathbf{E})$ must have a cycle, cycle vertices of $\mathbf{V}$ can be reordered so this cycle is monotone (the remaining vertices of $\mathbf{V}$ can be arbitrarily ordered). Ordering vertices $v_j \in V_i \in \mathbf{V}$ consecutively induces a reordering on the whole vertex set V of $G(A_0)$ and $G(A)$; this corresponds to a (block) permutation PAP^T and PA_0P^T. Clearly $\mathbf{G}(A_0)$ remains strongly connected; however, the blocks in the original L and $U(N)$ may migrate across the diagonal and, hence, the reordering also redefines L and $U(N)$. We summarize in

PROPOSITION 3. *Let A be an irreducible sM-matrix and $A = M-N$ be a block splitting such that D is the block diagonal of M. Suppose each diagonal block D_i is an irreducible nonsingular M-matrix. Then there exist (block) permutation matrices P such that PAP^T has an R-regular splitting with the same diagonal blocks.*

Proof. The conclusions follow from the discussion above. We make a remark. If D is the block diagonal of A itself and each D_i is irreducible, then D_i is an irreducible nonsingular M-matrix as shown in [BP, Thm. 4.16, p. 156]. □

Our main result is

THEOREM 1. *Any R-regular splitting of an irreducible sM-matrix, A, is convergent. Furthermore, given any block splitting of A, there exist (block) permutation matrices P such that PAP^T has an R-regular splitting with the same diagonal blocks.*

Proof. In the next section we will show that the graph of the iteration matrix, H, of an R-regular splitting is strongly primitive. The first statement then follows from Proposition 2 while the second statement follows from Proposition 3. □

In analogy with the nonsingular M-matrix case 9 (see [V, p. 92]), we can show that SOR splittings for $0 < \omega < 1$ are R-regular, hence convergent, if the block Gauss-Seidel case, $\omega = 1$, is R-regular. However, the monotonicity of $\rho(H_\omega)$ as in [V, p. 92, eqn. 3.79] is unsettled.

COROLLARY 1. *Let $A = D-L-U$ be an irreducible sM-matrix with R-regular splitting $M = D-L$, $N = U$. Then the SOR splitting*

$$M_\omega = \omega^{-1}(D-\omega L) \ , \tag{3.6a}$$

$$N_\omega = U + \omega^{-1}(1-\omega)D \ , \tag{3.6b}$$

$0 < \omega < 1$, *is also an R-regular splitting, hence convergent.*

Proof. Note $M_\omega = D_\omega - L$, $D_\omega = \omega^{-1}D$, and $U(N_\omega) = U$, $D(N_\omega) = \omega^{-1}(1-\omega)D$. Clearly, $A_{0,\omega} = \omega^{-1}D-L-U$ is irreducible and $\mathrm{G}(A_{0,\omega}) = \mathrm{G}(A)$ so verification of R-regularity is straightforward. □

COROLLARY 2. *Let A_0 be as in the definition of an R-regular splitting, and write $A_0 = (A_{ij}^0)$. If $A_{ij}^0 \neq 0$ and $A_{ji}^0 \neq 0$ for some $i \neq j$, then $\mathrm{G}(A_0)$ has a monotone cycle. In particular if $A = M-N$ is any block (point) Gauss-Seidel splitting of an irreducible sM-matrix with irreducible diagonal blocks such that $A_{ij} \neq 0$ and $A_{ji} \neq 0$ ($a_{ij} \neq 0$ and $a_{ji} \neq 0$), then this splitting is R-regular, hence convergent.*

Proof. Any cycle of length $l = 2$ with two distinct vertices is monotone. The remark in the proof of Proposition 3 insures that the diagonal blocks D_{ii} have $D_{ii}^{-1} > 0$. □

Recall that a block splitting must have at least $p = 2$ blocks; for technical reasons our exposition excludes the case of a single block, M itself, with $M^{-1} > 0$ (each entry positive). Such an M might arise naturally by choosing M to be an irreducible "submatrix" of A as we see in

THEOREM 2. *Any regular splitting of an irreducible sM-matrix A with $M^{-1} > 0$ is convergent. For example, if M is obtained from A by setting some off diagonal entries of A to zero and such an M is irreducible, then the corresponding splitting is regular and convergent.*

Proof. Let A have regular splitting $A = M-N$, $M^{-1} > 0$, $N \geqslant 0$. Then the kth column of $H = M^{-1}N$ has all entries positive or all entries zero, and there exists a permutation matrix P such that

$$H_1 = PHP^T = \begin{bmatrix} Z_1 & H_1 \\ Z_2 & H_2 \end{bmatrix}$$

where $Z_i = 0$, $H_i > 0$, $i = 1{,}2$ and Z_1 and H_2 are $n_1 \times n_1$ and $n_2 \times n_2$, respectively. Thus the strong components of $G(H_1)$ consist of n_1 trivial vertices and the complete graph on n_2 vertices, so $G(H_1)$ is clearly strongly primitive.

For the second statement we have from Theorem 0 that $A = DQ = M-N$, D diagonal positive and Q a Q-matrix. Thus $Q = M_1 - N_1$, $M_1 = D^{-1}M$, $N_1 = D^{-1}N$, M_1 is irreducible and M_1 is obtained from Q by setting some off-diagonal entry of Q, say q_{ij}, to zero. Note $Q^Te = 0$ implies $M_1^Te = N_1^Te \geqslant 0$ with strict equality in the jth entry. Hence M_1^T is irreducibly diagonally dominant implying $M_1^{-1} > 0$ and $M^{-1} > 0$; see [V, pp. 23, 85]. □

4. Iteration graphs. Let A be an irreducible sM-matrix and write

$$A = D-L-U \tag{4.1}$$

where $D = (d_{ij})$, $L = (l_{ij})$, and $U = (u_{ij})$ are diagonal, strictly lower triangular, and strictly upper triangular nonnegative matrices, respectively. The iteration matrix

$$(4.2) \qquad H = (D-L)^{-1}U = (I-L_1)^{-1}U_1, \qquad L_1 = D^{-1}L, \qquad U_1 = D^{-1}U$$

is called the *point Gauss-Seidel iteration matrix.* The second equality of (4.2) says there is no loss in considering matrices

$$(4.3) \qquad H = (I-L)^{-1}U\,.$$

In this section we study the graph of H, which we call an *iteration graph,* and relate it to the graph of $A = I-L-U$. We will also see that we can reduce more general splittings to this context.

Let $G = (V,E)$ be a graph with ordered vertices $V = \{v_i\}$. A path $p = [i_1,i_2,\ldots,i_l]$ will be called *monotone* if $i_j \geqslant i_{j+1}$, $1 \leqslant j \leqslant l-1$; that is, vertices are nonincreasing along the i_1,i_l path. The path p is 1*-monotone* if $p_1 = [i_1,\ldots,i_{l-1}]$ is monotone and $i_{l-1} < i_l$. Thus a 1-monotone path has length $l \geqslant 2$, and a *monotone cycle* is a 1-monotone path with $i_l = i_1$.

PROPOSITION 4. *Let $L = (l_{ij})$ be a strictly lower triangular nonnegative matrix. Then (i,j) is an edge of $G((I-L)^{-1})$ if and only if there is a monotone (i,j) path in $G(I-L)$.*

Proof. We use induction of the size, k, of the vertex V of G; the case $k = 1$ is trivial since (1,1) is a loop of both graphs. Write

$$(4.4) \qquad (I-L) = \begin{bmatrix} 1 & 0 \\ -l_1 & (I-L_1) \end{bmatrix}$$

where both the $(n-1) \times 1$ matrix l_1 and the $(n-1) \times (n-1)$ strictly lower triangular matrix L_1 are nonnegative. Note that $G(I-L_1)$ is the subgraph of G induced by the vertex set $V_1 = \{2,3,\ldots,n\}$. Now

$$(4.5) \qquad (I-L)^{-1} = \begin{bmatrix} 1 & 0 \\ (I-L_1)^{-1}l_1 & (I-L_1)^{-1} \end{bmatrix}.$$

If $j \neq 1$ the result follows from the induction hypothesis applied to the $k = n-1$ vertex graphs $G(I-L_1)$ and $G((I-L_1)^{-1})$.

Suppose $j = 1$ and let $(I-L_1)^{-1} = (t_{kl})$ and $l_1 = (s_2,\ldots,s_n)^T$; note both $(I-L_1)^{-1}$ and l_1 are nonnegative. Then $(i,1) \in E$ of $G((I-L)^{-1})$ if and only if $\sum t_{ip}s_p \neq 0$; that is, iff there is a monotone i,p path (corresponding to t_{ip}) followed by an edge $(p,1)$, $p > 1$. Thus $(i,1) \in E$ iff there is a monotone $(i,1)$ path in $G(I-L)$. □

PROPOSITION 5. *Let L be as in Proposition* 4 *and $U = (u_{ij})$ be strictly upper triangular and nonnegative. Then (i,j) is an edge of $H = (I-L)^{-1}U$ iff there is a* 1*-monotone (i,j) path in $G(A = I-L-U)$.*

Proof. Let $(I-L)^{-1} = (l_{ij}^*)$, $H = (h_{ij})$, and consider the entries $h_{ij} = \Sigma l_{ik}^* u_{kj}$, $k \leqslant \min(i,j)$. Thus $h_{ij} \neq 0$ iff some $l_{ik}^* \neq 0$ and $u_{kj} \neq 0$; that is, iff there exists a monotone (i,k) path in $G(I-L)$ and an edge (k,j) in $G(U)$. Since $G(I-L)$ and $G(U)$ are edge disjoint, this implies $h_{ij} \neq 0$ iff there exists a 1-monotone (i,j) path in $G(A)$. □

Note that Proposition 5 implies that any edge (i,j) with $i < j$ of $G(A)$ (also $G(U)$) is an edge in $G(H)$ since $p = (i,i,j)$ is 1-monotone in $G(A)$. In addition

Proposition 5 immediately implies the following

COROLLARY 3. *If $G(A)$ has a monotone (i,i) cycle, then $G(H)$ has a loop (i,i).*

PROPOSITION 6. *Let A,L,U, and H be as in Propositions 4 and 5. Then there exists an i,j path $p = [k_1,k_2,...,k_{l+1}]$ of length $l \geqslant 1$ and $k_l < k_{l+1} = j$, called a special (i,j) path, in $G(A)$ iff there exists an (i,j) path in $G(H)$.*

Proof. Both parts are by induction on path length l. Suppose the special i,j path, p, exists in $G(A)$. If p has length $l = 1$ then (i,j) is an edge of both $G(A)$ and $G(H)$. If $l \geqslant 2$ let q be the first index such that $k_{q-1} < k_q$. If $q = l+1$, then p is 1-monotone and (i,j) is an edge of $G(H)$. Otherwise $p_1 = [k_1,k_2,...,k_q]$ and $p_2 = [k_q,...,k_{l+1}]$ are both shorter special paths in $G(A)$, hence (k_1,k_q) and (k_q,k_{l+1}) paths exist in $G(H)$ by induction. Thus $G(H)$ has an (i,j) path.

Conversely suppose $p = [k_1,k_2,...,k_{l+1}]$ is an i,j path in $G(H)$. If $l = 1$, (i,j) is an edge of $G(H)$ hence a 1-monotone (i,j) path exists in $G(A)$. Otherwise $p_1 = [k_1,...,k_q]$ and $p_2 = [k_q,...,k_{l+1}]$, for any $2 \leqslant q \leqslant l$, are shorter paths in $G(H)$ implying, by induction, special paths in $G(A)$ and, hence, a special (i,j) path. □

Consider (4.1) and (4.2) where A has a block Gauss-Seidel splitting $A = (D-L)-U$. Then, as in (4.2), the iteration matrix $H = (D-L)^{-1}U$ can be rewritten as $H = (I-L_1)^{-1}U_1$ where $L_1 = D^{-1}L$ and $U_1 = D^{-1}U$. In the present context D^{-1} will be a block diagonal matrix with full diagonal blocks since $D_{ii}^{-1} > 0$. This means that L_1 and U_1 will be less sparse than L and U. More precisely, if $L_{ij}(U_{ij})$ has a nonzero in the kth column, then $D_{ii}^{-1}L_{ij}$ $(D_{ii}^{-1}U_{ij})$ has its kth column completely nonzero (in fact, positive). We will use Propositions 5 and 6 to examine the relation between $G(A)$, $G(A_1 = I-L_1-U_1)$ and $G(H)$.

Consider a block square matrix A and graphs $G(A) = (V,E)$ and $\mathbf{G}(A) = (\mathbf{V},\mathbf{E})$. Let $(V_i,V_j) \in \mathbf{E}$, $i \neq j$. Then $(v_k,v_l) \in E$ for some $v_k \in V_i$ and $v_l \in V_j$. We will say that $\mathbf{G}(A)$ *faithfully represents* $G(A)$ if: $(x,y) \in E$ for some $x \in V_i$, $y \in V_j$ with $i \neq j$ implies $(v,y) \in E$ for all $v \in V_i$. It is clear from our discussion above that $\mathbf{G}(A_1)$ faithfully represents $G(A_1)$ when $A_1 = I-L_1-U_1$ and

$$L_1 = D^{-1}L,\ U_1 = D^{-1}U\ . \tag{4.6}$$

If $G(A) = (V,E)$ is strongly connected, then $\mathbf{G}(A) = \mathbf{G}(A_1)$ is strongly connected but $G(A_1)$ is usually not. Note that in $G(A_1)$ no vertex $x \in V_i$ is adjacent to any vertex $y \in V_i$ except $y = x$; that is, the subgraphs of $G(A_1)$ induced by the V_i contain k_i nonadjacent loops where D_i is $k_i \times k_i$. For each V_i let $W_i = \{w \in V_i | (x,w) \in E, x \in V_j, i \neq j\}$; W_i can be viewed as the set of input vertices to V_i in $G(A_1)$. The structure of $G(A_1)$ is as follows.

PROPOSITION 7. *Let $A = D-L-U$ and $A_1 = I-L_1-U_1$ with D block diagonal, $D_{ii}^{-1} > 0$ and $i \geqslant 2$, L and U as in Proposition 5, and L_1 and U_1 as in (4.6). Then $\mathbf{G}(A_1)$ faithfully represents $G(A_1)$. If there exists an x,y path in $G(A)$ for $x \in V_i$ and $y \in V_j$, $i \neq j$, then there exist x,y paths in $G(A_1)$ for all $x \in V_i$. Furthermore, if $G(A)$ is strongly connected, the subgraph of $G(A_1)$ induced by $S = \bigcup_i W_i$ is strongly connected; each vertex $v \in V-S$ is a strong component loop adjacent only to a nonempty set of vertices in S.*

Proof. The first statement summarizes the previous discussion.

Suppose there exists an x,y path in $G(A)$ with $x \in V_i$, $y \in V_j$, $i \neq j$. Then there exists a V_i,V_j path in $\mathbf{G}(A) = \mathbf{G}(A_1)$, say $p = [V_i = V_{k_1},V_{k_2},...,V_{k_t} = V_j]$, with each vertex V_{k_t} distinct. Since $\mathbf{G}(A_1)$ faithfully represents A_1, each vertex of V_{k_t} is adjacent to the same vertices of $W_{k_{l+1}}$ in $G(A_1)$. Hence x,y paths exist in $G(A_1)$ for all $x \in V_i$.

To show that the subgraph of $G(A_1)$ induced by S is strongly connected we need only consider any $x \in W_i \subset V_i$ and $y \in W_j$ with $i \neq j$ since $i \geqslant 2$. An x,y path exists in $G(A)$ implying, as shown above, an x,y path in $G(A_1)$ using only vertices in the W_k.

Since $\mathrm{G}(A_1)$ is strongly connected and faithful each vertex in V_i, and hence $V_i - W_i$, is adjacent to each vertex in some W_j, $i \neq j$. Finally, since distinct vertices in V_i are not adjacent and since no vertex in $V_i - W_i$ is adjacent to any $V_j - W_j$, $i \neq j$, by definition of the W_i, each vertex in $V-S$ is a strong component adjacent only to vertices in S. □

We see now that for strongly connected $G(A)$ and A_1 as above, the iteration graph $G(H)$, $H = (I-L_1)^{-1}U_1$ has a structure like $G(A_1)$.

THEOREM 3. *Let A and $A_1 = I - L_1 - U_1$ be as in Proposition 7 with $G(A)$ strongly connected and $H = (I-L_1)^{-1}U_1$. Then $G(H)$ has a unique nontrivial strong component, $C = (Z,F)$. Each vertex $v \in V-Z$ is a trivial strong component adjacent only to a nonempty set in Z.*

Proof. We combine the results of Proposition 7 and Proposition 6 applied to $G(A_1)$. Note first that S of Proposition 7 has a least two vertices; since $G_1[S]$ is strongly connected, there exists a special k,k path (cycle) and hence a k,k cycle in $G(H)$ where $k = \max(i \mid i \in S)$. Let $C = (Z,F)$ be the strong component of $G(H)$ containing k. Let $x \in V-Z$ and $y \in V$. We show that if there exists an x,y path in $G(H)$, then $y \in Z$.

Note that the structure of $G(A_1)$ implies that special v,k paths exist in $G(A_1)$ for all $v \in V$. Hence there are v,k paths in $G(H)$ for all $v \in V$. Suppose there is an x,y path in $G(H)$; i.e., a special x,y path in $G(A_1)$. Then x and y can be restricted to the set S since there are no x,y paths in $G(A_1)$ for $y \in V-S$ (except loops) and any x,y path for $x \in V-S$, $y \in S$ implies an x,y path where $x \in S$ and $y \in S$ by Proposition 7. Furthermore this x,y path is a special x,y path since vertices in any V_j are ordered sequentially. So impose this restriction and note that since $G_1[S]$ is strongly connected there exists a k,x path in $G(A_1)$. Combining this k,x path and the special x,y path implies a special k,y path in $G(A_1)$. Hence there is a k,y path in $G(H)$, as well as the y,k path, so $y \in Z$. □

The structure of $G(H)$ is further revealed by the following observation. If any $w \in W_i$ of Proposition 7 is in $V-Z$ then $W_i \subset V-Z$ and hence $V_i \subset V-Z$. This follows since the vertices in the V_i are ordered sequentially, and if there is no special k,w_i path for some $w_i \in W_1$ there is no special k,w_i for any $w_i \in W$. Clearly, there is no k,v path for $v \in V_i - W_i$ at all.

So far we have considered the setting $A = D-L-U$ where D is block diagonal and L and U are strictly block triangular. We now consider more general block splitting, $A = (D-L)-N$, as discussed in § 3 with D and L as before and N as in (3.5). We write

$$N = L(N) + D(N) + U(N) \tag{4.7}$$

and

$$N_1 = D^{-1}N \equiv L_1(N) + D_1(N) + U_1(N)\ . \tag{4.8}$$

$L(N)$ and $L_1(N)$ are strictly block lower triangular, $D(N)$ and $D_1(N)$ are block diagonal, and $U(N)$ and $U_1(N)$ are strictly block upper triangular. All these matrices are nonnegative as are L, D, and $L_1 = D^{-1}L$. Note that we may write

$$(4.9) \qquad (I-L_1)^{-1} = I+L_2\,, \qquad L_2 \geqslant 0\,.$$

Theorem 3 provides the structure of $H_1 = (I-L_1)^{-1}U_1(N)$ where $G(A)$ is strongly connected. We need to examine $H = (I-L_1)^{-1}N_1$; observe that

$$(4.10) \qquad H = H_1+H_2+H_3\,,$$

H_1 as above, and,

$$(4.11) \qquad H_2 = (I+L_2)L_1(N)+L_2D_1(N)\,,$$

$$(4.12) \qquad H_3 = D_1(N)\,.$$

Note that H_2 is strictly block lower triangular, and recall that H_3 is block diagonal with nonnegative diagonal blocks, say $H_{3i} = D_{ii}^{-1}D_{ii}(N)$, as specified in (4.7) and (4.8). An important point is that the entire kth column of H_{3i}, especially the k,k diagonal entry, if any entry in the kth column of $D_{ii}(N)$ is positive.

Graph-theoretically, the situation is as follows. $G(H_1) = (V,F_1)$ is a graph as described by Theorem 3. To this graph edges F_2 and F_3 will be added (unioned) corresponding to edges in $G(H_2)$ and $G(H_3)$ to produce $G(H)$. Note that $G(H_2)$ is acyclic; even more, edges in F_2 are of the form (v_i,v_j), $v_i \in V_i$ and $v_j \in V_j$, $i \neq j$, the V_k as in the discussion surrounding Theorem 3. Since $G(H_2)$ is acyclic, the added edges F_2 will cause no cycles between the vertex sets $V_i \subset V-Z$, although some of the vertices in these subsets may merge into the strong component containing Z. Still there remains only one nontrivial strong component in $G = (V,F_1 \cup F_2)$.

Consider now the addition of the edges in F_3 to G. Again we need only be concerned with vertex sets $V_i \subset V-Z$. Suppose some edge $(x,y) \in F_3$ causes a cycle when added to G, $y \in V_j \subset V-Z$, say. Then there is a positive entry, say d_{kl}, of the matrix $D_{jj}(N)$ with y represented by the column index l, and hence $H_{3i} = (h_{ij})$ has $h_{kl} \neq 0$ for all k. Thus $(x,y) \in F_3$ for all $x \in V_j$, including the loop (y,y). This means that any new strong components caused by adding F_3 to make the final graph $G(H)$ will be primitive. We summarize the situation in

PROPOSITION 8. *Let $A = D-L-N$ and $A_1 = I-L_1-N_1$, D and L as in Proposition 7 and N and N_1 as in (4.7) and (4.8). Then $G(H)$, $H = (I-L_1)^{-1}N_1$, is strongly primitive if $G(H_1)$, $H_1 = (I-L_1)^{-1}U_1(N)$ is strongly primitive.*

We are now in a position to complete the proof of Theorem 1, § 3. We will show $G(H)$ is strongly primitive where

$$H = M^{-1}N = (D-L)^{-1}N = (I-L_1)^{-1}N_1\,,$$

N_1 as in (4.8).

Proof of Theorem 1.

[1] The matrix $A_0 = D-L-U(N)$ of condition (iii) in the definition of an R-regular splitting will play the role of A in the application of Theorem 3. $A_1 = I-L_1-U_1(N)$. Since A_0 is irreducible, $G(H_1)$, $H_1 = (I-L_1)^{-1}U_1(N)$ has the structure of $G(H)$ of Theorem 3.

[2] We show $G(H_1)$ is strongly primitive. First recall that $\mathbf{G}(A_0)$ has a monotone cycle and note $\mathbf{G}(A_0) = \mathbf{G}(A_1)$. This implies that $G(A_1)$ also has a monotone cycle since vertices in the subsets V_k of Proposition 7 and Theorem 3 are ordered sequentially. Applying Corollary 3 with $A = A_1$ insures that $G(H_1)$ has a loop. Theorem 3 asserts that this loop must be in the nontrivial strong component of $G(H_1)$. Hence $G(H_1)$ is strongly primitive.

[3] By Proposition 8 $G(H)$ is strongly primitive where $H = (I-L_1)^{-1}N_1$. □

REFERENCES

[AHU] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[BP] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[G] F. R. Gantmacher, *The Theory of Matrices*, Vols. I and II. Chelsea, New York, 1959.

[KGW] L. Kaufman, B. Gopinath and E. Wunderlich, *Analysis of packet network congestion control using sparse matrix algorithms*, IEEE Trans. Communications, 29 (1981), pp. 453-465.

[KSM] L. Kaufman, J. Serry and J. Morrison, *Overflow models for dimension PBX feature packages*, Bell Syst. Tech. J., 60 (1981), pp. 661-676.

[V] R. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

SIAM J. ALG. DISC. METH.
Vol. 5, No, 2, June 1984

001

THREE VERSIONS OF A GROUP TESTING GAME*

F. K. HWANG†

Abstract. We consider a game between two players G and H such that G is to choose a subset from a given set and H is to identify G correctly by using a sequence of symmetric group tests to gather information. We study three versions of the game varying in degrees of restrictions imposed on G in choosing the subset. Namely, in the first version, G assumes no restriction; in the second version, G can choose only the cardinality of the set; and in the third version, the choice is random. We determine the value of the game for the first version and give strategies for H which yield good upper bounds on the value of the game for the two other versions.

1. Introduction. We introduce a game which is derived from the symmetric group testing scheme as proposed by Kumar, Sobel and Blumenthal [1]. There are two players in this game. To emphasize their different roles, we will call them G (a female) and H (a male) respectively. The equipment for the game can be assumed to be a set of n coins; all look alike except that some coins are genuine and some are counterfeit. We will denote the set of coins by N and the *counterfeit subset* by D. Then the game is for G to choose D and for H to choose an algorithm which identifies D, where an algorithm here means a sequence of symmetric group tests. A *symmetric group test* can be applied on any subset of coins g, and there are three possible outcomes, labeled genuine, counterfeit and mixed. The *genuine outcome* reveals that all coins in g are genuine. The *counterfeit outcome* reveals that all coins in g are counterfeit, and the *mixed outcome* reveals that g contains some genuine and some counterfeit coins, but is otherwise ambiguous. Note that H can always identify the set D correctly by using the algorithm which tests the coins one by one. Therefore, we assume that H chooses from only those algorithms which identify D correctly, and an algorithm is evaluated by the number of tests it uses.

The game can be played in three different ways:

(i) G is almighty (*A-version*). In this version G always knows in advance what algorithm H selects and G makes sure that the set D is chosen in such a way that it is most unfavorable to the particular algorithm chosen. In game theoretic terminology, the *payoff* to G given T is

$$\max_D f(D, T),$$

where $f(D, T)$ is the number of tests consumed by T when D is the counterfeit subset. The *value* of the game to G is

$$V(A) = \min_T \max f(D, T).$$

(ii) G is just (*J-version*). In this version G wants to play a fair game and therefore chooses D without using her prescience power. Since the coins all look alike to H, G, by relinquishing her foresight, has also no reason to distinguish the choice of one subset over another for D as long as they have the same cardinality. Therefore, we can consider the payoff to G given T as

$$\max_d f(d, T).$$

* Received by the editors November 18, 1980, and in final revised form March 25, 1983.

† Bell Laboratories, Murray Hill, New Jersey 07974.

$f(d, T)$ is the expectation of $f(D, T)$ over all D with cardinality d. The value of the game to G is

$$V(J) = \min_T \max_d f(d, T),$$

(iii) G is merciful (*M-version*). In this version G does not choose D to maximize $f(D, T)$. Instead, she chooses D randomly from all subsets of N and informs H so. Strictly speaking, a game in the game theoretic sense no longer exists when G does not act as an adversary of H. But we use the same terminology as before. The payoff to G given T is now simply the expected number of tests:

$$Ef(D, T) = \sum_{D \subseteq N} \frac{1}{2^n} f(D, T) = \sum_{d=0}^{n} \frac{\binom{n}{d}}{2^n} f(d, T).$$

We also define

$$V(M) = \min_T \sum_{d=0}^{n} \frac{\binom{n}{d}}{2^n} f(d, T).$$

In this paper we study all three versions of the symmetric group testing game. We first prove that $V(A) = n$. Then we give two algorithms for the J-version and study their payoff functions. We also give an algorithm which we prove to be near optimal and which we conjecture to be optimal for the M-version. Finally we mention that the M-version can be generalized to include the binomial population case studied in [1]. We give an asymptotic analysis for two algorithms for the binomial case.

2. Some preliminary remarks. A *tertiary tree* is a rooted tree such that each node has indegree one (except the root, which has indegree zero), and outdegree at most three. Nodes with outdegree zero are called *terminal nodes* and nodes with positive outdegrees are called *internal nodes*. The *path* for a node v is the alternating sequence of nodes and links connecting the root with v, but not including v. The *length* of a path is the number of nodes in it. The sum of all terminal node path lengths is also known as the *cost* of a tree. Let $C(T)$ denote the cost of a tree T rooted at v with n terminal nodes. Let $v_i, i = 1, 2, \cdots, j, j \leqq 3$ denote the nodes at the other ends of the j outlinks of v and let $T_i, i = 1, 2, \cdots, h$, denote the subtrees of T rooted at v_i. Then it can be easily verified that:

THE DECOMPOSITION LEMMA. $C(T) = \sum_{i=1}^{j} C(T_i) + n$.

A symmetric group testing algorithm can be represented by a tertiary tree where each internal node v is associated with a test $g(v)$ and the three outlinks of v are associated with the three possible outcomes of the test. The *test history* at a node v is the sequence of tests and outcomes associated with the path for v. Any subset of N which is consistent with the test history at v is called a *solution point* at v. $S(v)$, the set of all solution points at v, is called the *solution space* at v. Note that the solution space at any terminal node must consist of a single solution point.

From now on an algorithm will always be interpreted in its tree form. Suppose that H chooses the algorithm T and G chooses the counterfeit subset D. The number of tests needed to identify the n coins is simply the path length of the terminal node v of T with $S(v) = \{D\}$. The payoff to G is then the maximum path length of T for the A-version, and the average path length of T for the M-version. For the J-version, let $l(d)$ denote the average path length over all terminal nodes v such that $S(v)$

consists of a solution point with d elements. Then the payoff to G is the maximum of $l(d)$ over all d.

Let T denote an algorithm for identifying the counterfeit subset in a set of n coins. Since the solution space of the root of T contains 2^n points and each test is at most a 3-partition, the maximum path length of T is at least $\lceil \log_3 2^n \rceil$, where $\lceil x \rceil$ denotes the smallest integer not less than x, and the average path length is at least $\log_3 2^n \cong .631\, n$. Define

$$I_n = \begin{cases} \lceil \log_3 2^n \rceil & \text{for the } A\text{-version,} \\ \log_3 2^n & \text{for the } J\text{- and the } M\text{-versions.} \end{cases}$$

Then I_n is a lower bound for the value of the symmetric group testing game. We call it the *information bound.*

An algorithm is called *regular* if it satisfies the following three conditions:

(i) A test will not be performed if the outcome is known in advance,

(ii) No test will include coins already identified,

(iii) Two coins known to have the same identification will not be included in the same test.

It is clear that any irregular algorithm can be made regular without increasing the number of tests for any D chosen. Therefore, from now on, we assume that we only deal with regular algorithms. Note that condition (iii) implies that we can never rule out the possibility of a mixed outcome for any test involving more than one coin.

3. The A-version.

In this section we prove that $V(A) = n$, the number of coins in the given set N.

THEOREM 1. *In the A-version, n tests are necessary and sufficient to identify n coins.*

Proof. The sufficiency of n tests is immediate, since this can always be achieved by testing each coin individually.

To show that n tests are required, consider the variant of the problem in which, whenever a "mixed" outcome is obtained, G also provides H with a specific pair of coins that have opposite identifications within that mixed group (but without specifying which is genuine and which is counterfeit). Certainly this cannot increase the maximum number of tests required, since H now has more information. We shall show that even in this case, given any algorithm T, there exists a consistent sequence of responses for G that forces at least n tests to be used.

The strategy that G uses is essentially to respond "mixed" whenever possible. In order to keep track of the responses made so far and their consequences, G maintains a *graph structure*, whose vertices correspond to the unidentified coins, plus the *sets* of coins already identified as genuine and counterfeit. Two vertices in the graph will be joined by an edge if and only if they are known to have opposite identifications. Thus, by transitivity, any two vertices connected by a path with an odd number of edges must have different identifications and hence will be joined by an edge. It follows from this that each connected component of the graph will be a complete bipartite subgraph.

Now we describe how G will respond to each test in sequence and how the graph will be altered in the process. If a single coin is tested, then G responds "genuine". At this point all coins in the same component with that coin have been identified (coins on the same side of the bipartite graph are genuine and coins on the other side are counterfeit), so the entire component is removed from the graph.

If a group test is applied, the regularity conditions imply that it must involve coins from more than one component. In this case, G responds "mixed" and chooses two coins from different components that she informs H have opposite identifications. This adds to the graph the edge joining the corresponding two vertices and all edges that follow from this, merging the two components into a single complete bipartite subgraph.

It is now a simple observation that each test can only reduce the number of connected components in the graph by 1. Since the graph initially has n connected components, each an isolated vertex, and since the last test must reduce the number of connected components to zero (all coins identified), it follows that at least n tests are required. The proof is complete.

Notice that, if H is restricted to algorithms that test only one or two element sets, then he will be in a situation where a "mixed" outcome *automatically* provides a pair of oppositely identified items. Thus, in this case, the graph structure described in the proof provides a particularly convenient way for H to keep track of what he knows from the tests performed so far. In the following sections we will be especially interested in the class P of such testing algorithms, and we will see that this viewpoint is a useful one.

4. The *J*-version. We give two algorithms, both in class P, for the J-version.

Assume that the n coins are labeled by the numbers 1 to n. The *chain algorithm*:

(i) Test coins 1 and 2 as a pair.

(ii) Suppose the current test is on coins i and $i+1$. If the outcome is mixed, test the pair of coins $i+1$ and $i+2$ next; if not, test the pair of coins $i+2$ and $i+3$.

The *star algorithm*:

(i) Test coins 1 and 2 as a pair.

(ii) Suppose the current test is on coins i and j with $i<j$. If the outcome is mixed, test coins i and $j+1$ next; if not, test coins $j+1$ and $j+2$ next.

It should be understood that for both procedures, if some coin in the prescribed pair is not available, then we simply test what is available (and do not test if nothing is available).

Note that, for both these procedures, all coins which have been tested but not yet identified belong to the same connected component of the previously mentioned graph structure. Therefore any test with a nonmixed outcome identifies all coins previously tested and results in a subproblem with smaller values of n and d. Furthermore this recurrence is bound to happen except in the following two cases (without loss of generality assume $d \leqq n/2$):

(i) $2d-1 \leqq n \leqq 2d$ for the chain algorithm,

(ii) $d=1$ for the star algorithm.

These observations allow us to write down the recurrent equations (where the subscript C is for the chain algorithm, the subscript S is for the star algorithm and E is for expectation):

$$E_C(0,n)=\left\lceil \frac{n}{2} \right\rceil \qquad (\lceil x \rceil \text{ denotes the smallest integer} \geqq x),$$

$$E_C(d,n)=\sum_{i=1}^{d} \frac{\binom{n-2i}{d+1-i}}{\binom{n}{d}} \{E_C(d+1-i,\, n-2i)+2i-1\}$$

$$+\sum_{i=1}^{d-1}\frac{\binom{n-2i}{d-1-i}}{\binom{n}{d}}\{E_C(d-1-i,n-2i)+2i-1\}$$

$$+\sum_{i=1}^{d}\frac{\binom{n-1-2i}{d-i}}{\binom{n}{d}}\{E_C(d-i,n-1-2i)+2i\}$$

$$+\sum_{i=1}^{d-1}\frac{\binom{n-1-2i}{d-1-i}}{\binom{n}{d}}\{E_C(d-1-i,n-1-2i)+2i\}+\frac{2}{\binom{n}{d}}\delta n,$$

where $\delta=1$ if $2d-1\leqq n\leqq 2d$, $\delta=0$ otherwise;

$$E_S(0,n)=\left\lceil\frac{n}{2}\right\rceil,$$

$$E_S(d,n)=\sum_{i=0}^{d}\frac{\binom{n-2-i}{d-i}}{\binom{n}{d}}\{E_S(d-i,n-2-i)+i+1\}$$

$$+\sum_{i=0}^{n-d}\frac{\binom{n-2-i}{d-2}}{\binom{n}{d}}\{E_S(d-2,n-2-i)+i+1\}+\frac{1}{\binom{n}{d}}\delta n,$$

where $\delta=1$ if $d=1$, and $\delta=0$ otherwise. (It is understood that $\binom{x}{y}=0$ for any negative x or y.) Define

$$C(d,n)=E_C(d,n)\binom{n}{d}$$

and

$$S(d,n)=E_S(d,n)\binom{n}{d}.$$

Then the recurrence equations for $E_C(d,n)$ and $E_S(d,n)$ can be changed into recurrence equations for $C(d,n)$ and $S(d,n)$ to allow us to work with integral numbers.

Values of $C(d,n)$ and $S(d,n)$ for $0\leqq d\leqq n\leqq 10$ are given in Tables 1 and 2 respectively ($d_T(n)$ is defined as those $d\leqq n/2$ for which $E_T(d,n)$ is maximum).

Three conjectures seem to emerge from Tables 1 and 2:

(i) $(1/n)E(d_C(n),n)\downarrow_n$.

(ii) $d_C(n)=\lfloor n/2\rfloor$ ($\lfloor x\rfloor$ denotes the largest integer $\leqq x$).

(iii) $(1/n)E(d_S(n),n)\geqq(1/n)E(d_S(n+2),n)$.

Table 3 gives values of $d_C(n)$, $d_S(n)$, $(1/n)E(d_C(n),n)$, $(1/n)E(d_S(n),n)$ for $11\leqq n\leqq 20$. The credibility of the above three conjectures is enhanced.

TABLE 1
$C(d, n)$

d \ n	0	1	2	3	4	5	6	7	8	9	10
0	0	1	1	2	2	3	3	4	4	5	5
1		1	4	7	12	17	24	31	40	49	60
2			1	7	18	37	64	102	151	214	291
3				2	12	37	88	176	316	523	816
4					2	17	64	176	400	802	1467
5						3	24	102	316	802	1776
6							3	31	151	523	1467
7								4	40	214	816
8									4	49	291
9										5	60
10											5
$d_C(n)$	0	0	1	1	1, 2	2	3	3	4	4	5
$(1/n)E(d_C(n), n)$	0	1	1	.778	.75	.74	.733	.718	.714	.707	.705

TABLE 2
$S(d, n)$

d \ n	0	1	2	3	4	5	6	7	8	9	10
0	0	1	1	2	2	3	3	4	4	5	5
1		1	4	7	13	18	27	34	46	55	70
2			1	7	16	36	60	102	148	220	295
3				2	13	36	90	173	323	525	845
4					2	18	60	173	380	788	1423
5						3	27	102	323	788	1778
6							3	34	148	525	1423
7								4	446	220	845
8									4	55	295
9										5	70
10											5
$d_S(n)$	0	0	1	1	1	1, 2	1, 3	3	3	4	5
$(1/n)E(d_S(n), n)$	0	1	1	.778	.813	.72	.75	.706	.721	.695	.706

TABLE 3
An extended table

n	11	12	13	14	15	16	17	18	19	20
$d_C(n)$	6	6	7	7	8	8	9	9	10	10
$(1/n)E(d_C(n), n)$	.700	.698	.6949	.6936	.6912	.6893	.6883	.6875	.6860	.6854
$d_S(n)$	5	5	5	7	7	7	8	9	9	9
$(1/n)E(d_S(n), n)$	.690	.696	.6859	.6902	.6835	.6855	.6810	.6829	.6796	.6803

Furthermore, values in Table 3 throw support to two new conjectures.

(iv) $E(d_C(n), n) > E(d_S(n), n)$ for $n > 11$.

(v) $d_S(n) = 2\lceil n/4 \rceil - 1$ except for $n \equiv 1 \pmod 4$.

For large n, we can divide the n coins into smaller piles, say, n_i coins for the ith pile, and use an algorithm designed for n_i coins for the ith pile. Then the expected number of tests for the n coins is the sum of the expected numbers of tests over all piles. For example, if each pile contains 19 coins, then

$$\frac{1}{n}\max_d E_S(d, n) = \tfrac{1}{19}\max_d E_S(d, 19) = .6796,$$

which is less than eight percent over the information bound .631.

5. The *M*-version. Let P' be the subclass of algorithms in P in which a single coin is tested only for the last test, i.e., when all unidentified coins belong to the same connected component. We show that any algorithm in P' has the same expected number of tests for the M-version and we derive this number explicitly.

Let T be an algorithm in P'. Suppose that at the internal node v, T tests the two coins X and Y. Then X and Y must be in separate components by the regularity condition (iii). Certainly both X and Y can be either genuine or counterfeit. Furthermore, since they are in separate components, their identifications are independent. Therefore all four possible combinations can occur. We now show that the solution points which are consistent with each of the four combinations are equinumerous. Let s be a solution point consistent with one of the four combinations. Let s_X be obtained from s by changing the identification of X and all other coins in its component. Similarly we can define s_Y and s_{XY}. Then s, s_X, s_Y and s_{XY} all correspond to different combinations of possibilities for X and Y, and hence the four solution subspaces corresponding to these four combinations are equinumerous. This implies that a test of a pair always distributes one fourth of the solution points to each of the genuine outcome and the counterfeit outcome, and two fourths to the mixed outcome. Since the cardinality of the solution space at the root of T is 2^n, the cardinality of the solution spaces at all internal nodes must be a power of two. Let $f(i)$ denote the cost of a tree rooted at a node whose solution space has 2^i elements. Then

$$f(0) = 0,$$

$$f(1) = 2,$$

$$f(i) = 2f(i-2) + f(i-1) + 2^i \quad \text{for } 2 \leqq i \leqq n, \text{ by the decomposition lemma.}$$

Let $F(t)$ denote the generating function of $f(i)$. A straightforward analysis shows

$$F(t) = \frac{2t}{(1-2t)^2(1+t)}.$$

Furthermore, it is easy to verify that

$$f(n) = \frac{2}{3}\left(n2^n + \left\lfloor \frac{2^n+1}{3} \right\rfloor\right).$$

Therefore,

$$\tfrac{2}{3}n2^n \text{ is a good asymptotic approximation for } f(n).$$

Note that

$$\frac{f(n)/2^n}{I(n)}=\frac{2/3(n2^n+\lfloor(2^n+1)/3\rfloor)}{(\log_3 2)n2^n}\leqq\begin{cases}1.41 & \text{for } n\geqq 1,\\ 1.10 & \text{for } n\geqq 10,\\ 1.06 & \text{for } n\geqq 100.\end{cases}$$

In fact, we conjecture that $V(M)=f(n)/2^n$.

6. Some concluding remarks. We can extend the M-version to cases where G chooses D with a probability distribution and informs H about it. A random choice of D is equivalent to a uniform probability distribution. Another simple case is the binomial distribution in which

$$P(D=N')=\binom{n}{n'}p^{n'}(1-p)^{n-n'}\quad\text{for all } N'\subseteq N,$$

where n' is the cardinality of N' and p is the probability of any coin being defective. This is the case studied in [1]. While several procedures had been proposed in [1], due to computational difficulties, the expected numbers of tests for these procedures were given only for small values of n.

Note that when p is small, then the advantage of a symmetric group test over an ordinary group test (which does not have the counterfeit outcome) is much less since it is unlikely that all coins in the test group are counterfeit. The "large p" case is mathematically equivalent to the "small p" case by reversing the definitions of genuine and counterfeit. Therefore, the interesting case for symmetric group testing is for p to be close to .5, or similarly, the number of genuine coins and the number of counterfeits are close. The algorithms in the class P' are particularly suitable for such situations. We now give an asymptotic analysis for the chain algorithm and the star algorithm.

For n large, both the chain algorithm and the star algorithm can be treated as recurrent processes where the recurrent points are coincident with either a genuine outcome or a counterfeit outcome. Therefore, the expected number of tests per coin can be approximated by the ratio

$$\frac{\text{expected number of tests between two recurrent points}}{\text{expected number of coins identified between two recurrent points}}$$

$$=1-\frac{1}{\text{expected number of coins identified between two recurrent points}}$$

since 1 + numerator = denominator.

For the chain procedure, the denominator equals

$$\begin{aligned}&2(p^2+q^2)+3(p^2q+q^2p)+4(p^3q+q^3p)+5(p^3q^2+q^3p^2)+\cdots\\ &\quad=2(p^2+q^2)(1+2pq+3(pq)^2+\cdots)+3pq(1+pq+(pq)^2+\cdots)\\ &\qquad+2(pq)^2(1+pq+(pq)^2+\cdots)\\ &\quad=\frac{2(p^2+q^2)}{(1-pq)^2}+\frac{3pq}{1-pq}+\frac{2(pq)^2}{(1-pq)^2}=\frac{2+pq}{(1-pq)}.\end{aligned}$$

Therefore, the expected number of tests per coin is $(1+2pq)/(2+pq)$.

For the star procedure, the denominator equals

$$\begin{aligned}&2(p^2+q^2)+3(p^2q+q^2p)+4(p^2q^2+q^2p^2)+5(p^2q^3+q^2p^3)+\cdots\\&=p^2(2+3q+4q^2+\cdots)+q^2(2+3p+4p^2+\cdots)\\&=\frac{p^2(2-q)}{(1-q)^2}+\frac{q^2(2-p)}{(1-p)^2}=3.\end{aligned}$$

Therefore the expected number of tests per coin is $\frac{2}{3}$. Note that

$$\frac{1+2pq}{2+pq}\leqq\frac{2}{3}\quad\text{for all } p.$$

Hence the chain algorithm is always better than the star algorithm for n large (actually it is also true for small n), an interesting contrast to conjecture (iv) of § 4.

Another interesting observation is that as the options of G decrease and the information to H increases from the A-version to the J-version, to the M-version, and to the binomial population case, the value of the game to H seems to decrease steadily as suggested from the upper bounds we obtain, i.e., from n to .68 to $\frac{2}{3}$ to $(1+2pq)/(2+pq)$.

Acknowledgment. The author wishes to thank M. R. Garey for a careful reading of the paper and useful suggestions.

REFERENCE

[1] M. Sobel, S. Kumar and S. Blumenthal, *Symmetric binomial group-testing with three outcomes*, in Statistical Decision Theory and Related Topics, S. S. Gupta and J. Yackel, eds., Academic Press, New York, 1971.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

PERFORMANCE OF HEURISTICS FOR A COMPUTER RESOURCE ALLOCATION PROBLEM*

MICHAEL A. LANGSTON†

Abstract. Attention is given to the problem of maximizing the number of pieces packed into a fixed set of bins whose capacities may differ. Instances of this problem arise in the allocation of computer resources such as processors and memory. Since the problem is NP-hard, the worst-case behavior of several approximation algorithms is investigated. In particular it is shown that the asymptotic worst-case performance bound of the first-fit-increasing rule is 2, while the iterated first-fit-decreasing heuristic can be implemented so that its asymptotic bound does not exceed $\frac{3}{2}$.

Key words. approximation algorithms, worst-case analysis, bin-packing, storage allocation, multiprocessor scheduling

AMS subject classification. 68-E99

1. Introduction. In a variant of the classical one-dimensional bin-packing problem, we seek to maximize the number of pieces packed into a fixed collection of finite-capacity bins. Motivations for the study of this problem include the desire to maximize the number of variable-length records stored in a multivolume memory and to maximize the number of independent tasks which can be executed by a multiprocessor system in a given time period.

For the abstract bin-packing model we consider in this paper, it is assumed that the bins may differ in size. Thus, our problem is mathematically equivalent to the optimization criteria mentioned above when storage volumes may have unequal capacities or when processors may differ in speed.

Since it is easily confirmed [GJ] that our problem is NP-hard, we focus our attention on efficient (polynomial-time) approximation algorithms in an effort to guarantee near-optimal packings. Worst-case performance ratios are employed to gauge the relative merit of several packing heuristics. Our research parallels the pioneering efforts of two earlier publications [CLT] and [CL], wherein the investigation was restricted to problem instances involving bins of equal size.

The next section introduces the notation we use to discuss and examine the behavior of packing algorithms. Sections 3 and 4 contain worst-case analyses for the smallest-piece-first and the first-fit-increasing rules, respectively. We show that each may asymptotically pack only half the optimal number of pieces, but no fewer. In § 5 we study the iterated first-fit-decreasing heuristic. We prove that it can be implemented so as to guarantee asymptotically packing at least two-thirds the optimal number of pieces. In the final section we list a few conclusions drawn from this effort.

2. Notation. Let $L=\{p_1, p_2, \cdots, p_N\}$ denote a list of N pieces. Let $B=\{B_1, B_2, \cdots, B_M\}$ designate a set of M bins. We use a function $s: L \cup B \rightarrow R^+$ to identify the size of each piece and bin.

Let $n_{\text{ALG}}(L, B)$ represent the number of pieces of L packed into B by an algorithm ALG. Let $n_0(L, B)$ stand for the maximum achievable number. We define the asymptotic worst-case performance bound of ALG as the least real number R such that, for all L and B, $n_0(L, B) \leqq R n_{\text{ALG}}(L, B) + C$ where C is some constant.

* Received by the editors March 1, 1982, and in revised form April 20, 1983.

† Computer Science Department, Washington State University, Pullman, Washington 99164.

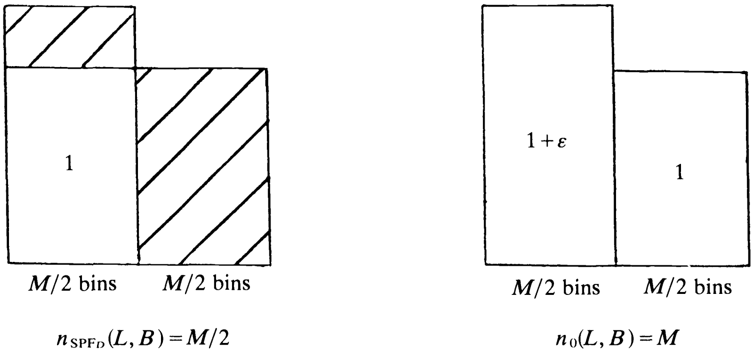
We will use the term "increasing" for the more precise but cumbersome "nondecreasing". Similarly, we will employ "decreasing" rather than "nonincreasing". ALG_I (ALG_D) denotes the implementation of ALG in which B is first sorted into a sequence of increasing (decreasing) bin sizes.

3. Smallest-piece-first. Observe that an optimal packing can, theoretically, always be constructed using only the $n_0(L, B)$ smallest pieces of L. Although we typically cannot afford the time required to generate such a packing, its existence suggests that assigning the pieces of L in order of increasing size should yield an effective approximation algorithm.

The smallest-piece-first (SPF) heuristic initially sorts L into an increasing sequence of piece sizes then assigns each piece in turn to the bin having the least filled space, with ties broken in favor of the bin having the lowest index. In other words, if we view B as arranged from left to right, SPF packs L in "horizontal" levels. The algorithm halts when the first piece is encountered which is too large to fit in any bin. This is natural, since no other unpacked piece can fit either. SPF is of time complexity $O(N \log N)$, dominated by the initial sort.

It is known [CLT] that when the bins of B have the same size the asymptotic worst-case performance bound for SPF is 2. That is, as N and M grow large, instances exist for which the ratio of the number of pieces contained in an optimal packing to the number in an SPF packing approaches arbitrarily close to 2. We now demonstrate that relaxing the aforementioned restriction on B has no significant effect on the worst-case behavior of SPF. It will be shown in the next two sections that the same cannot be said for some other packing algorithms.

To construct L and B such that $n_0(L, B) = 2n_{\mathrm{SPF}_D}(L, B)$, let $M = N = 2k$ for any positive integer k. Let $s(p_i) = s(B_{k+i}) = 1$ and $s(p_{k+i}) = s(B_i) = 1+\varepsilon$ for $1 \leqq i \leqq k$ and for some ε in the range $(0, 1)$. See Fig. 3.1. It is not difficult to prove that 2 is an upper limit as well for SPF_D's performance bound, but we turn our attention instead to SPF_I and show that its worst-case bound is slightly better (i.e., has the same asymptotic ratio but a better additive constant).



$n_{\mathrm{SPF}_D}(L, B) = M/2$ $\qquad$ $n_0(L, B) = M$

FIG. 3.1. *Packing instance in which* $n_0(L, B) = 2n_{\mathrm{SPF}_D}(L, B)$.

Using the example described in [CLT, Thm. 1], we display L and B such that $n_0(L, B) = 2n_{\mathrm{SPF}_I}(L, B) - 1$. Let $N = 2M - 1$, let $s(B_1) = s(B_2) = \cdots = s(B_M) = 1$, and let L contain M pieces of size $1/M$ and $M-1$ pieces of size 1. See Fig. 3.2. We now prove that this bound is tight.

THEOREM 3.1. *For all L and B, $n_0(L,B) \leqq 2n_{\mathrm{SPF}_I}(L,B)-1$.*

Proof. First we will show that we need only consider instances in which the SPF_I packing contains no empty bins. Suppose there are empty bin(s), the first of which is B_i, $1 \leqq i \leqq M$. Since $s(B_1) \leqq s(B_2) \leqq \cdots \leqq s(B_i)$, only the $i-1 \geqq 0$ smallest pieces of L will fit into $B_1 \cup B_2 \cup \cdots \cup B_i$ in any packing. Thus we may rearrange an optimal packing so that it leaves B_i empty as well. We may then delete B_i from B without affecting $n_{\mathrm{SPF}_I}(L,B)$ or $n_0(L,B)$. Repeated applications of this construction remove all empty bins from the SPF_I packing.

Hence $n_{\mathrm{SPF}_I}(L,B) \geqq M$. Simple capacity arguments guarantee that an optimization algorithm cannot find room for more than $M-1$ additional pieces as large as the first piece SPF fails to pack. □

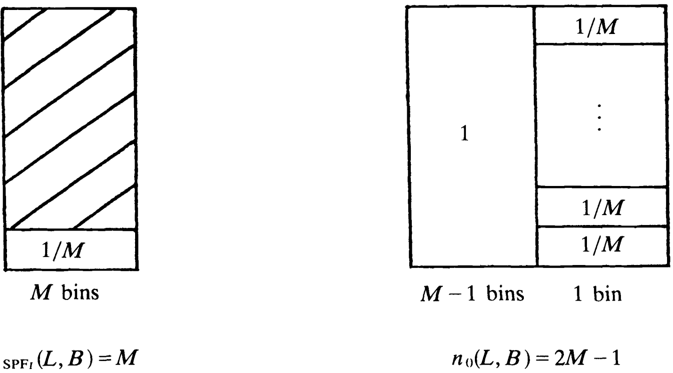


FIG. 3.2. *Packing instance in which $n_0(L,B)=2n_{\mathrm{SPF}_I}(L,B)-1$.*

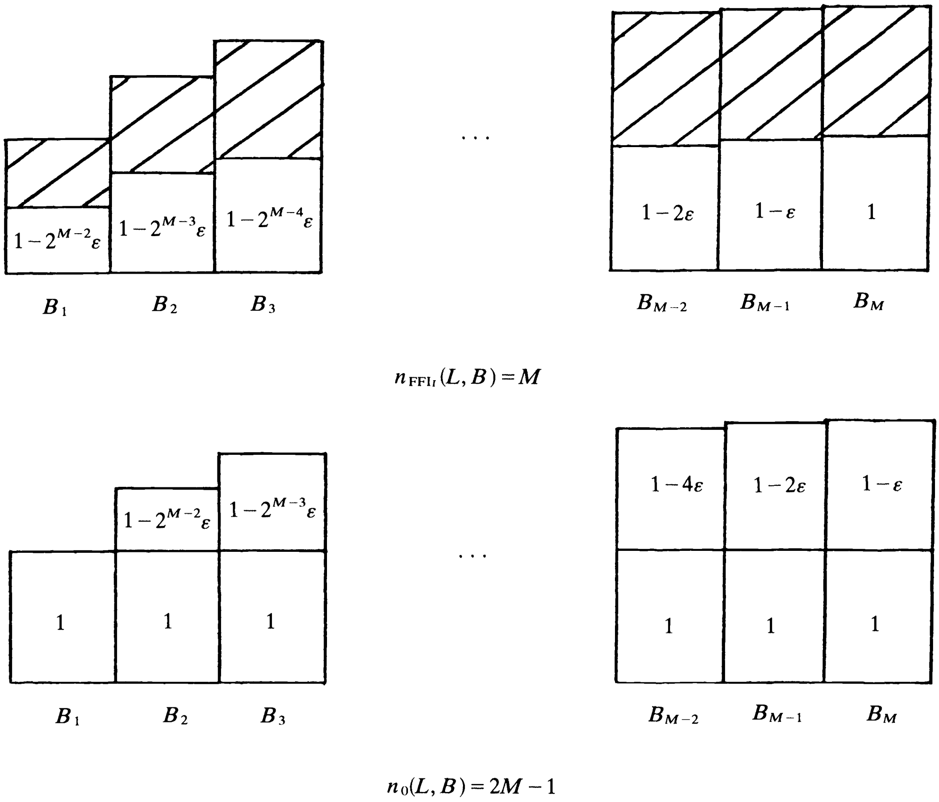
4. First-fit-increasing. Like SPF, the first-fit-increasing (FFI) rule is an attempt to insure good results by packing the pieces of L in order of increasing size. But FFI assigns each piece in turn to the lowest-indexed bin into which it will fit. In other words, FFI packs L "vertically", a bin at a time. The procedure terminates when it encounters a piece it cannot pack. Clearly FFI is of time complexity $O(N \log N)$.

When all bins must have equal capacity, FFI is substantially superior to SPF. With this restriction the worst-case performance bound for FFI is $\frac{4}{3}$ (see [CLT]). Allowing bin sizes to differ has a disastrous impact on FFI: we will show that FFI's asymptotic worst-case performance ratio is no better than that of SPF, namely 2.

For an L and B such that $n_0(L,B)=2n_{\mathrm{FFI}_D}(L,B)$, refer to Fig. 3.1 of the last section. It is relatively simple to show that this worst-case coefficient cannot exceed 2, but let us concentrate instead on FFI_I since its worst-case bound turns out to be slightly more attractive (i.e., has the same asymptotic ratio but a smaller additive constant).

A careful review of the proof of Theorem 3.1 shows that, with minor modifications, its arguments hold for FFI_I as well. Hence we know that, for all L and B, $n_0(L,B) \leqq 2n_{\mathrm{FFI}_I}(L,B)-1$.

We now demonstrate that this bound can in fact be achieved. Let $N=2M-1$. Select an ε in the range $(0, 2^{1-M}]$. Let $s(B_1)=1$ and let $s(B_j)=2-(2^{M-j})\varepsilon$ for $2 \leqq j \leqq M$. Let $s(p_j)=1-(2^{M-j-1})\varepsilon$ for $1 \leqq j \leqq M-1$ and $s(p_j)=1$ for $M \leqq j \leqq 2M-1$. See Fig. 4.1.



$n_{\mathrm{FFI}_I}(L,B)=M$

$n_0(L,B)=2M-1$

FIG. 4.1. *Packing instance in which* $n_0(L,B)=2n_{\mathrm{FFI}_I}(L,B)-1$.

5. Iterated first-fit-decreasing. Despite our intuition that a successful approximation algorithm must pack a sublist L' containing the N' shortest pieces of L, we have observed that the worst-case behavior of both SPF and FFI suffer due to the early and irreversible placement of the smallest pieces of L'.

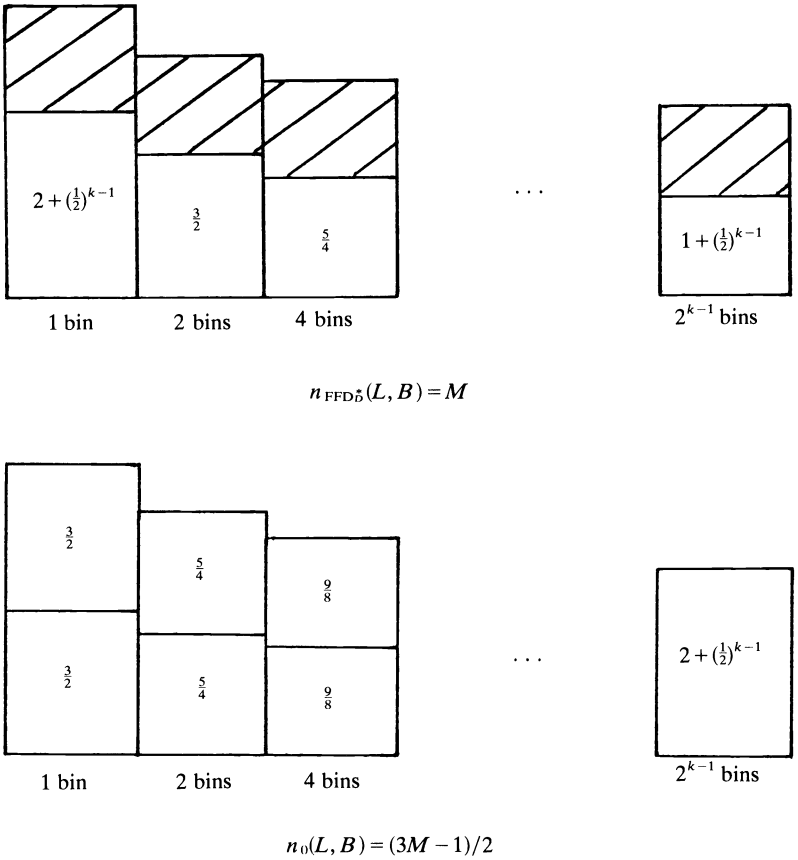
Consider then the iterated first-fit-decreasing (FFD*) algorithm. FFD* also sorts L into an increasing sequence of piece sizes. Next an upper bound on $n_0(L,B)$, say n_{ub}, is computed using bin and piece capacities. That is, we sum the increasing sequence of piece sizes until the inclusion of one more piece would exceed the total capacity of B. An attempt is made to pack a sublist containing the n_{ub} smallest pieces of L, in a *decreasing* sequence of piece sizes, by assigning each piece in turn to the lowest-indexed bin into which it will fit. If a piece is encountered which cannot be packed, the largest piece in the sublist is discarded and a new packing attempted. FFD* terminates when the sublist has been shortened sufficiently to allow all of its remaining pieces to be packed.

The performance of FFD* has been studied in [CL] when bin capacities are equal. For this special case, no more than M iterations are necessary, making its time complexity $O(N \log N + MN \log M)$; FFD* always packs as many pieces as FFI; its asymptotic worst-case bound lies in the range $[\frac{8}{7}, \frac{7}{6}]$.

Allowing the bin sizes to differ, it is easy to see that still no more than M iterations are required (consult the capacity arguments in [CL, Thm. 2]). Hence the time complexity of FFD* is unaffected. But the interested reader should have no difficulty in constructing packing instances for which FFI$_D$ (FFI$_I$) packs more pieces than FFD$_D^*$ (FFD$_I^*$).

In the following example, $n_0(L,B) > \frac{3}{2}n_{\mathrm{FFD}_D^*}(L,B)-1$. Let $M=2^k-1$ for any integer $k>1$. Let $N=3(2^{k-1})-2$. Let $s(B_j)=2+(\frac{1}{2})^i$ for $2^i \leqq j < 2^{i+1}$ and $0 \leqq i < k$. Let $s(p_j)=2+(\frac{1}{2})^{k-1}$ for $1 \leqq j \leqq 2^{k-1}$, let $s(p_j)=1+(\frac{1}{2})^i$ for $2^{k-1}-1+2^i \leqq j < 2^{k-1}-1+2^{i+1}$ and $1 \leqq i < k$. See Fig. 5.1.

Thus we know that $\frac{3}{2}$ is a lower limit on the asymptotic worst-case performance bound of FFD$_D^*$. We have not seriously attempted to establish an upper limit on the FFD$_D^*$ bound since, as we will soon show, the asymptotic bound for FFD$_I^*$ can be no higher (and is, we think, likely to be lower).

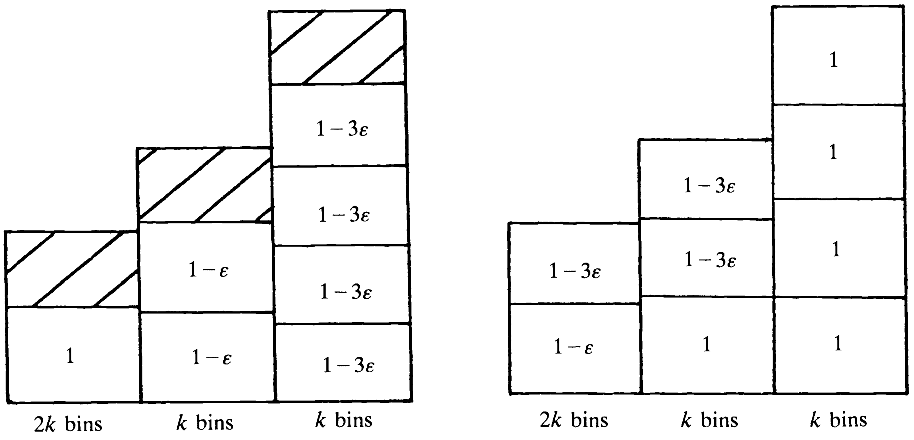


$n_{\mathrm{FFD}_D^*}(L,B)=M$

$n_0(L,B)=(3M-1)/2$

FIG. 5.1. *Packing instance in which* $n_0(L,B) > \frac{3}{2}n_{\mathrm{FFD}_D^*}(L,B)-1$.

The next example and its accompanying figure depict the worst set of problem instances we have been able to contrive for FFD$_I^*$. Let $M=4k$ for any positive integer k. Let $N=11k$. Select an ε in the range $(0,\frac{1}{15})$. Let $s(B_j)=2-4\varepsilon$ for $1 \leqq j \leqq 2k$, $s(B_j)=3-6\varepsilon$ for $2k<j \leqq 3k$ and $s(B_j)=4$ for $3k<j \leqq 4k$. Let $s(p_j)=1$ for $1 \leqq j \leqq 5k$, $s(p_j)=1-\varepsilon$ for $5k<j \leqq 7k$ and $s(p_j)=1-3\varepsilon$ for $7k<j \leqq 11k$. See Fig. 5.2. Thus $n_0(L,B)=\frac{11}{8}n_{\mathrm{FFD}_I^*}(L,B)$. Note that this ratio holds for an entire family of instances, although we can set $k=1$, reduce $s(B_4)$ to 3 and realize a single instance in which $n_0(L,B)=\frac{11}{8}n_{\mathrm{FFD}_I^*}(L,B)+\frac{3}{8}$.



$n_{\mathrm{FFD}_f^*}(L, B) = 8k$ $n_0(L, B) = 11k$

FIG. 5.2. *Packing instance in which* $n_0(L, B) = \frac{11}{8} n_{\mathrm{FFD}_f^*}(L, B)$.

THEOREM 5.1. *For all L and B*, $n_0(L, B) \leqq \frac{3}{2} n_{\mathrm{FFD}_f^*}(L, B) + 1$.

Proof. We assume the existence of a counterexample and derive a contradiction. Suppose there exists a list of pieces $L = \{p_1, p_2, \cdots, p_N\}$ and a collection of bins $B = \{B_1, B_2, \cdots, B_M\}$ such that $n_0(L, B) > \frac{3}{2} n_{\mathrm{FFD}_f^*}(L, B) + 1$. For notational convenience assume $s(p_1) \geqq s(p_2) \geqq \cdots \geqq s(p_N)$ and $s(B_1) \leqq s(B_2) \leqq \cdots \leqq s(B_M)$. Thus at each iteration, the FFD* heuristic attempts to pack the sublist $\{p_j, p_{j+1}, \cdots, p_N\}$ for some $1 \leqq j \leqq N$.

Let f represent the index of the first piece examined by FFD* during its next-to-last iteration. Note that we may assume N is the index of the last piece examined. Therefore during this iteration the sublist $\{p_f, p_{f+1}, \cdots, p_{N-1}\}$ is packed. Letting L' denote $\{p_f, p_{f+1}, \cdots, p_N\}$, it is easy to see that $n_0(L, B) > \frac{3}{2} n_{\mathrm{FFD}_f^*}(L', B) + 1$. From the set of all counterexamples we select one with minimum M.

We now restrict our attention to the comparison of an arbitrary optimal packing of L versus the FFD packing of L'. When no confusion results, the reference to L or L' will henceforth be omitted but implied.

For simplicity, normalize piece and bin sizes so that $s(p_f) = 1$ and assume, as we may, that $s(p_1) = s(p_2) = \cdots = s(p_{f-1}) = 1$.

CLAIM 1. *The* FFD *packing contains no empty bins.*

Proof of Claim 1. The size of such a bin must be less than $s(p_N)$, implying that it is also empty in the optimal packing. Hence, we can delete the bin from B, contradicting the presumed minimality of M. □

CLAIM 2. *Each bin of the optimal packing contains at least two pieces.*

Proof of Claim 2. No bin of the optimal packing can be empty, else we can delete it from B and delete any pieces it had contained in the FFD packing from both L and L', contradicting the minimality of M.

Suppose some bin B_i, $1 \leqq i \leqq M$, contains a single piece p_j in the optimal packing. Consider where FFD packed p_j (or p_f if $j < f$).

If $p_j(p_f)$ is packed by FFD in a bin B_k, $k \geqq i$, then the FFD packing of B_i must contain a piece p_j' as large as p_j. Suppose we delete B_i from B and delete any pieces it had contained in the FFD packing (most notably p_j') from L and L'. If the optimal packing had contained p_j' as well, we move p_j (if it remains) to the position formerly occupied by p'_j; otherwise we merely delete p_j from the optimal packing. Either construction contradicts the minimality of M.

If $p_j(p_f)$ is packed by FFD in a bin B_k, $k<i$, then we delete B_k from B, delete the piece(s) it had contained in the FFD packing from L and L', and in the optimal packing move the remaining former contents of B_k to the now empty B_i (since $s(B_k) \leqq s(B_i)$), contradicting the minimality of M.

Finally, if the FFD packing does not contain $p_j(p_f)$, it must be that $j=N$ and it is easy to see that the minimality of M is again contradicted. □

CLAIM 3. $s(p_N) > \frac{2}{3}$.

Proof of Claim 3. From Claim 2 we know that $n_0(L,B) \geqq 2M$. This and the fact that $n_0(L,B) > \frac{3}{2} n_{\mathrm{FFD}^*_I}(L',B)+1$ implies $n_0(L,B) - n_{\mathrm{FFD}}(L',B) > 2M/3$. Thus the optimal packing must find room for an additional $2M/3$ pieces, each of size 1. The FFD packing therefore includes at least one bin B_i, $1 \leqq i \leqq M$, such that $s(B_i) - s$ (the contents of $B_i) \geqq \frac{2}{3}$ and hence $s(p_N) > \frac{2}{3}$. □

As an immediate consequence of the last two claims, we observe that $s(B_1) > \frac{4}{3}$.

Let us scrutinize the FFD packing. We will use the term k-bin to designate a bin containing exactly k pieces. Let $B1$ represent the collection of one-bins and let I denote the number of bins in $B1$.

CLAIM 4. $I > 0$.

Proof of Claim 4. If $I=0$, then $n_{\mathrm{FFD}}(L',B) \geqq 2M$ and simple capacity arguments imply $n_0(L,B) < 3M$. □

Clearly B_I is the last one-bin (i.e., all one-bins are "left-justified"). Moreover, each B_i, $1 \leqq i \leqq I$, contains p_{f+i-1}. We denote the single piece in B_I, namely p_{f+I-1}, by z.

CLAIM 5. *The optimal packing of $B1$ requires exactly $2I$ pieces, each of size less than $s(z)$.*

Proof of Claim 5. Immediate from Claim 2 and the fact that $3s(p_N) > s(z)+s(p_N) > s(B_I) \geqq s(B_{I-1}) \geqq \cdots \geqq s(B_I)$. □

Let J represent the number of pieces in $L' - \{p_N\}$ not contained in $B1$ in either the FFD or the optimal packing. Let K denote the difference between the number of pieces the optimal packing places in $B-B1$ and the number placed there by the FFD rule.

CLAIM 6. $K \leqq J/2$.

Proof of Claim 6. Suppose B_i, $I<i \leqq M$, is a k-bin, $k \geqq 2$, of $B-B1$. Suppose further that B_i contains $k+r$, $r \geqq 1$, pieces in the optimal packing and that less than $2r$ of them are in the FFD packing of $B-B1$. From the optimal packing of B_i we know $(k-r+1)s(z)+(2r-1)s(p_N) \leqq s(B_i)$. From the FFD packing of B_i we know $ks(z)+s(p_N) > s(B_i)$. Combining these and using the fact that $2s(p_N)-s(z) > 0$, we conclude that $r<1$, a contradiction. □

We now complete the proof of Theorem 5.1 by observing that $4I+J+K \geqq n_0(L,B) > \frac{3}{2} n_{\mathrm{FFD}^*_I}(L',B)+1 \geqq \frac{3}{2}(3I+J-1)+1$ and by Claim 6 we derive $I \leqq 0$, contradicting Claim 4. □

6. Concluding remarks. The superior worst-case behavior of FFD$^*_I$ makes it worth its added complexity. We conjecture that its true performance bound is of the form $n_0(L,B) \leqq \frac{11}{8} n_{\mathrm{FFD}^*_I}(L,B)+C$ for all L and B and for some positive constant C. The proof of Theorem 5.1 was straightforward and should, we hope, leave the reader with an understanding of how we arrived at the result. This may unfortunately not be the case for attempts at proving tighter bounds, particularly if one must resort to the adoption of a *weighting argument* approach (see, for example, [CGJ], [CL], [DFL] or [Jo]). In any event, such a proof cannot rely on Claim 3 and hence must deal with several bothersome issues including "regular" two-bins which may contain four pieces in an optimal packing and "fallback" two-bins which may contain three.

Additionally, we remark that for each class of algorithm considered it was advantageous to first sort B into an increasing sequence of bin sizes. The author has participated in one other work involving bins of different sizes, [FL], and in that effort it was similarly profitable to use increasing bin sizes. It would be interesting to learn of some general result addressing when, if ever, it is more appropriate to arrange bins by decreasing size.

REFERENCES

[CGJ] E. G. Coffman Jr., M. R. Garey and D. S. Johnson, *An application of bin packing to multiprocessor scheduling*, SIAM J. Comput., 7 (1978), pp. 1–17.

[CL] E. G. Coffman Jr. and J. Y-T. Leung, *Combinatorial analysis of an efficient algorithm for processor and storage allocation*, SIAM J. Comput., 8 (1979), pp. 202–217.

[CLT] E. G. Coffman Jr., J. Y-T. Leung and D. W. Ting, *Bin packing: maximizing the number of pieces packed*, Acta Informatica, 9 (1978), pp. 263–271.

[DFL] B. L. Deuermeyer, D. K. Friesen and M. A. Langston, *Scheduling to maximize the minimum processor finish time in a multiprocessor system*, this Journal, 3 (1982), pp. 190–196.

[FL] D. K. Friesen and M. A. Langston, *Bounds for* MULTIFIT *scheduling on uniform processors*, SIAM J. Comput., 12 (1983), pp. 60–70.

[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.

[Jo] D. S. Johnson, *Near optimal bin packing algorithms*, Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, 1973.

IT'S HARD TO COLOR ANTIRECTANGLES*

ANDY BOUCHER†

Abstract. The maximum number of elements in an antirectangle of a convex board equals the minimum number of rectangles it takes to cover that board. It is shown here that the dual of this theorem, that the minimum number of antirectangles needed to cover a convex board equals the maximum size of any rectangle of the board, is not generally true.

A *board* is a finite subset of the set of all unit squares whose corners are integer lattice points in R^2. A board B is said to be (horizontally and vertically) *convex* if, whenever two unit squares in B are on the same horizontal or vertical line, all unit squares on that line between the two are also in B. A *rectangle* of B is a subset of B whose members form a rectangle. An *antirectangle* of B is a subset of B, no two of whose elements are in the same rectangle.

Chaiken, Kleitman, Saks, and Shearer [1] have shown that, for convex boards, the maximum number of elements in an antirectangle of B equals the minimum number of rectangles it takes to cover B (i.e. $\alpha = \theta$). The dual of this theorem is that the minimum number of antirectangles needed to cover B equals the maximum size of any rectangle of B (i.e. that $\chi = \omega$). It is shown here that this dual need not hold for all convex boards. Specifically, it does not hold for the board in Fig. 1. Notice that this board is convex and $\omega = 144$. Now, a covering by antirectangles of any board is equivalent to a coloring of that board, no two squares in the same rectangle receiving the same color. So, suppose the board could be colored with 144 colors. Then we may arbitrarily assign the colored squares of $ABCD$ the colors shown, necessarily no two being the same. Now, what color may we give square E? Since $WXYZ$ has 126

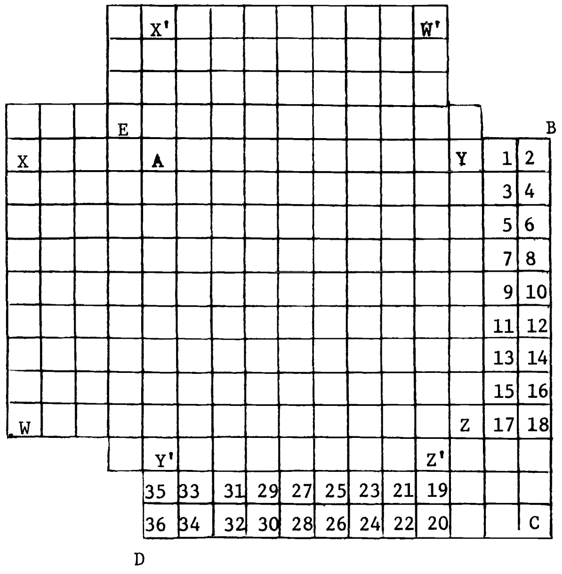


FIG. 1

* Received by the editors July 22, 1980, and in final revised form July 7, 1983. This research was done at Smith College, supported in part by the National Science Foundation under grant SPI 79 26984.

† Balliol College, Oxford, England.

squares, and since it cannot use any of the colors 1–18, $WXYZ$ must use all the colors 19–144. But then E, in a rectangle with all the squares of $WXYZ$, must not be colored using any of these, i.e. it must use a color among 1–18. Symmetrically, considering $W'X'Y'Z'$, E must use a color 19–36. This is impossible, so B is uncolorable, that is, $\chi > \omega$.

Acknowledgment. The author wishes to thank M. O. Albertson for his assistance given throughout the work on this problem.

BIBLIOGRAPHY

[1] S. Chaiken, D. J. Kleitman, M. Saks and J. Shearer, *Covering regions with rectangles*, this Journal, 2 (1981), pp. 394–410.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

ITERATIVE METHODS FOR COMPUTING STATIONARY DISTRIBUTIONS OF NEARLY COMPLETELY DECOMPOSABLE MARKOV CHAINS*

J. R. KOURY,† D. F. McALLISTER‡ AND W. J. STEWART‡

Abstract. We propose new methods which combine aggregation with point and block iterative techniques for computing the stationary probability vector of a finite ergodic Markov chain. These techniques are also compared numerically with several methods which have recently appeared in the literature for the class of nearly completely decomposable Markov chains.

Key words. point and block iterative methods, aggregation, nearly completely decomposable Markov chains, stationary probability factor

1. Introduction.

1.1. Background.

The application of finite Markov chains throughout the biological, physical and social sciences, as well as in business and engineering, is well documented. For example, queueing networks have been used extensively in modeling computer systems. They provide the basis upon which new designs may be evaluated and a means to estimate the change in performance due to an increase in capacity of one or several of the system components. The effect of changes in both workload characteristics and system software may be determined from appropriate queueing network models. However, since the class of queueing networks amenable to solution by analytical techniques is very restrictive, it often becomes necessary to numerically analyze the underlying Markov chain.

This is the approach adopted by Kaufman [Kauf81b], Kaufman, Gopinath and Wunderlich [Kauf81] and Kaufman, Seery and Morrison [Kauf81a] in the study of queueing network problems that commonly arise in communications theory. Such problems include networks with overflowing queues where the associated transition probability matrix has a symmetric zero-structure and order approaching 30,000.

Another important application area related to finite Markov chains is the solution of homogeneous systems of linear equations arising from compartmental models of biological and physical systems. Such models have been studied extensively, for example, at the Oak Ridge National Laboratory (see Sheppard and Householder [Shep51], Funderlic and Heath [Fund71], or Funderlic and Mankin [Fund81]). Applications include studies of the increase of atmospheric carbon dioxide as a result of increased fossil fuel combustion and the study of the dynamics of carbon through closed ecological systems.

A third area of considerable current importance is that of reliability analysis. Unlike the previously described application areas, it becomes necessary to study the transient behavior of the Markov chain rather than its stationary behavior. The methods that are described herein may have potential for application in the study of the transient behavior.

The Markov chains that arise in these applications are typically finite, homogeneous and ergodic. Furthermore, they usually involve a very large number of states and a

* Received by the editors December 29, 1982, and in revised form March 24, 1983. This research was supported in part by the National Science Foundation under grant MCS80-04345. A version of this paper was presented at the Sparse Matrix Symposium, 1982, Fairfield Glade, Tennessee, October 24–27, 1982.

† IBM, Raleigh, North Carolina.

‡ Department of Computer Science, North Carolina State University, Raleigh, North Carolina 27650.

relatively small number of nonzero transition probabilities, thus making the transition probability matrix $\mathbf{Q}$ large and sparse. The fundamental problem is to compute and analyze the unique stationary distribution vector $\mathbf{x}=(x_1, x_2, \cdots, x_n)^T$ which has the properties:

$$\mathbf{x}^T\mathbf{Q}=\mathbf{x}^T, \quad x_i>0\ \forall i, \quad \sum_{i=1}^{n} x_i=1=\|\mathbf{x}\|_1. \tag{1.1}$$

In the case where $\mathbf{Q}$ is large and exhibits a nearly completely decomposable (NCD) [Cour77] structure, a method of "aggregation" can be incorporated into some steps of a standard iteration procedure (such as the power method or the method of Gauss–Seidel). The main thrust of this paper is to show that aggregation combined with point and block iterative techniques can produce methods which converge rapidly to the stationary probability vector $\mathbf{x}$. We also compare these new methods with some methods which have recently appeared in the literature. For a discussion of other methods relevant to this problem see [Harr83] and [Fund83].

We first review some numerical techniques currently in use for computing stationary distributions of Markov chains.

1.2. Current numerical methods. Let M be a finite n state aperiodic Markov chain with transition probability matrix $\mathbf{Q}$ of order n. We will assume that $\mathbf{Q}$ is irreducible and hence by the Perron–Frobenius theory of positive matrices the eigenvalue 1 is a simple dominant eigenvalue of $\mathbf{Q}$. In this case, M has a stationary probability distribution $\mathbf{x}$ which satisfies (1.1).

There are many methods in the literature which compute $\mathbf{x}$ or an approximation to $\mathbf{x}$.

Iterative methods. The most well known and perhaps oldest methods are iterative and are variants of the power method:

Let $\mathbf{x}^{(0)}$ be a positive starting vector with norm 1. Then the power method is defined by the sequence

$$\mathbf{x}^{(i+1)}=\mathbf{Q}^T\mathbf{x}^{(i)}/\|\mathbf{Q}^T\mathbf{x}^{(i)}\|_1, \qquad i=0,1,2,\cdots.$$

It may be shown that if the eigenvalues of $\mathbf{Q}$ are arranged in decreasing order

$$\lambda_1=1>|\lambda_2|\geqq|\lambda_3|\geqq\cdots\geqq|\lambda_n| \tag{1.2}$$

then, asymptotically, we have

$$\|\mathbf{x}^{(i+1)}-\mathbf{x}\|_1\sim|\lambda_2|\ \|\mathbf{x}^{(i)}-\mathbf{x}\|_1, \tag{1.3}$$

which implies that the sequence converges linearly to $\mathbf{x}$ with convergence factor $|\lambda_2|$, the magnitude of the subdominant eigenvalue of $\mathbf{Q}$. This is the technique used in the Recursive Queue Analyzer of Wallace and Rosenberg [Wall66]. As is obvious from relationship (1.3), if $|\lambda_2|$ is close to 1, convergence may be excruciatingly slow. Unfortunately, this phenomenon is almost assured in the case that n is large since all of the eigenvalues of $\mathbf{Q}$ lie in the unit circle of the complex plane. This is also the case if $\mathbf{Q}$ is nearly completely decomposable (NCD).

Let M be a perturbation of a completely decomposable Markov chain M^* and let $\mathbf{Q}$ and $\mathbf{Q}^*$ be respectively the corresponding transition probability matrices. Then M is said to be NCD if

$$\mathbf{Q}=\mathbf{Q}^*+\varepsilon\mathbf{C}, \tag{1.4}$$

where $0<\varepsilon\ll 1$, $\mathbf{Q}^*$ is a block diagonal stochastic matrix and the row sums of $\mathbf{C}$ are zero. If $\mathbf{Q}^*$ has N blocks, then 1 is an eigenvalue of $\mathbf{Q}^*$ of multiplicity at least N. Since the eigenvalues of a matrix are continuous functions of the elements, it must be the case that $\mathbf{Q}$ has at least $N-1$ eigenvalues which are close to 1. In particular $|\lambda_2|$ must be close to 1.

A variant of the power method, called simultaneous iteration [Stew81] is used for finding stationary distributions in the Queueing Network Analysis Package (QNAP), [Merl78]. Rather than iterate with a single vector as in the power method, iteration is carried out with $m>1$ trial vectors and it can be shown that the rate of convergence is asymptotically proportional to $|\lambda_{m+1}|$.

Single step approximation methods.

i. *Courtois aggregation.* The decomposition and aggregation technique of Simon and Ando [Simo61] may be used to obtain an approximate solution when $\mathbf{Q}$ is NCD.

Let $\mathbf{Q}$ and $\mathbf{x}$ be partitioned according to the blocks of $\mathbf{Q}^*$, i.e. if

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{Q}_{11}^* & & & 0 \\ & \mathbf{Q}_{22}^* & & \\ & & \ddots & \\ 0 & & & \mathbf{Q}_{NN}^* \end{bmatrix}$$

then

$$(1.5) \qquad \mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} & \cdots & \mathbf{Q}_{1N} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \cdots & \mathbf{Q}_{2N} \\ \vdots & \vdots & & \vdots \\ \mathbf{Q}_{N1} & \mathbf{Q}_{N2} & \cdots & \mathbf{Q}_{NN} \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}.$$

The scalar ε is defined to be

$$\varepsilon = \left\| \mathbf{Q} - \begin{bmatrix} \mathbf{Q}_{11} & & & 0 \\ & \mathbf{Q}_{22} & & \\ & & \ddots & \\ 0 & & & \mathbf{Q}_{NN} \end{bmatrix} \right\|_\infty .$$

Courtois [Cour77] refers to ε as the maximum degree of coupling of the subsystems represented by the diagonal blocks of $\mathbf{Q}$. He suggests approximating $\mathbf{Q}^*$ by distributing the probability mass in a given row and outside the diagonal block into the elements of the row within the diagonal block. Unfortunately it is not known how this should best be done. (It is known that there is an uncountable set of block diagonal stochastic matrices $\mathbf{Q}^*$ with the given block structure which has the same stationary probability vector $\mathbf{x}$ as $\mathbf{Q}$ [Dodd81a] but the vector $\mathbf{x}$ must be known in order to find such a $\mathbf{Q}^*$.)

After forming a matrix $\mathbf{Q}^*$ from $\mathbf{Q}$ we compute the stationary probability vector $\mathbf{x}_I^*$ of each block $\mathbf{Q}_{II}^*$:

$$(1.6) \qquad \mathbf{x}_I^{*T}\mathbf{Q}_{II}^* = \mathbf{x}_I^{*T}, \qquad \|\mathbf{x}_I^*\|_1 = 1, \qquad 1 \leqq I \leqq N.$$

We now perform an aggregation step which may be described as follows.

Let $\mathbf{1}$ denote a vector all of whose elements are 1. Define the $N\times N$ matrix $\mathbf{P}$ as the matrix whose $(I-J)$th element is given by:

$$(1.7) \qquad P_{IJ} = \mathbf{x}_I^{*T}\mathbf{Q}_{IJ}\mathbf{1}.$$

Then $\mathbf{P}$ is a stochastic matrix. Furthermore it is irreducible and hence has a stationary vector $\mathbf{X}$, such that

$$\mathbf{X}^T\mathbf{P}=\mathbf{X}^T \quad \text{and} \quad \|\mathbf{X}\|_1=1.$$

An approximation $\hat{\mathbf{x}}$ to $\mathbf{x}$ may now be computed from using $\hat{\mathbf{x}}_I=X_I\mathbf{x}_I^*$, $1\leqq I\leqq N$. Courtois shows that $\hat{\mathbf{x}}$ is an $O(\varepsilon)$ approximation to $\mathbf{x}$.

ii. *Inverse iteration.* The eigenvector problem can be reformulated as a linear system problem by noting that if $\mathbf{x}^T\mathbf{Q}=\mathbf{x}^T$ then

$$\mathbf{S}\mathbf{x}=(\mathbf{I}-\mathbf{Q}^T)\mathbf{x}=0, \qquad \|\mathbf{x}\|_1=1. \tag{1.8}$$

Hence $\mathbf{x}$ lies in the null space of the transition rate matrix

$$\mathbf{S}=\mathbf{I}-\mathbf{Q}^T.$$

Since 0 is a simple eigenvalue of $\mathbf{S}$, rank $(\mathbf{S})=n-1$ and inverse iteration can be applied to solve system (1.8) as follows:

Let $\mathbf{x}^{(0)}$ be any initial approximation to $\mathbf{x}$. For $k=1,2,\cdots$

$$\text{solve } \mathbf{S}\mathbf{y}^{(k)}=\mathbf{x}^{(k)} \text{ for } \mathbf{y}^{(k)} \text{ and set } \mathbf{x}^{(k+1)}=\mathbf{y}^{(k)}/\|\mathbf{y}^{(k)}\|_1.$$

In this case the sequence converges in a single iteration even if $\mathbf{x}^{(0)}$ is a random vector, since the corresponding eigenvalue is known exactly. If $\mathbf{A}$ is large, general linear equation solvers are not appropriate for solving (1.8) and it is necessary to exploit the structure of the problem. It is often the case that the matrix $\mathbf{A}$ can be partitioned and permuted so that it has a (block) banded structure [Mull80], [Stew78] and one can employ a sparse or banded matrix solver with a direct matrix factorization routine as a subprocedure. Although convergence is immediate, storage requirements are high and the calculation is expensive. We will not include this technique in our numerical tests.

2. New methods.

2.1. Single step methods.

i. *G. W. Stewart's method* (GWSD). An approach somewhat similar to that of Courtois was proposed by G. W. Stewart [Stew80]. His method has the advantage that it does not require the distribution of the probability mass which is outside the diagonal blocks to within the diagonal blocks. In addition he provides an error analysis of the procedure which results in effectively computable error bounds.

Let $\mathbf{Q}$ be partitioned as given in (1.5) and assume that each of the blocks $\mathbf{Q}_{II}$, $1\leqq I\leqq N$, is irreducible. Then, again by the Perron–Frobenius theory each has a simple dominant eigenvalue or Perron root λ_I, $1\leqq I\leqq N$, and corresponding left and right eigenvectors $\mathbf{v}_I$, $\mathbf{u}_I$ respectively. Assume $\mathbf{v}_I$ and $\mathbf{u}_I$ have been normalized so that $\mathbf{v}_I^T\mathbf{u}_I=1$. Define the $N\times N$ matrix $\mathbf{B}$ as the matrix whose $(I-J)$th element is given by:

$$B_{IJ}=\mathbf{v}_I^T\mathbf{Q}_{IJ}\mathbf{u}_J, \qquad 1\leqq I, \quad J\leqq N.$$

Then $\mathbf{B}$ is irreducible and has a left eigenvector $\mathbf{z}$ corresponding to its unique dominant eigenvalue. Stewart shows that the vector $\bar{\mathbf{x}}$ defined by

$$\bar{\mathbf{x}}_I=z_I\mathbf{v}_I, \qquad 1\leqq I\leqq N,$$

is an $O(\varepsilon)$ approximation to $\mathbf{x}$.

ii. *Vantilborgh's $O(e^2)$ approximation* (VANTD). Using Courtois' method, Vantilborgh [Vant81a,b] has developed an algorithm to produce an $O(e^2)$ approximation to $\mathbf{x}$.

Let $\hat{\mathbf{x}}$ be an approximation to the left Perron eigenvector, $\mathbf{x}$, of $\mathbf{Q}$ and define $B_I = \|\mathbf{x}_I\|_1$. Courtois [Cour77] shows that a necessary and sufficient condition for $\mathbf{x}=\hat{\mathbf{x}}$ is

$$\mathbf{x}_I^* = B_I^{-1}\hat{\mathbf{x}}_I \quad \text{for } I = 1, 2, \cdots, N. \tag{2.1}$$

This motivates the search for an ideal matrix $\mathbf{Q}^*_{[I,\text{ideal}]}$. The $\mathbf{Q}^*_{[I,\text{ideal}]}$ is any row stochastic matrix formed from $\mathbf{Q}_I$ by adding the off-diagonal block row mass of $\mathbf{Q}$ into the diagonal submatrix in such a way that (2.1) holds.

Courtois [Cour75] and Courtois and Louchard [Cour76] describe procedures for obtaining an $O(e^2)$ approximation to the steady state probability vector. One of these involves modifying the results obtained from an $O(e)$ approximation without regard to the original amalgamation of the diagonal blocks. Vantilborgh approaches the problem by finding a better way to amalgamate the off diagonal blocks into the diagonal blocks. He shows that all three methods are mathematically equivalent.

Let the vectors be partitioned as follows:

$$\begin{array}{cccc} \mathbf{a} = [& \hat{\mathbf{a}} & \vdots & \mathbf{a}_N \;]. \\ & n-n(N) & & n(N) \\ & \text{elements} & & \text{elements} \end{array}$$

Consider the matrix

$$\begin{array}{cc} \left[\begin{array}{c|c} \hat{\mathbf{Q}} & \hat{\mathbf{Q}}_C \\ \hline \hat{\mathbf{Q}}_L & \mathbf{Q}_{NN} \end{array}\right] & \begin{array}{l} n-n(N) \text{ rows} \\ \\ n(N) \text{ rows} \end{array} \\ \begin{array}{cc} n-n(N) & n(N) \\ \text{columns} & \text{columns} \end{array} & \end{array}$$

The following results will be given in terms of the Nth diagonal block. The results can be easily extended to treat the Ith aggregate, $1 \leqq I \leqq N-1$.

From [Vant81] we have

$$\mathbf{x}_N^T(\mathbf{Q}_{NN} + \hat{\mathbf{Q}}_L(\mathbf{I}-\hat{\mathbf{Q}})^{-1}\hat{\mathbf{Q}}_C) = \mathbf{x}_N^T. \tag{2.2}$$

Equation (2.2) defines the matrix $\mathbf{Q}^*_{[N,\text{ideal}]}$ which will be approximated.

Using matrix forms of Taylor and Laurent series, Vantilborgh determines the following approximations to $\mathbf{Q}^*_{[N,\text{ideal}]}$

$$\hat{\mathbf{X}}^{*T} = \frac{1}{1-X_N}[X_1, \cdots, X_{N-1}]^T.$$

If

$$\mathbf{Q}_N^* = \mathbf{Q}_{NN} + \frac{1}{\hat{\mathbf{X}}^{*T}\mathbf{p}\mathbf{1}}\mathbf{q}\mathbf{1}\hat{\mathbf{X}}^{*T}\mathbf{p}, \quad \text{then } \mathbf{x}_N^* = B_N^{-1}\mathbf{x}_N + O(e^2), \tag{2.3}$$

where

$$\mathbf{p} = \begin{bmatrix} \mathbf{x}_1^{*T}\mathbf{Q}_{1N} \\ \vdots \\ \mathbf{x}_{N-1}^{*T}\mathbf{Q}_{N-1,N} \end{bmatrix} \quad \text{and} \quad \mathbf{q} = [\mathbf{Q}_{N1}\mathbf{1}, \cdots, \mathbf{Q}_{N,N-1}\mathbf{1}].$$

Having formed this approximation to $\mathbf{Q}^*_{[N,\text{ideal}]}$, its left Perron eigenvector must be determined. Using the resulting subvector with Courtois's normalization procedure produces an $O(e^2)$ approximation of $B_N^{-1}\mathbf{x}_N$.

2.2. Iterative methods. The preceding methods for approximating the stationary probability vectors of stochastic matrices have been direct "one step" methods. In many cases this approximation will be sufficient for the application at hand. For more accurate solutions, iterative algorithms can be employed.

Classical methods of computing the stationary probability vector require time complexity $O(n^3)$. When modeling queueing systems, the state space quickly becomes too large to make such algorithms tractable. However, the special structure of nearly completely decomposable systems allows the application of iterative algorithms with time complexities considerably less than $O(n^3)$ for n large. In this paper we will investigate combining the aggregation procedure embedded in Courtois's direct method with existing iterative algorithms such as the power method [Dodd81] and block iterative techniques which have the effect of greatly increasing the convergence rate for most nearly completely decomposable systems. We will compare these with Vantilborgh's extension of his $O(e^2)$ direct approximation to an iterative process which produces an error of $O(e^k)$ on the kth iteration.

We first describe the application of Courtois aggregation as a subprocedure in an iterative algorithm (cf. [Dodd81]).

Let $\hat{\mathbf{x}}$ be the approximation to $\mathbf{x}$ produced by the current iteration of the iterative technique being employed. We form the $N \times N$ stochastic matrix $\mathbf{P}$:

$$\mathbf{P}_{IJ} = (\hat{\mathbf{x}}_I^T \mathbf{Q}_{IJ} \mathbf{1}) / \|\hat{\mathbf{x}}_I\|_1, \qquad 1 \leqq I, J \leqq N,$$

where $\mathbf{1}$ is a vector of 1's of conformable length. Let $\mathbf{Y}$ be the stationary probability vector of $\mathbf{P}$:

$$\mathbf{Y}^T \mathbf{P} = \mathbf{Y}^T.$$

We then modify $\hat{\mathbf{x}}$ using Courtois aggregation as follows:

$$\hat{\mathbf{x}}_I \leftarrow \frac{\hat{\mathbf{x}}_I}{\|\hat{\mathbf{x}}_I\|_1} \mathbf{Y}_I, \qquad 1 \leqq I \leqq N.$$

We note that this method requires $\hat{\mathbf{x}}$ to be on the unit hypercube.

In combining aggregation with the above classical methods we have the following generic algorithm:

1. Initialize $\mathbf{x}^{(0)}$ using Courtois's direct method.
2. Perform some number of either point Gauss–Seidel, block Gauss–Seidel or power method iterations.
3. Perform a Courtois aggregation.
4. Perform some number of iterations for the method being used in step 2.
5. If convergence has been reached, stop. Else go to step 3 with the new approximation.

In [Mull80], Muller has described an iterative procedure for solving (1.8) with $\mathbf{S}$ replaced by a perturbed matrix $\hat{\mathbf{S}}$. The n, n element of $\hat{\mathbf{S}}$ becomes

$$\hat{\mathbf{S}}_{nn} \leftarrow |\mathbf{S}_{nn}| + k,$$

where k is a small, real, positive scalar. We note that Courtois aggregation requires that $\mathbf{x}$ be a probability vector. Muller has modified Courtois aggregation to treat the solution of nonhomogeneous linear systems [Koury83]. In [Koury83] it is shown that this technique is inferior to solving the homogeneous system and we will not consider it further.

We now define the classical iteration schemes which we use with Courtois aggregation. For the initial vector in each case we set $\mathbf{x}^{(0)}$ to the vector produced by Courtois's direct method. The vector $\mathbf{x}^{(k)}$ is the vector obtained on the kth iterate.

i. *Block Gauss–Seidel* (BGS). For the homogeneous system (1.8), block or group Gauss–Seidel becomes [Berm79]:

$$\mathbf{S}_{II}^T\mathbf{x}_I^{(k)} = -\sum_{J=1}^{I-1} \mathbf{S}_{IJ}^T\mathbf{x}_J^{(k)} - \sum_{J=I+1}^{N} \mathbf{S}_{IJ}^T\mathbf{x}_J^{(k-1)}, \qquad I = 1, \cdots, N,$$

where **S** has been partitioned like **Q**. For each iteration, after computing the right-hand side, we use an *LU* decomposition of the diagonal blocks to solve for $\mathbf{x}_I^{(k)}$. Note that this implies that the diagonal blocks of **S** must be nonsingular.

ii. *Point Gauss–Seidel* (GS). A special case of block Gauss–Seidel is point Gauss–Seidel. If the blocks of S are 1×1, we have

$$x_j^{(k)} = -1/s_{jj}\left(\sum_{j=1}^{i-1} s_{ij}x_j^{(k)} + \sum_{j=i+1}^{n} s_{ij}x_j^{(k-1)}\right), \qquad j = 1, \cdots, n.$$

Note here that the diagonal elements of **S** must be nonzero.

iii. *Power method* (DMS). In the power method, $\mathbf{x}^{(k)}$ is obtained by multiplying the matrix **Q** on the left by the iteration vector:

$$\mathbf{x}^{T(k)} = \mathbf{x}^{T(k-1)}\mathbf{Q}.$$

Its convergence properties were discussed in § 2.

The three preceding classical algorithms have been studied previously. The analysis of their convergence properties can be found in several sources (e.g., [Varga62], [Young71], [Berm79]). Point Gauss–Seidel and block Gauss–Seidel are known to converge for diagonally dominant M-matrices [Varga62, p. 91] and the structure of NCD queueing networks provides a natural partitioning of the state space which we use in block Gauss–Seidel.

iv. *Vantilborgh's iterative method* (VANTI). In § 2 we described Vantilborgh's direct method which produces an $O(e^2)$ approximation to **x**. Vantilborgh shows that if $\mathbf{x}^*$ is within $O(e^k)$ of **x**, then using this approximation to compute the matrices $\mathbf{Q}_I^*$, $1 \leqq I \leqq N$, produces an $O(e^{k+1})$ approximation to **x**. Therefore, Vantilborgh's $O(e^2)$ technique can be extended naturally to an iterative method. The iteration scheme is as follows:

1. Initialize k to 1.
2. Perform CD (Courtois's direct method) to compute $\mathbf{x}^{(1)}$.
3. $k \leftarrow k+1$. Use $\mathbf{x}_I^{*(k)}$, $1 \leqq I \leqq N$, to form matrices $\mathbf{Q}_I^*$ (equation (2.3)).
4. Determine Perron vectors of each $\mathbf{Q}_I^*$ and concatenate subvectors as in Vantilborgh's direct method.
5. Normalize $\mathbf{x}^{(k)}$ using Courtois's procedure to obtain an $O(e^k)$ approximation of the stationary probability vector.
6. If the desired convergence has been reached, stop. Else go to step 3.

3. Test matrices. Matrices of four different types were used to test the various algorithms (cf. [Koury83]). The first is an 8×8 matrix taken from [Cour77, p. 85] with .00045 subtracted from $q_{6,2}$ to make the matrix stochastic. A set of five randomly generated matrices comprise the second test group. The third group of test matrices was taken from a queueing network model of a multiprogramming system analyzed in [Vant81]. Kaufman, in [Kauf81], describes a model of two nodes of a packet switching network which we used to produce our last group of test matrices.

i. *Courtois's matrix* (COURTOIS). An 8×8 matrix [Cour77, p. 85] has appeared in several places in literature relevant to NCD systems (e.g., [Stew80], [Vant81], [Dodd81]). For this matrix, $e = 10^{-3}$. The subdominant eigenvalue, λ_2 is close to one which implies that the power method will converge slowly.

ii. *Randomly generated matrices* (N). These five matrices are of order 16. They were produced by first generating numbers from a uniform distribution on [0, 1]. These numbers were then scaled according to their corresponding blocks by multiplying them by the scaling factors shown in the diagram below with α successively set to 1, .1, .01, .001, and .0001 producing the matrices $N1$ through $N5$:

$$\begin{bmatrix} 1 & \alpha & \alpha/2 & \alpha/5 \\ \alpha & 1 & \alpha & \alpha/2 \\ \alpha/2 & \alpha & 1 & \alpha \\ \alpha/5 & \alpha/2 & \alpha & 1 \end{bmatrix}.$$

Each row was then multiplied by the appropriate scalar to produce a row stochastic matrix.

The maximum degree of coupling, e, varied from .816 for $\alpha=1$ to .00044 for $\alpha=.0001$. The eigenvalues tended to show a large separation between the Nth and $(N+1)$st eigenvalue and the matrices were more dense than those produced by queueing network models.

iii. *Multiprogramming model matrices* (V). These matrices were generated from the multiprogramming model of [Vant81]. Similar models have been studied in [Konh76], [Konh78], and [Gele78]. A diagram of the model is shown in Fig. 1.

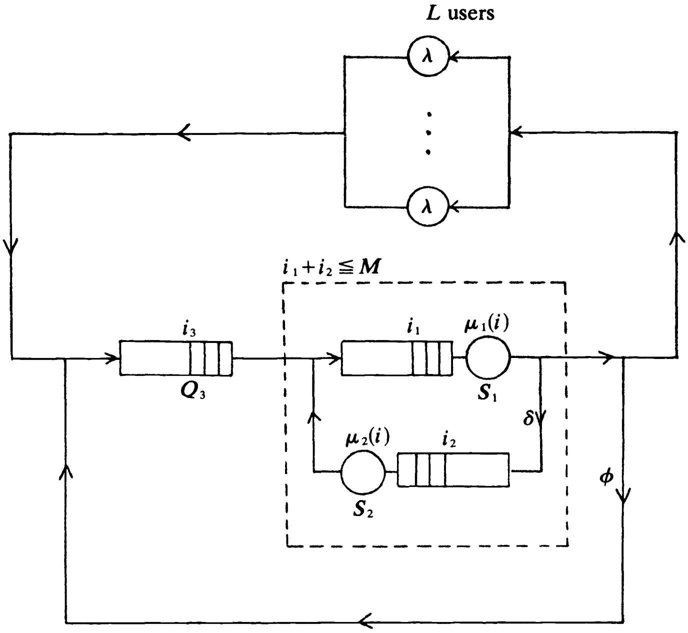


FIG. 1

In the model there are U users which produce jobs according to a state independent Poisson process, each user having rate λ. In the inner model, there is a CPU represented by S_1 with state dependent exponential mean service rate $\mu_1(i_1)$ where i_1 is the number of customers in queue S_1. The analogous parameters for the secondary memory are S_2, i_2 and $\mu(i_2)$.

When a job is produced by a user, it immediately enters the inner model if $i_1 + i_2 < M$, otherwise it enters queue Q_3. When a job completes service at the CPU it goes to secondary memory with probability δ and to Q_3 or S_1 with probability ϕ. When $\phi + \delta \cong 1$, which is usually the case for computer systems [Cour77], the matrices produced by the model are NCD. The parameter λ also effects the decomposability of the system. When arrivals to the inner model are slow relative to the service rates of the CPU and secondary memory, many state transitions occur within the inner model before the value $i_1 + i_2$ changes. Hence, each state with constant values of $i_1 + i_2$ represents a subsystem which can be analyzed separately.

Two sets of service rates were used for the CPU and secondary memory [Vant81]. They are

$$(3.1) \qquad \mu_1(i) = \frac{64}{i+16}, \qquad \mu_2(i) = \frac{3i}{i+6},$$

and

$$(3.2) \qquad \mu_1(i) = \frac{112}{i+16}, \qquad \mu_2(i) = \frac{24i}{i+6}.$$

By varying these parameters, matrices with different maximum degrees of coupling and dimension could be produced. The matrices derived from this model are block tridiagonal and the blocks themselves are tridiagonal. There will be $m+1$ diagonal blocks each with size $1 + i_1 + i_2$.

iv. *Packet switching network model matrices* (K). In [Kauf81] Kaufman describes several models of a two-node packet switching network. We used the six-dimensional model with buffer reservation and processor sharing. The model consists of two symmetric nodes.

Local and foreign packets arrive according to a Poisson distribution with respective parameters λ_l and λ_f. The trunk transmission rate is exponential with mean $\underline{m}_t$. Packets received from the trunk are serviced at exponential rate $\underline{m}_t$. New packets are processed at exponential rate $\underline{m}_s$, or $\underline{m}_s - \underline{m}_t$ if a packet from the trunk is being serviced. When a packet is determined to be foreign, it must be stored in the output section of the buffer until an acknowledgment is received from the other node verifying its reception. Packets wait in the input section of the buffer until being processed. The buffer has size B with a sliding partition separating the input and output sections. The number of packets in the input section is designated by i; j denotes the number in the output section, and k is the number of outstanding acknowledgments for packets transmitted across the trunk. The indices i', j', and k' have the same meaning for node two. It is possible for k and k' to be as large as W, the window size.

The network model is Markovian and any particular state can be described by a 6-tuple (i, j, k, i', j', k'). As local traffic predominates and the service rates for foreign traffic decrease, the activity at the two nodes becomes more independent. Decomposability can be forced by having 99% or more local traffic. The resulting matrices are block diagonally dominant with each block having constant values for j, k, j', and k'.

This model produced matrices with some interesting properties. First, there were more diagonal blocks produced than were desirable, causing the solution of the aggregate system to be too costly. This problem was alleviated by concatenating adjacent diagonal blocks together to form larger blocks. In addition, a block can be very nearly cyclic, thus producing an eigenvalue arbitrarily close to -1 for some matrices.

Finally, it should be noted that these matrices grow very fast as the values of B and W are increased, e.g., with $B=10$ and $W=4$, the state space has dimensions 34,225.

4. Numerical results.

The following acronyms will be used:

GWSD—G. W. Stewart's direct method;
CD—Courtois's $O(e)$ direct method;
VANTD—Vantilborgh's $O(e^2)$ direct method;
VANTI—Vantilborgh's iterative method;
BGSA—block Gauss–Seidel iteration with aggregation;
BGS—block Gauss–Seidel iteration;
GSA—point Gauss–Seidel iteration with aggregation;
GS—point Gauss–Seidel iteration;
DMSA—Dodd, McAllister, and Stewart's power method with aggregation.

The results of the algorithms applied to the test matrices are given in Appendix I and Tables 1 and 2. The following conventions are used. Courtois's matrix is designated by his name. The V's refer to matrices generated from the multiprogramming model described in [Vant81], with A denoting service rates given by equation (3.1) and B denoting service rates given by equation (3.2). K denotes matrices derived from Kaufman's packet switching network where the smaller diagonal blocks were

TABLE 1
Matrices run using true probability vector as stopping criterion.

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
N1	16	0.817D+00	GWSD		622	0.292D+00	0.358D+00
			CD		321	0.296D+00	0.362D+00
			VANTD		725	0.308D−01	0.461D−01
			VANTI	9	6648	0.141D−11	0.869D−11
			BGSA	8	4015	0.293D−11	0.148D−10
			BGS	10	2557	0.247D−11	0.187D−10
			GSA	11	5282	0.338D−11	0.312D−10
			GS	12	2901	0.227D−11	0.257D−10
			DMSA	13	6184	0.628D−11	0.871D−10
N2	16	0.308D+00	GWSD		622	0.577D−01	0.187D+00
			CD		321	0.760D−01	0.246D+00
			VANTD		725	0.204D−02	0.215D−01
			VANTI	6	4539	0.835D−12	0.969D−09
			BGSA	5	2662	0.242D−11	0.866D−09
			BGS	9	2342	0.610D−12	0.241D−07
			GSA	12	5733	0.358D−11	0.483D−05
			GS	60	13221	0.706D−11	0.314D+20
			DMSA	15	7086	0.421D−11	0.193D−03

"*" means that EPS**ITER<10**–35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
N3	16	0.427D−01	GWSD		622	0.652D−02	0.153D+00
			CD		321	0.990D−02	0.232D+00
			VANTD		725	0.298D−04	0.164D−01
			VANTI	4	3133	0.147D−12	0.442D−07
			BGSA	4	2211	0.175D−12	0.528D−07
			BGS	8	2127	0.155D−11	0.140D+00
			GSA	11	5282	0.957D−11	0.111D+05
			GS	200	43321	0.836D−06	*
			DMSA	15	7086	0.780D−11	0.273D+10
N4	16	0.444D−02	GWSD		622	0.661D−03	0.149D+00
			CD		321	0.102D−02	0.230D+00
			VANTD		725	0.312D−06	0.158D−01
			VANTI	3	2430	0.112D−13	0.128D−06
			BGSA	3	1760	0.175D−12	0.200D−05
			BGS	7	1912	0.195D−11	0.572D+05
			GSA	10	4831	0.935D−11	0.314D+13
			GS	200	43321	0.149D−03	*
			DMSA	14	6635	0.264D−11	0.228D+22
N5	16	0.446D−03	GWSD		622	0.662D−04	0.149D+00
			CD		321	0.103D−03	0.230D+00
			VANTD		725	0.313D−08	0.158D−01
			VANTI	2	1727	0.201D−12	0.101D−05
			BGSA	3	1760	0.221D−13	0.250D−03
			BGS	6	1697	0.104D−11	0.132D+09
			GSA	9	4380	0.562D−11	0.808D+19
			GS	200	43321	0.319D−04	*
			DMSA	12	5733	0.665D−11	*
V1A	141	0.141D+00	GWSD		22837	0.298D−02	0.212D−01
			CD		8067	0.496D−01	0.352D+00
			VANTD		17094	0.201D−02	0.101D+00
			VANTI	7	86719	0.467D−11	0.428D−05
			BGSA	7	62441	0.554D−11	0.507D−05
			BGS	31	42545	0.640D−11	0.161D+16
			GSA	40	292107	0.952D−11	0.111D+24
			GS	200	132267	0.195D−03	*
			DMSA	150	1073217	0.958D−11	*
V2A	141	0.230D−01	GWSD		22837	0.425D−02	0.185D+00
			CD		8067	0.523D−01	0.228D+01
			VANTD		17094	0.208D−02	0.395D+01
			VANTI	7	86719	0.611D−11	0.182D+01
			BGSA	8	69982	0.652D−12	0.843D+01
			BGS	200	221854	0.226D−09	*
			GSA	40	292107	0.981D−11	*
			GS	200	132267	0.638D−02	*
			DMSA	134	959601	0.931D−11	*

"*" means that EPS**ITER < 10**-35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
V3A	141	0.973D−02	GWSD		22837	0.187D−02	0.192D+00
			CD		8067	0.581D−02	0.597D+00
			VANTD		17094	0.322D−04	0.340D+00
			VANTI	4	53011	0.977D−11	0.109D−02
			BGSA	5	47359	0.405D−11	0.464D−01
			BGS	25	36179	0.935D−11	*
			GSA	26	192693	0.637D−11	*
			GS	200	132267	0.853D−04	*
			DMSA	81	583248	0.907D−11	*
V4A	141	0.174D+00	GWSD		22837	0.194D−02	0.111D−01
			CD		8067	0.164D+00	0.942D+00
			VANTD		17094	0.113D−01	0.371D+00
			VANTI	9	109191	0.191D−11	0.128D−04
			BGSA	9	77523	0.169D−11	0.113D−04
			BGS	42	54216	0.869D−11	0.621D+21
			GSA	35	256602	0.733D−11	0.257D+16
			GS	200	132267	0.344D−04	*
			DMSA	125	895692	0.959D−11	*
V5A	141	0.334D−01	GWSD		22837	0.129D−01	0.386D+00
			CD		8067	0.180D+00	0.540D+01
			VANTD		17094	0.122D−01	0.109D+02
			VANTI	9	109191	0.148D−11	0.284D+02
			BGSA	10	85064	0.650D−12	0.374D+03
			BGS	200	221854	0.902D−08	*
			GSA	35	256602	0.542D−11	*
			GS	200	132267	0.114D−01	*
			DMSA	109	782076	0.826D−11	*
V6A	141	0.172D−01	GWSD		22837	0.198D−02	0.115D+00
			CD		8067	0.743D−02	0.433D+00
			VANTD		17094	0.490D−04	0.166D+00
			VANTI	5	64247	0.260D−12	0.175D−03
			BGSA	5	47359	0.657D−11	0.441D−02
			BGS	17	27691	0.968D−11	0.995D+19
			GSA	20	150087	0.791D−11	*
			GS	200	132267	0.987D−06	*
			DMSA	60	434127	0.819D−11	*
V4B	141	0.494D+00	GWSD		22837	0.524D−01	0.106D+00
			CD		8067	0.140D+00	0.284D+00
			VANTD		17094	0.108D−01	0.441D−01
			VANTI	11	131663	0.379D−111	0.878D−08
			BGSA	12	100146	0.147D−11	0.689D−08
			BGS	25	36179	0.974D−11	0.432D−03
			GSA	37	270804	0.974D−11	0.202D+01
			GS	200	132267	0.151D−08	*
			DMSA	200	1428267	0.425D−10	*

"*" means that EPS**ITER < 10**–35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
V5B	141	0.108D+00	GWSD		22837	0.140D+00	0.131D+01
			CD		8067	0.231D+00	0.215D+01
			VANTD		17094	0.202D−01	0.175D+01
			VANTI	12	142899	0.139D−11	0.581D+00
			BGSA	12	100146	0.454D−11	0.190D+01
			BGS	174	194268	0.972D−11	*
			GSA	38	277905	0.757D−11	*
			GS	200	132267	0.217D−01	*
			DMSA	135	966702	0.935D−11	*
V6B	141	0.441D−01	GWSD		22837	0.108D−01	0.245D+00
			CD		8067	0.182D−01	0.412D+00
			VANTD		17094	0.933D−03	0.479D+00
			VANTI	7	86719	0.746D−11	0.228D−01
			BGSA	8	69982	0.494D−11	0.343D+00
			BGS	46	58460	0.802D−11	*
			GSA	28	206895	0.629D−11	*
			GS	200	132267	0.190D−03	*
			DMSA	93	668460	0.958D−11	*
V7A	141	0.162D+00	GWSD		22837	0.313D−02	0.193D−01
			CD		8067	0.179D+00	0.110D+01
			VANTD		17094	0.182D−01	0.697D+00
			VANTI	11	131663	0.942D−12	0.474D−03
			BGSA	10	85064	0.972D−11	0.791D−03
			BGS	69	82863	0.815D−11	*
			GSA	42	306309	0.978D−11	0.164D+23
			GS	200	132267	0.319D−04	*
			DMSA	152	1087419	0.896D−11	*
V21A	38	0.788D−01	GWSD		2001	0.117D−01	0.149D+00
			CD		780	0.113D−01	0.143D+00
			VANTD		1742	0.203D−03	0.326D−01
			VANTI	6	8574	0.716D−12	0.299D−05
			BGSA	6	5762	0.314D−11	0.131D−04
			BGS	28	6646	0.541D−11	0.424D+20
			GSA	14	11280	0.458D−11	0.128D+05
			GS	200	31380	0.186D−04	*
			DMSA	55	42030	0.860D−11	*
V22A	38	0.178D−01	GWSD		2001	0.103D−01	0.580D+00
			CD		780	0.931D−02	0.523D+00
			VANTD		1742	0.140D−03	0.444D+00
			VANTI	5	7275	0.444D−11	0.249D−02
			BGSA	6	5762	0.372D−11	0.117D+00
			BGS	59	12939	0.988D−11	*
			GSA	14	11280	0.271D−11	0.848D+13
			GS	200	31380	0.112D−02	*
			DMSA	50	38280	0.712D−11	*

"*" means that EPS**ITER < 10**–35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
V24A	38	0.107D−01	GWSD		2001	0.668D−04	0.624D−02
			CD		780	0.289D−03	0.270D−01
			VANTD		1742	0.748D−07	0.652D−03
			VANTI	3	4677	0.326D−14	0.266D−08
			BGSA	3	3362	0.350D−12	0.285D−06
			BGS	5	1977	0.704D−12	0.501D−02
			GSA	6	5280	0.397D−11	0.263D+01
			GS	200	31380	0.334D−07	*
			DMSA	20	15780	0.626D−11	*
V25A	38	0.106D−01	GWSD		2001	0.667D−05	0.627D−03
			CD		780	0.291D−04	0.274D−02
			VANTD		1742	0.764D−09	0.674D−05
			VANTI	2	3378	0.173D−13	0.153D−09
			BGSA	2	2562	0.149D−11	0.132D−07
			BGS	3	1571	0.176D−12	0.146D−06
			GSA	3	3030	0.600D−12	0.498D−06
			GS	200	31380	0.316D−09	*
			DMSA	12	9780	0.719D−11	0.341D+13
V8A	141	0.378D−01	GWSD		22837	0.499D−01	0.132D+01
			CD		8067	0.555D−01	0.147D+01
			VANTD		17094	0.406D−02	0.285D+01
			VANTI	9	109191	0.150D−11	0.954D+01
			BGSA	10	85064	0.130D−11	0.220D+03
			BGS	162	181536	0.878D−11	*
			GSA	38	277905	0.689D−11	*
			GS	200	132267	0.897D−02	*
			DMSA	122	874389	0.954D−11	*
V9A	141	0.237D−01	GWSD		22837	0.187D−02	0.790D−01
			CD		8067	0.690D−02	0.291D+00
			VANTD		17094	0.441D−04	0.784D−01
			VANTI	5	64247	0.122D−12	0.162D−04
			BGSA	5	47359	0.298D−11	0.396D−03
			BGS	13	23447	0.713D−11	0.940D+10
			GSA	19	142986	0.826D−11	0.610D+20
			GS	200	132267	0.882D−07	*
			DMSA	57	412824	0.920D−11	*
K1	81	0.217D+00	CD		2763	0.435D+00	0.201D+01
			VANTD		6304	0.123D+00	0.261D+01
			VANTI	20	122163	0.577D−11	0.111D+03
			BGSA	19	35074	0.737D−11	0.307D+02
			BGS	15	17628	0.300D−11	0.276D−01
			GSA	54	64485	0.747D−11	*
			GS	92	40023	0.794D−11	*
			DMSA	200	231363	0.255D−06	*

"*" means that EPS**ITER < 10**-35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
K1NOC	81	0.283D+00	GWSD		18010	0.100D+00	0.354D+00
			CD		6288	0.397D+00	0.140D+10
			VANTD		13027	0.744D−01	0.927D+00
			VANTI	9	76650	0.131D−11	0.112D−06
			BGSA	9	62349	0.315D−11	0.268D−06
			BGS	15	15078	0.777D−12	0.128D−03
			GSA	19	120630	0.351D−11	0.896D−01
			GS	90	42738	0.860D−11	*
			DMSA	82	499764	0.769D−11	*
K2	81	0.503D−01	CD		2763	0.421D+00	0.836D+01
			VANTD		6304	0.999D−01	0.395D+02
			VANTI	20	122163	0.504D−11	0.472D+15
			BGSA	20	36668	0.207D−11	0.194D+15
			BGS	17	19340	0.394D−11	0.469D+11
			GSA	200	231363	0.676D−09	*
			GS	200	83763	0.312D−05	*
			DMSA	200	231363	0.125D−01	*
K2NOC	81	0.646D−01	GWSD		18010	0.356D−01	0.552D+00
			CD		6288	0.953D−01	0.148D+01
			VANTD		13027	0.521D−02	0.125D+01
			VANTI	6	53196	0.185D−11	0.255D−04
			BGSA	6	43887	0.354D−11	0.488D−04
			BGS	16	15619	0.914D−11	0.100D+09
			GSA	20	126648	0.482D−11	0.304D+13
			GS	200	87288	0.672D−06	*
			DMSA	78	475692	0.856D−11	*
K3	16	0.167D−01	CD		162	0.357D+00	0.214D+02
			VANTD		407	0.374D−01	0.135D+03
			VANTI	13	5167	0.580D−11	0.758D+12
			BGSA	9	1688	0.548D−11	0.552D+05
			BGS	9	995	0.683D−11	0.688D+05
			GSA	200	26762	0.195D−05	*
			GS	200	11362	0.316D−03	*
			DMSA	200	26762	0.228D−01	*
K3NOC	16	0.190D−01	GWSD		877	0.284D−02	0.149D+00
			CD		332	0.219D−03	0.115D−01
			VANTD		719	0.159D−05	0.438D−02
			VANTI	3	1760	0.183D−11	0.265D−06
			BGSA	4	1802	0.286D−13	0.217D−06
			BGS	7	786	0.811D−12	0.892D+00
			GSA	4	1752	0.595D−12	0.452D−05
			GS	200	11532	0.109D−06	*
			DMSA	200	71332	0.184D−08	*

"*" means that EPS**ITER < 10**–35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
K4	16	0.700D−02	CD		162	0.324D+00	0.463D+02
			VANTD		407	0.427D−01	0.871D+03
			VANTI	13	5167	0.837D−11	0.864D+17
			BGSA	9	1688	0.878D−11	0.218D+09
			BGS	10	1078	0.613D−12	0.217D+10
			GSA	200	26762	0.521D−04	*
			GS	200	11362	0.144D−01	*
			DMSA	200	26762	0.893D−01	*
K4NOC	16	0.800D−02	GWSD		877	0.123D−02	0.154D+00
			CD		332	0.469D−03	0.586D−01
			VANTD		719	0.175D−05	0.274D−01
			VANTI	3	1760	0.482D−12	0.941D−06
			BGSA	4	1802	0.931D−14	0.227D−05
			BGS	7	786	0.340D−11	0.162D+04
			GSA	4	1752	0.722D−13	0.176D−04
			GS	200	11532	0.198D−04	*
			DMSA	71	25537	0.933D−11	*
K5	100	0.600D+00	CD		3521	0.511D+00	0.851D+00
			VANTD		7996	0.129D+00	0.357D+00
			VANTI	17	129661	0.615D−11	0.363D−07
			BGSA	14	35829	0.203D−11	0.259D−08
			BGS	19	26050	0.352D−11	0.578D−07
			GSA	21	36617	0.554D−11	0.253D−06
			GS	38	22521	0.598D−11	0.161D−02
			DMSA	109	175305	0.839D−11	0.127D+14
K5NOC	100	0.700D+00	GWSD		49252	0.194D+00	0.277D+00
			CD		16757	0.290D+00	0.414D+00
			VANTD		34026	0.620D−01	0.127D+00
			VANTI	10	201497	0.275D−11	0.975D−10
			BGSA	11	201075	0.498D−11	0.252D−09
			BGS	15	27063	0.995D−11	0.210D−08
			GSA	18	314693	0.952D−11	0.585D−08
			GS	35	34257	0.603D−11	0.159D−05
			DMSA	87	1456781	0.769D−11	0.230D+03
K6	100	0.167D+00	CD		3521	0.337D+00	0.202D+01
			VANTD		7996	0.902D−01	0.325D+01
			VANTI	16	122241	0.421D−11	0.119D+02
			BGSA	14	35829	4.470D−11	0.368D+00
			BGS	18	24993	0.553D−11	0.561D+03
			GSA	83	134329	0.865D−11	*
			GS	129	68021	0.842D−11	*
			DMSA	200	318721	0.180D−04	*

"*" means that EPS**ITER < 10**-35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
K6NOC	100	0.175D+00	GWSD		49252	0.758D−01	0.433D+00
			CD		16757	0.174D+00	0.995D+00
			VANTD		34026	0.251D−01	0.819D+00
			VANTI	8	164549	0.363D−11	0.412D−05
			BGSA	8	150999	0.166D−11	0.189D−05
			BGS	18	28983	0.154D−11	0.650D+02
			GSA	20	347797	0.450D−11	0.620D+04
			GS	119	76257	0.856D−11	*
			DMSA	81	1357469	0.779D−11	*
K7	100	0.346D−01	CD		3521	0.361D+00	0.104D+02
			VANTD		7996	0.958D−01	0.799D+02
			VANTI	16	122241	0.317D−11	0.739D+12
			BGSA	15	37962	0.314D−11	0.254D+11
			BGS	19	26050	0.230D−11	0.129D+17
			GSA	200	318721	0.118D−05	*
			GS	200	103521	1.109D−03	*
			DMSA	200	318721	0.311D−01	*
K7NOC	100	0.350D−01	GWSD		49252	0.170D−01	0.487D+00
			CD		16757	0.489D−01	0.140D+01
			VANTD		34026	0.180D−02	0.147D+01
			VANTI	6	127601	0.124D−12	0.675D−04
			BGSA	6	117615	0.626D−13	0.340D−04
			BGS	18	28983	0.258D−11	0.416D+15
			GSA	20	347797	0.331D−11	0.435D+18
			GS	200	116757	0.543D−05	*
			DMSA	76	1274709	0.803D−11	*
K8	100	0.174D−01	CD		3521	0.365D+00	0.210D+02
			VANTD		7996	0.974D−01	0.322D+03
			VANTI	16	122241	0.331D−11	0.466D+17
			BGSA	15	37962	0.358D−11	0.877D+15
			BGS	19	26050	0.260D−11	0.694D+22
			GSA	200	318721	0.320D−03	*
			GS	200	103521	0.685D−02	*
			DMSA	200	318721	0.882D−01	*
K8NOC	100	0.175D−01	GWSD		49252	0.863D−02	0.493D+00
			CD		16757	0.257D−01	0.147D+01
			VANTD		34026	0.488D−03	0.160D+01
			VANTI	5	109127	0.300D−12	0.183D−03
			BGSA	5	100923	0.288D−11	0.175D−02
			BGS	17	28343	0.946D−11	0.699D+19
			GSA	19	331245	0.590D−11	0.142D+23
			GS	200	116757	0.382D−03	*
			DMSA	73	1225053	0.961D−11	*

"*" means that EPS**ITER < 10**–35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
K13	16	0.498D−02	CD		162	0.663D+00	0.133D+03
			VANTD		407	0.105D+00	0.423D+04
			VANTI	15	5937	0.555D−11	0.196D+24
			BGSA	12	2168	0.151D−11	0.655D+16
			BGS	12	1244	0.219D−11	0.954D+16
			GSA	200	26762	0.605D−04	*
			GS	200	11362	0.207D−01	*
			DMSA	200	26762	0.285D+00	*
K13NOC	16	0.500D−02	GWSD		877	0.232D−02	0.464D+00
			CD		332	0.180D−02	0.360D+00
			VANTD		719	0.414D−05	0.166D+00
			VANTI	3	1760	0.357D−11	0.285D−04
			BGSA	4	1802	0.124D−12	0.198D−03
			BGS	9	906	0.646D−11	0.331D+10
			GSA	4	1752	0.359D−13	0.574D−04
			GS	200	11532	0.469D−04	*
			DMSA	41	14887	0.941D−11	*
K14	100	0.560D−02	CD		3521	0.104D+01	0.185D+03
			VANTD		7996	0.235D+00	0.748D+04
			VANTI	17	129661	0.727D−11	*
			BGSA	16	40095	0.400D−11	*
			BGS	20	27107	0.399D−11	*
			GSA	200	318721	0.314D−01	*
			GS	200	103521	0.939D−01	*
			DMSA	200	318721	0.693D+00	*
K15	81	0.424D−02	CD		2763	0.362D+00	0.853D+02
			VANTD		6304	0.121D+00	0.672D+04
			VANTI	21	128133	0.382D−11	*
			BGSA	19	35074	0.677D−11	*
			BGS	18	20196	0.278D−11	*
			GSA	200	231363	0.520D−01	*
			GS	200	83763	0.875D−01	*
			DMSA	200	231363	0.259D+00	*
K16	16	0.572D−02	CD		162	0.829D+00	0.145D+03
			VANTD		407	0.811D−01	0.248D+04
			VANTI	15	5937	0.291D−11	0.126D+23
			BGSA	11	2008	0.888D−11	0.413D+14
			BGS	12	1244	0.242D−11	0.197D+16
			GSA	200	26762	0.291D−04	*
			GS	200	11362	0.276D−01	*
			DMSA	200	26762	0.289D+00	*

"*" means that EPS**ITER < 10**-35.

TABLE 1 (*Continued*)

Matrix	n	EPS	ALG	ITER	EOC	ERR	ERR/EPS**ITER
K16NOC	16	0.597D-02	GWSD		877	0.323D-02	0.541D+00
			CD		332	0.242D-04	0.406D-02
			VANTD		719	0.224D-07	0.629D-03
			VANTI	3	1760	0.285D-13	0.134D-06
			BGSA	3	1443	0.915D-12	0.430D-05
			BGS	7	786	0.251D-11	0.930D+04
			GSA	3	1397	0.140D-12	0.657D-06
			GS	200	11532	0.301D-06	*
			DMSA	200	71332	0.137D-05	*
COURTOIS	8	0.100D-02	GWSD		129	0.170D-03	0.170D+00
			CD		71	0.726D-03	0.726D+00
			VANTD		173	0.104D-06	0.104D+00
			VANTI	2	399	0.615D-11	0.615D-05
			BGSA	3	367	0.594D-14	0.594D-05
			BGS	5	301	0.793D-11	0.793D+04
			GSA	12	1163	0.997D-11	*
			GS	200	8271	0.188D-03	*
			DMSA	62	5713	0.977D-11	*

“*” means that EPS**ITER<10**-35.

TABLE 2

Matrices run using difference between successive iterations as stopping criterion.

Matrix	n	EPS	ALG	ITER	EOC	ERR
K30	400	0.153D-01	VANTI	29	2159781	0.980D-11
			BGS	34	289813	0.891D-11
			BGSA	28	569299	0.774D-11
K30NOC	400	0.179D-01	VANTI	6	2474500	0.132D-11
			BGS	30	460944	0.576D-11
			BGSA	7	2725975	0.262D-13
K31	1156	0.179D-01	VANTI	46	21826305	0.962D-11
			BGS	49	1834670	0.946D-11
			BGSA	44	4476077	0.811D-11
V30	1701	0.252D-01	VANTI	10	8206058	0.199D-11
			BGS	200	8221355	0.279D-05
			BGSA	10	3695875	0.623D-11
V31	1701	0.172D-01	VANTI	10	8206058	0.206D-11
			BGS	200	8221355	0.124D-03
			BGSA	11	3992747	0.349D-12
V32	1701	0.408D-01	VANTI	10	8206058	0.215D-11
			BGS	200	8221355	0.340D-11
			BGSA	10	3695875	0.340D-11

concatenated. K##NOC indicates Kaufman matrices with nonconcatenated diagonal blocks (i.e., the natural partitioning was used). The order of the matrix is denoted by n. The maximum degree of coupling is denoted in the column labeled EPS. The number of iterations required for the relative error to be less than 10^{-11} is given in the column labeled ITER. Computations were halted at 200 iterations if convergence had not been reached. The column labeled EOC denotes the total estimated number of floating point operations required. Details can be found in [Koury83]. The last column labeled ERR/EPS**ITER is the relative error divided by e^k where k is the number of iterations.

Direct methods. In Table 1, when applying CD, the off-diagonal block probability mass was added to the diagonal element in the same row to approximate $\mathbf{Q}_I^*$. As is clear from the universal results VANTD is superior in accuracy and comparable in total work and hence is the direct method of choice.

Iterative methods. For the smaller matrices (i.e., dimension < 150), convergence was assumed when $\|\mathbf{x}^{(k)}-\mathbf{x}\|_\infty/\|\mathbf{x}\|_\infty<10^{-11}$ where $\mathbf{x}$ is the true probability vector computed using the QR algorithm. Some larger matrices of order approximately 2000 were also tested. For these matrices, convergence was assumed when $\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|_\infty/\|\mathbf{x}^{(k+1)}\|_\infty<10^{-11}$. When the relative difference between successive iterations was used as the convergence criterion, an L was appended to the algorithm acronym (e.g., BGSA → BGSAL).

Vantilborgh shows that on the kth iteration, we have an approximation of $O(e^k)$ using an absolute convergence criterion. However, we cannot be assured that the approximation has accuracy $O(e^k)$ if we use a relative error convergence criterion. In addition, if the state space has size 10^6, then knowing that $\|\hat{\mathbf{x}}-\mathbf{x}\|_\infty$ is of order $e=10^{-3}$ is meaningless when the largest element of the stationary distribution may be of order 10^{-4}.

In general, BGSA and BGS were found to produce ten digits of accuracy with fewer operations than the other algorithms. VANTI was a close second. However, we note that BGS and BGSA require more storage than VANTI since LU decompositions of the diagonal blocks must be stored. We note that the diagonal blocks are column diagonally dominant M-matrices [Berm79], so that the LU decomposition of the diagonal blocks is guaranteed to be stable with no pivoting.

Since the aggregation step can be expensive when the number of blocks is large, we tried to mitigate this problem by concatenating smaller blocks to form larger diagonal blocks. However, this concatenation procedure, in some cases, degraded the performance of some of the algorithms and resulted in reducible blocks. For CD, we must have irreducible diagonal blocks in order to determine the $\mathbf{x}_I^*$, $1\leqq I\leqq N$. This problem can be overcome by distributing the off-diagonal block mass into the reverse diagonal. Since this distribution is arbitrary, one would not expect CD to produce a good approximation. Indeed, for the K matrices, the results for CD and VANTD are poor.

For the iterative methods, no distribution of off-diagonal row mass into the diagonal blocks is necessary. We note that block concatenation does however increase the storage requirements of BGS and BGSA. In our tests, there were some cases where block concatenation reduced the number of operations needed in BGSA significantly, but no a priori criterion could be established to predict this. The balance between decreasing operations as one increases storage with block concatenation may be crucial on some systems when using large matrices.

Matrices V21A through V25A were produced by fixing the number of thinking users to 10, the multiprogramming level to 3, d to .95, p to 0.03, and then varying λ

from 0.1 for V21A to 0.00001 in V25A in multiples of 10. For V22A, aggregation is seen to accelerate convergence of BGS dramatically. For matrix V5A, $e=.0334$. In this case BGS performs poorly, and BGSA performs very well.

BGS converged rapidly for all of the K matrices. However, using an aggregation step sometimes accelerated convergence, especially when the natural partitioning of the matrix was observed. For example, for matrix K5NOC, BGS converged in 15 iterations, and BGSA converged in 11 iterations in spite of the fact that $e=.7$. For matrix K3NOC, BGS converged in 7 iterations and BGSA converged in 4 iterations, with $e=0.019$.

The only time aggregation degraded the convergence rate of the iterative schemes was when the natural partitioning of the system was violated. And in these cases, the degradation of the convergence rates was not severe.

Our tests indicate that one would not want to use GS on NCD systems. GSA performed better than GS in every case. If the storage required for BGS or BGSA is prohibitively expensive, VANTI may be the algorithm of choice since its complexity was close to that of BGS and BGSA in most cases.

DMSA is only of theoretical interest. The convergence rate is seen to be asymptotic to the $N+1$ eigenvalue of the matrix [Dodd81].

Our test results show that aggregation can be a powerful tool for accelerating the convergence of iterative methods used in determining the stationary probability vector of queueing networks.

It is likely that performing an aggregation step on every iteration is unnecessary. An investigation showing just how often aggregation should be performed would be helpful.

Miranker and Pan [Mir80] have suggested an aggregative procedure for accelerating the convergence of iterative solutions of general linear systems. In this procedure they pick two arbitrary different partitionings of the coefficient matrix. Information gathered from both of these partitionings is used for an aggregative acceleration process. In our algorithms we tried to reduce the cost of aggregation by concatenating diagonal blocks. In some cases, this procedure did not work well. However, it may be that using two different partitions with concatenated blocks of the state space may be a tenable procedure for aggregation.

REFERENCES

[Berm79] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[Chan75] K. M. CHANDY, U. HERZOG AND L. WOO, *Approximate analysis of queueing networks*, IBM J. Res. Dev., 19 (1975).

[Cour75] P. J. COURTOIS, *Error analysis in nearly-completely decomposable stochastic systems*, Econometrica, 43 (1975), pp. 691–709.

[Cour76] P. J. COURTOIS AND G. LOUCHARD, *Approximation of eigencharacteristics in nearly-completely decomposable stochastic systems*, Stochastic Process. Appl., 4 (1976), pp. 283–296.

[Cour77] P. J. COURTOIS, *Decomposability: Queueing and Computer System Applications*, Academic Press, New York, 1977.

[Dodd81] S. L. DODD, D. F. MCALLISTER AND W. J. STEWART, *An iterative method for the exact solution of Coxian queueing networks*, Proceedings ACM/SIGMETRICS Conf. on Measurement and Modeling of Computer Systems, September 1981, pp. 97–104.

[Dodd81a] ———, *Decomposition and perturbation in stochastic matrices*, TR 81-03, Dept. Computer Science Technical Report, North Carolina State Univ., Raleigh, NC, January, 1981.

[Fund71] R. E. FUNDERLIC AND M. T. HEATH, *Linear compartmental analysis of ecosystems*, Oak Ridge Nat. Lab. Rept. IBP-74-4, Oak Ridge, TN, 1971.

[Fund81] R. E. FUNDERLIC AND J. B. MANKIN, *Solution of homogeneous systems of linear equations arising from compartmental models*, SIAM J. Sci. Stat. Comput., 2 (1981), pp. 375–383.

[Fund83] R. E. FUNDERLIC AND R. J. PLEMMONS, *A combined direct-iterative method for certain M-matrix linear systems*, this Journal, (to appear).

[Gele75] E. GELENBE, *On approximate computer system models*, J. Assoc. Comput. Mach., 22 (1975), pp. 261–269.

[Gele78] E. GELENBE AND A. KURINCKX, *Random injection control of multiprogramming in virtual memory*, IEEE Trans. Software Engr., SE-4 (1978), pp. 2–17.

[Harr83] W. J. HARROD AND R. J. PLEMMONS, *Comparison of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. and Stat. Computing, 5 (1984), pp. 453–469.

[Karl75] S. KARLIN AND H. TAYLOR, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, Inc., New York, 1975.

[Kauf81] L. KAUFMAN, B. GOPINATH AND E. F. WUNDERLICH, *Sparse matrix algorithms for a packet network control analysis*, IEEE Trans. Comm., COM-29 (1981), pp. 453–465.

[Kauf81a] ———, *Analysis of racket network congestion control using sparse matrix algorithms*, IEEE Trans. Comm., 29 (1981), pp. 453–465.

[Kauf81b] L. KAUFMAN, J. B. SEERY AND J. A. MORRISON, *Overflow models for dimension* PBX *feature packages*, Bell. Syst. Tech. J., 60 (1981), pp. 661–676.

[Kauf81c] L. KAUFMAN, *Solving large linear systems arising in queueing problems*, Proc. Bielefield Conference on Large Linear Systems, Springer-Verlag, New York.

[Koba74] H. KOBAYASHI, *Application of the diffusion approximation to queueing networks* I: *equilibrium queue distributions*, J. Assoc. Comput. Mach., 21 (1974), pp. 316–328.

[Konh76] A. G. KONHEIM AND M. REISER, *A queueing model with finite waiting room and blocking*, J. Assoc. Comput. Mach., 23 (1976), pp. 210–229.

[Konh78] ———, *Finite capacity queueing systems with applications in computer modeling*, SIAM J. Comput., 7 (1978), pp. 210–229.

[Kour83] J. R. KOURY, *A comparison of numerical techniques for solving nearly completely decomposable Markovian queueing networks*, M.Sc. thesis, North Carolina State Univ., Raleigh, 1983.

[Marc64] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[Mari78] R. MARIE, *Modélisation par réseaux de files d'altente*, Thesis: Docteur-es-Sciences Mathématiques, Université de Rennes, France, November 1978.

[Merl78] D. MERLE, D. POTIER AND M. VERAN, *A tool for computer system performance analysis*, Centre Scientifique CII-HB, Grenoble, France, 1978.

[Mira80] W. L. MIRANKER AND V. Y. PAN, *Methods of aggregation*, Linear Alg. and Appl., 29 (1980), pp. 231–257.

[Mull80] B. MÜLLER, *An iterative method for the exact solution of general queueing networks*, Dissertation, universitat Dortmund. Abt. Informatik, 1980.

[Orte72] J. M. ORTEGA, *Numerical Analysis, a Second Course*, Academic Press, New York, 1972.

[Shep51] C. W. SHEPPARD AND A. S. HOUSEHOLDER, *The mathematical basis of the interpretation of tracer experiments in closed steady-state systems*, J. Appl. Phys., 22 (1951), pp. 510–520.

[Simo61] H. A. SIMON AND A. ANDO, *Aggregation of variables in dynamic systems*, Econometrica, 20 (1961), pp. 111–132.

[Stew73] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[Stew76] ———, *Algorithm* 506, HQR3 *and* EXCHNG: *Fortran subroutines for calculation and ordering the eigenvalues of a real upper Hessenberg matrix* [*F*2], ACM Trans. Math. Software, 2 (1976), pp. 275–280.

[Stew80] G. W. STEWART, *Computable error bounds for aggregated Markov chains*, Tech. Rep. 901, Computer Science Sept., Univ. Maryland, College Park, May 1980.

[Stew82] G. W. STEWART AND W. J. STEWART, *An iterative method for the solution of nearly decomposable Markov chains*, Tech. Rep., Dept. Computer Science, Univ. of Maryland, College Park, 1982.

[Stew78] W. J. STEWART, *A comparison of numerical techniques in Markov modelling*, Comm. ACM, 21 (1978), pp. 144–151.

[Stew81] W. J. STEWART AND A. JENNINGS, *A simultaneous iteration algorithm for real matrices*, ACM Trans. Math. Software, 7 (1981), pp. 184–198.

[Vant81] H. VANTILBORGH, *The error of aggregation in decomposable systems*, (R 453), MBLE Research Laboratory, Brussels, Belgium, 1981.

[Varg62] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[Wall66] V. L. WALLACE AND R. S. ROSENBERG, RQA-1, *the recursive queue analyzer*, Tech. Rep. 2, Dept. Electrical Engineering, Univ. Michigan, Ann Arbor, February, 1966.

[Wilk71] J. H. WILKINSON AND C. REINSCH, *Handbook of Automatic Computation, Vol. 2, Linear Algebra*, Springer-Verlag, Berlin, 1971.

[Youn71] D. M. YOUNG, *Iterative solutions of large linear systems*, Academic Press, London, 1971.

[Youn73] D. M. YOUNG AND R. T. GREGORY, *A Survey of Numerical Mathematics, Vol.* II, Addison-Wesley, Reading, MA, 1973.

[Zarl76] R. L. ZARLING, *Numerical solution of nearly decomposable queueing networks*, Ph.D. thesis, Dept. Computer Science, Univ. North Carolina, Chapel Hill, 1976.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

THE GENERALIZED TODA FLOW, THE QR ALGORITHM AND THE CENTER MANIFOLD THEORY*

MOODY T. CHU†

Abstract. A continuous version of the classical QR algorithm, known as the Toda flow, is generalized to complex-valued, full and nonsymmetric matrices. It is shown that this generalized Toda flow, when sampled at integer times, gives the same sequence of matrices as the QR algorithm applied to the matrix $\exp(G(X_0))$. When $G(X)=X$, global convergence is deduced for the case of distinct real eigenvalues. This convergence property can also be understood locally by the center manifold theory. It is shown that the manifold of upper triangular matrices with decreasing main diagonal entries is the stable center manifold for the Toda flow. One interesting example is given to demonstrate geometrically the dynamical behavior of this flow.

Key words. generalized Toda flow, QR-algorithm, upper Hessenberg matrix, QR-decomposition, isospectral family, center manifold theorem

1. Introduction. For over ten years, the QR algorithm has been recognized as the most efficient way of finding eigenvalues for small matrices. The process usually involves repeated applications of a rather complicated similarity transformation to the underlying matrix. To be more precise, we recall the following facts about this algorithm [6], [8].

LEMMA 1.1. *For any matrix X (in $\mathbb{C}^{n\times n}$), there exists unitary matrix Q such that $X=QR$ where R is an upper triangular matrix with real nonnegative diagonal entries. Moreover, Q is unique if X is nonsingular.*

Let X_0 be the given matrix whose eigenvalues are to be found; then, the QR algorithm generates a sequence of matrices $\{X_k\}$ by

$$\begin{aligned} &X_1=X_0, \\ &X_k=Q_kR_k, \qquad X_{k+1}=R_kQ_k, \qquad k=1,2,\cdots. \end{aligned} \tag{1.1}$$

Observe that

$$X_{k+1}=R_kQ_k=Q_k^*X_kQ_k=\cdots=(Q_0\cdots Q_k)^*X_0(Q_0\cdots Q_k), \tag{1.2}$$

and that the matrix $Q_0\cdots Q_k$ is still unitary; so, this sequence of matrices is isospectral. Indeed, we know [6], [8]

LEMMA 1.2. *If X_0 is nonsingular and its eigenvalues have distinct moduli, then the sequence $\{X_k\}$ in (1.1) converges essentially in the sense that the entries below the principal diagonal tend to zero, the moduli of those above the diagonal tend to fixed values and the entries on the principal diagonal tend to the eigenvalues of X_0.*

Motivated by the discrete case (1.2), one might want to construct a one-parameter family of unitary matrix $Q(t)$ with $Q(0)=I$ so that the isospectral family of matrices,

$$X(t)=Q^*(t)X_0Q(t), \tag{1.3}$$

would also have some asymptotic behavior as t becomes large. Toward this end, we

* Received by the editors September 21, 1982, and in revised form April 22, 1983. This research was supported in part by a grant from the Faculty Research and Professional Development Fund at North Carolina State University.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

first realize that the flow $X(t)$ defined by (1.3) will satisfy the differential equation

$$\begin{aligned}\dot{X}(t)&=\dot{Q}^*(t)X_0Q(t)+Q^*(t)X_0\dot{Q}(t)\\&=-Q^*(t)\dot{Q}(t)Q^*(t)Q_0X(t)+Q^*(t)X_0\dot{Q}(t)\\&=-Q^*(t)\dot{Q}(t)X(t)+X(t)Q^*(t)\dot{Q}(t).\end{aligned}\tag{1.4}$$

If we call

$$B(t)=Q^*(t)\dot{Q}(t),\tag{1.5}$$

then $X(t)$ is the solution to the differential system

$$\dot{X}=[X,B]\triangleq XB-BX,\qquad X(0)=X_0.\tag{1.6}$$

Since $Q(t)$ is a unitary matrix, it is easy to see that $B(t)$ has to be skew-hermitian, i.e., $B^*(t)=-B(t)$ for every t. But, we also realize that the above argument can be reversed. Namely, if $B(t)$ is a family of skew-hermitian matrices, let $Q(t)$ be the solution to the system

$$\dot{Q}(t)=Q(t)B(t),\qquad Q(0)=I,\tag{1.7}$$

then $Q(t)$ is unitary for all t and the solution $X(t)$ to (1.6) can be expressed as in (1.3). Thus, the problem whether $X(t)$ has nice and useful asymptotic behavior really depends on the choice of $B(t)$. The obvious choice of B, a constant skew-hermitian matrix, gives a family of isospectral matrices

$$X(t)=e^{-Bt}X_0\,e^{Bt},\tag{1.8}$$

which apparently leads us to nothing.

Recently the study of a special nonlinear wave equation, known as the Toda lattice, has suggested a feasible choice of this $B(t)$ for real Jacobi matrix X_0 (see, e.g., [4], [5], [10]). The asymptotics of the resulting differential system (1.6) have been extensively studied—the solution flow converges to a diagonal matrix. Furthermore, the standard QR algorithm can be recovered as the evaluation at integer times of this flow under the exponential transformation.

In the first part of this paper, we generalize this classical Toda flow to complex-valued, full and nonsymmetric matrices. In addition, we allow a much larger flexibility in selecting this skew-hermitian matrix $B(t)$. As a result, the standard QR algorithm still can be interpreted as this flow sampled at integer times. Global convergence for the case of real eigenvalues can be deduced easily when $B(t)$ is chosen properly. In fact, in the second part of this paper, it is shown that the manifold of upper triangular matrices with decreasing main diagonal entries is the stable center manifold for this generalized Toda flow.

In [7] a variety of isospectral flows on band matrices are considered. The asymptotics for hermitian matrices are also detailed there. These results can be viewed as special cases of ours in this paper and in [2], [3].

This paper is organized as follows. In § 2, we discuss the generalized Toda flow, its relevance to the QR algorithm and its global convergence property. In § 3, we cite without proof the center manifold theorems and then demonstrate its application to the Toda flow by one simple example. Section 4 contains the general treatment of this application in higher dimensional space. Finally, we study the dynamical behavior of the simplest Toda flow both qualitatively and quantitatively. This reveals very interesting geometrical insights to the general convergence theorem of the QR algorithm.

2. Preliminary results. The following lemma is the cornerstone of our development while its proof is almost trivial.

LEMMA 2.1. *The general space* $\mathbb{C}^{n\times n}$ *is the direct sum of the space* $U(n)$ *of all upper triangular matrices with real diagonal elements and the space* $O(n)$ *of all skew-hermitian matrices. To be more specific, given any* $X \in \mathbb{C}^{n\times n}$, *let*

$$X = X^+ + X^0 + X^-, \tag{2.1}$$

where X^- *and* X^+ *are the strictly lower and upper triangular matrices of* X, *respectively, and*

$$X^0 = \mathrm{Re}\,(X^0) + i\,\mathrm{Im}\,(X^0)$$

is the diagonal matrix of X *with* $\mathrm{Re}\,(X^0)$ *and* $\mathrm{Im}\,(X^0)$ *denoting the real and the imaginary parts of* X^0, *respectively. Define*

$$\begin{aligned} \Pi_u(X) &= X^+ + X^{-*} + \mathrm{Re}\,(X^0), \\ \Pi_0(X) &= X^- - X^{-*} + i\,\mathrm{Im}\,(X^0); \end{aligned} \tag{2.2}$$

then $X = \Pi_u(X) + \Pi_0(X)$ *with* $\Pi_u(X) \in U(n)$ *and* $\Pi_0(X) \in O(n)$.

Given a matrix $X_0 \in \mathbb{C}^{n\times n}$ and fixed, let $G(z)$ be an analytic function defined on a domain Ω containing all eigenvalues of X_0. We shall consider the following initial value problem

$$\begin{aligned} \dot{X} &= [X, \Pi_0(G(X))] = X \cdot (\Pi_0(G(X))) - (\Pi_0(G(X))) \cdot X, \\ X(0) &= X_0, \end{aligned} \tag{2.3}$$

where $\dot{} = d/dt$ and $t \in \mathbb{R}$. This differential equation is known as the generalized Toda flow. Here, we have used the notation that $G(X)$ means the matrix-valued contour integral

$$G(X) = \frac{1}{2\pi i} \int_\Gamma G(\lambda)(\lambda I - X)^{-1}\, d\lambda, \tag{2.4}$$

where Γ is any contour that surrounds the spectrum of X in Ω. It is well-known [9] that this integral exists and is independent of the choice of Γ, provided only that Γ surrounds the spectrum in Ω.

LEMMA 2.2. *Let* $Q(t)$ *be the solution to the problem*

$$\dot{Q} = Q \cdot (\Pi_0(G(X))), \qquad Q(0) = I, \tag{2.5}$$

then $Q(t)$ *is a unitary matrix.*

Proof. Observe that

$$(QQ^*)^{\dot{}} = \dot{Q}Q^* + Q\dot{Q}^* = Q(\Pi_0(G(X)))Q^* + Q(\Pi_0(G(X)))^*Q^* = 0.$$

Since

$$Q(0)Q^*(0) = I,$$

it follows that, for every $t \in \mathbb{R}$, we have

$$Q(t)Q^*(t) = I.$$

LEMMA 2.3. *The flow* $X(t)$ *of Problem* (2.3) *satisfies*

$$X(t) = Q^*(t)X_0Q(t). \tag{2.6}$$

Proof. Let

$$Z(t)=Q(t)X(t)Q^*(t).$$

Then

$$\begin{aligned}\dot{Z}&=\dot{Q}XQ^*+Q\dot{X}Q^*+QX\dot{Q}^*\\&=Q(\Pi_0(G(X)))XQ^*+Q[X(\Pi_0(G(X)))-(\Pi_0(G(X)))X]Q^*\\&\quad-QX(\Pi_0(G(X)))Q^*=0.\end{aligned}$$

This shows that

$$Z(t)=Z(0)=X(0)=X_0.$$

Of course, the expression (2.6) does not solve the Problem (2.3) at all because the matrices $Q(t)$ are defined implicitly in (2.5). However, two important consequences follow from this lemma.

COROLLARY 2.1. *The maximal interval of existence for Problem* (2.3) *is* $(-\infty, \infty)$.

COROLLARY 2.2. *For any* $t\in\mathbb{R}$, $X(t)$ *is always unitarily similar to* X_0. *In particular, the flow* $X(t)$ *is isospectral.*

We can even identify the matrix $Q(t)$ in (2.5) through the following lemma.

LEMMA 2.4. *If the matrix* $e^{tG(X_0)}$ *has the QR-decomposition*

$$e^{tG(X_0)}=Q(t)R(t) \tag{2.7}$$

as defined in Lemma 2.1, *then* $Q(t)$ *solves Problem* (2.5).

Proof. Taking derivatives on both sides of (2.7), we have

$$(QR)^{\cdot}=\dot{Q}R+Q\dot{R}=(e^{tG(X_0)})^{\cdot}=G(X_0)QR.$$

It follows that

$$Q^*\dot{Q}+\dot{R}R^{-1}=Q^*G(X_0)Q=G(Q^*X_0Q). \tag{2.8}$$

Define

$$\tilde{X}(t)=Q^*(t)X_0Q(t).$$

Observe that in (2.8), $Q^*\dot{Q}\in O(n)$ and $\dot{R}R^{-1}\in U(n)$; thus, we conclude from Lemma 2.1 that

$$Q^*\dot{Q}=\Pi_0(G(\tilde{X})). \tag{2.9}$$

But, on the other hand, we have

$$\dot{\tilde{X}}=\dot{Q}^*X_0Q+Q^*X_0\dot{Q}=\dot{Q}^*Q\tilde{X}+\tilde{X}Q^*\dot{Q}=[\tilde{X},\Pi_0(G(\tilde{X}))]$$

and

$$\tilde{X}(0)=X_0.$$

This implies that

$$\tilde{X}(t)=X(t).$$

By (2.9), the assertion is proved.

Lemma 2.4 is the key connection between the differential equation (2.3) and the QR algorithm because now we can prove the following theorem [4], [10].

THEOREM 2.1. *Suppose* $X(t)$ *solves Problem* (2.3). *For* $k=0, \pm1, \pm2, \cdots$, *suppose*

$$e^{G(X(k))}=Q_kR_k. \tag{2.10}$$

Then,

$$e^{G(X(k+1))}=R_kQ_k. \tag{2.11}$$

Proof. It is known, from Lemma 2.3 and Lemma 2.4, that

$$X(t)=Q^*(t)X_0Q(t)$$

and

$$e^{tG(X_0)}=Q(t)R(t).$$

So,

$$R(t)Q(t)=Q^*(t)\,e^{tG(X_0)}Q(t)=e^{tG(Q^*(t)X_0Q(t))}=e^{tG(X(t))}.$$

For $k=0$, choose $t=1$; then,

$$e^{G(X(0))}=Q(1)R(1)$$

implies

$$e^{G(X(1))}=R(1)Q(1).$$

Since the Toda equation is autonomous, the assertion follows.

In other words, the above theorem asserts that the isospectral matrices produced by the QR-algorithm (2.1) are related to those isospectral matrices produced by evaluations of the Toda flow (2.3) at integer times. In particular, setting $G(z)=\ln z$ and supposing that X_0 satisfies the conditions in Lemma 1.2, we then recover the QR-algorithm from the Toda flow. The constantly shifted QR-algorithm is equivalent to the choice $G(z)=\ln(z-c)$. For the purpose of convenient computation we shall, henceforth, be interested only in the choice $G(z)=z$.

The classical QR-algorithm takes great advantage of the fact that one cycle of the simple QR process applied to a Hessenberg matrix results in a Hessenberg matrix. We now show that the same fact is preserved along the entire Toda flow (for the case $G(z)=z$).

LEMMA 2.5. *If X is an upper Hessenberg matrix, so is $\dot{X}=[X,\Pi_0X]$.*

Proof. As an upper Hessenberg matrix, it must be that $x_{ij}=0$, whenever $1\leqq j\leqq n-2$ and $j+2\leqq i\leqq n$. With i and j in this range and fixed, it is also true that $x_{ij}=0$ implies $x_{ik}=0$ for $k\leqq j$ and $x_{kj}=0$ for $i\leqq k$. Let b_{ij} denote the (i,j)-component of the matrix $\Pi_0(X)$. Then, from (2.3), we know

$$\dot{x}_{ij}=\sum_{k=1}^{n}x_{ik}b_{kj}-\sum_{k=1}^{n}b_{ik}x_{kj}=x_{i,j+1}b_{j+1,j}-b_{i,i-1}x_{i-1,j},$$

since $b_{kj}=0$ unless $k=j-1,j$ or $j+1$ and $b_{ik}=0$ unless $k=i-1,i$ or $i+1$. Note that $b_{j+1,j}=x_{j+1,j}$ and $b_{i,i-1}=x_{i,i-1}$. If $i=j+2$, then $\dot{x}_{ij}=0$. If $i>j+2$, then $x_{i,j+1}=x_{i-1,j}=0$ and still $\dot{x}_{ij}=0$.

Observe that, from Lemma 2.3, the trajectory $X(t)$ is bounded. Indeed $\|X(t)\|_2=\|X(0)\|_2$. It follows that its ω-limit set is nonempty, compact and connected. We are interested in finding this set. A special case is well known [4], [5], [10].

LEMMA 2.6. *If X_0 is a Jacobi matrix with positive off-diagonal elements, then X_0 has simple spectrum $\{\lambda_1>\lambda_2>\cdots>\lambda_n\}$ and $X(t)\to\operatorname{diag}\{\lambda_1,\cdots,\lambda_n\}$ exponentially as $t\to\infty$.*

Indeed we may have more general results.

THEOREM 2.2. *If the matrix $X_0\in\mathbb{R}^{n\times n}$ has distinct real eigenvalues $\{\lambda_1>\cdots>\lambda_n\}$, then the Toda flow $X(t)$ of (2.3) with $G(z)=z$ converges to an upper triangular matrix with the eigenvalues appearing on the diagonal in decreasing order.*

Remark. We do not assume any symmetric property on X_0. In fact, we may even weaken the hypothesis by assuming X_0 has real eigenvalues only.

Proof. The matrix e^{X_0} has positive eigenvalues $\{e^{\lambda_1} > \cdots > e^{\lambda_n}\}$. So, by Lemma 2.2 and Theorem 2.1, the sequence $e^{X(k)}$ converges to an upper triangular matrix with diagonal elements $\{e^{\lambda_1}, \cdots, e^{\lambda_n}\}$. But the continuity of the logarithm from (2.4) and the isospectral property from Lemma 2.3 imply that $X(k)$ must converge to an upper triangular matrix. Being an autonomous system with upper triangular matrices as critical points, the flow $X(t)$ to Problem (2.3) must converge as $t \to \infty$ by the continuous dependence property.

We shall investigate this property further in the next two sections.

3. Center manifold theory and the Toda flow in $\mathbb{R}^{2\times 2}$. First of all, we cite four results concerning the center manifold [1]. Given a system

$$\dot{x} = Ax + f(x, y), \qquad \dot{y} = By + g(x, y), \tag{3.1}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ and A, B are constant matrices such that all eigenvalues of A have zero real parts while all those of B have negative real parts, the functions f and g are C^2 with

$$\begin{aligned} f(0,0) &= 0, \qquad Df(0,0) = 0,\\ g(0,0) &= 0, \qquad Dg(0,0) = 0. \end{aligned} \tag{3.2}$$

An invariant manifold for (3.1) with the property

$$y = h(x), \qquad h(0) = 0, \qquad Dh(0) = 0, \tag{3.3}$$

is called a center manifold.

LEMMA 3.1. *There exists a center manifold for* (3.1) *with* $h \in C^2$.

LEMMA 3.2. *The flow on this center manifold is governed by*

$$\dot{u} = Au + f(u, h(u)). \tag{3.4}$$

LEMMA 3.3. *The stability of the zero solution of* (3.4) *is equivalent to the stability of the zero solution of* (3.1). *In particular, suppose that the zero solution of* (3.4) *is stable and that* $(x(t), y(t))$ *is a solution of* (3.1) *with* $(x(0), y(0))$ *sufficiently small. Then there exists a solution* $u(t)$ *of* (3.4) *such that as* $t \to \infty$,

$$\begin{aligned} x(t) &= u(t) + O(e^{-\mu t}),\\ y(t) &= h(u(t)) + O(e^{-\mu t}), \end{aligned} \tag{3.5}$$

where $\mu > 0$ *is a constant.*

LEMMA 3.4. *Let* ϕ *be a* C^1 *mapping of a neighborhood of the origin in* $\mathbb{R}^n$ *into* $\mathbb{R}^m$ *with* $\phi(0) = 0$ *and* $D\phi(0) = 0$. *Consider the operator* M *on* ϕ,

$$M\phi(x) = D\phi(x) \cdot (Ax + f(x, \phi(x))) - B\phi(x) - g(x, \phi(x)). \tag{3.6}$$

Suppose that as $x \to 0$, *there exists* $q > 1$ *such that*

$$M\phi(x) = O(|x|^q);$$

then, as $x \to 0$, *we have an approximation*

$$|h(x) - \phi(x)| = O(|x|^q).$$

We now consider an example. Denote a matrix $X \in \mathbb{R}^{2\times 2}$ as

$$X = \begin{bmatrix} x_{11}, & x_{12}\\ x_{21}, & x_{22} \end{bmatrix}.$$

Then the Toda equation (2.3), with $G(X)=X$, is given by

$$\dot{X}=\begin{bmatrix} x_{21}(x_{12}+x_{21}), & x_{21}(x_{22}-x_{11}) \\ x_{21}(x_{22}-x_{11}), & -x_{21}(x_{12}+x_{21}) \end{bmatrix}. \tag{3.7}$$

Identifying $\mathbb{R}^{2\times 2}$ as $\mathbb{R}^4$, we may rewrite the above equation as

$$\begin{bmatrix} x_{11} \\ x_{21} \\ x_{12} \\ x_{22} \end{bmatrix}=\left[\begin{array}{cc|cc} 0 & x_{21} & x_{21} & 0 \\ -x_{21} & 0 & 0 & x_{21} \\ \hline -x_{21} & 0 & 0 & x_{21} \\ 0 & -x_{21} & -x_{21} & 0 \end{array}\right]\begin{bmatrix} x_{11} \\ x_{21} \\ x_{12} \\ x_{22} \end{bmatrix}. \tag{3.8}$$

Let

$$T=\begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix}$$

be an arbitrary upper triangular matrix in $\mathbb{R}^{2\times 2}$ with $t_{22}<t_{11}$. Note that T is an equilibrium point of (3.7). We shall use, without causing any confusion, the same notation to represent both a matrix in $\mathbb{R}^{2\times 2}$ and a vector in $\mathbb{R}^4$ henceforth. Set

$$Y=X-T;$$

then

$$\dot{Y}=\begin{bmatrix} 0 & t_{12} & 0 & 0 \\ 0 & t_{22}-t_{11} & 0 & 0 \\ 0 & t_{22}-t_{11} & 0 & 0 \\ 0 & -t_{12} & 0 & 0 \end{bmatrix}\begin{bmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \end{bmatrix}+\begin{bmatrix} y_{21}(y_{21}+y_{12}) \\ y_{21}(y_{22}-y_{11}) \\ y_{21}(y_{22}-y_{11}) \\ -y_{21}(y_{21}+y_{12}) \end{bmatrix}. \tag{3.9}$$

Notice that this matrix $\partial[X,\Pi_0 X]/\partial X|_{X=T}$ has eigenvalues 0, 0, 0 and $t_{22}-t_{11}$. Indeed, if we let

$$P^T=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad E=\begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -\alpha \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $\alpha=t_{12}/(t_{22}-t_{11})$ and define $Z=E^{-1}P^T Y$, then

$$Z=\begin{bmatrix} z_{11} \\ z_{12} \\ z_{22} \\ z_{21} \end{bmatrix}=\begin{bmatrix} y_{11}-\alpha y_{21} \\ y_{12}-y_{21} \\ y_{22}+\alpha y_{21} \\ y_{21} \end{bmatrix}$$

and

$$\begin{bmatrix} z_{11} \\ z_{12} \\ z_{22} \end{bmatrix}=\begin{bmatrix} \\ 0 \\ \\ \end{bmatrix}\begin{bmatrix} z_{11} \\ z_{12} \\ z_{22} \end{bmatrix}+\begin{bmatrix} z_{21}(z_{12}+2z_{21})-\alpha z_{21}(z_{22}-z_{11}-2\alpha z_{21}) \\ 0 \\ -z_{21}(z_{12}+2z_{21})+\alpha z_{21}(z_{22}-z_{11}-2\alpha z_{21}) \end{bmatrix}, \tag{3.10}$$

$$\dot{z}_{21}=(t_{22}-t_{11})z_{21}+z_{21}(z_{22}-z_{11}-2\alpha z_{21}).$$

By Lemma 3.1, there exists a three-dimensional center manifold $z_{21}=h(z_{11},z_{12},z_{22})$.

To find an approximation to h, set

$$M\phi(z_{11}, z_{12}, z_{22}) = D\phi(z_{11}, z_{12}, z_{22})\begin{bmatrix} \phi(z_{12}+2\phi)-\alpha\phi(z_{22}-z_{11}-2\alpha\phi) \\ 0 \\ -\phi(z_{12}+2\phi)+\alpha\phi(z_{22}-z_{11}-2\alpha\phi) \end{bmatrix} \tag{3.11}$$
$$-(t_{22}-t_{11})\phi-\phi(z_{22}-z_{11}-2\alpha\phi)=0,$$

according to (3.6). Obviously, $\phi \equiv 0$ is a trivial solution to (3.11). By Lemma 3.2, the flow on this center manifold is governed by

$$\dot{u}=0. \tag{3.12}$$

Since with any initial value the solution to (3.12) is a constant, by Lemma 3.3, we know with any matrix sufficiently near 0 as the initial value for (3.10), the corresponding solution $Z(t)$ converges to a constant upper triangular matrix exponentially as $t \to \infty$. Equivalently, with any matrix sufficiently close to T as the initial value for (3.7), the corresponding flow $X(t)$ then converges to a constant upper triangular matrix exponentially.

We shall illustrate the above results numerically in § 5.

4. General treatment in $\mathbb{R}^{n\times n}$. In this section, we generalize the result in § 3 to the general case $\mathbb{R}^{n\times n}$. Essentially, all these proofs can even be extended to the complex case $\mathbb{C}^{n\times n}$.

Let T be a fixed real upper triangular matrix with diagonal elements $t_{nn} < \cdots < t_{22} < t_{11}$. We establish the major result through several lemmas.

LEMMA 4.1. *If the matrix $X \in \mathbb{R}^{n\times n}$ is identified as the vector in $\mathbb{R}^{n^2}$, then the Toda flow* (2.3) *can always be written as*

$$\dot{X}=f(X)=C(X)\cdot X, \tag{4.1}$$

where $C(X) \in \mathbb{R}^{n^2\times n^2}$ is linear in X. More precisely, if $C(X)$ is written in block form as an $n\times n$ array of $n\times n$ blocks, then its (i, i)-block is always the matrix $-\Pi_0 X$, its (i, j)-block is $-x_{ij}I$ if $i>j$ and $x_{ji}I$ if $i<j$.

LEMMA 4.2. *The derivative of f at X acting on B is given by*

$$Df(X)B=C(B)X+C(X)B. \tag{4.2}$$

Proof. This follows from the fact that $f(X+B)-f(X)=C(B)X+C(B)B+C(X)B$ and the fact that $C(B)B=o(\|B\|)$.

LEMMA 4.3. *The Jacobian $Df(X)$ evaluated at the upper triangular matrix T always has $n(n+1)/2$ zero eigenvalues and $n(n-1)/2$ eigenvalues of the form $t_{jj}-t_{ii}$ with $j>i$. Furthermore, this matrix is always diagonalizable.*

Proof. We first construct this $n^2\times n^2$ Jacobian matrix $Df(T)$ explicitly. Let e_{ij} be the standard elementary matrix whose only nonzero component 1 is at the (i, j)-position. From the definition of the function C, it is obvious that $C(T)=0$ and $C(e_{ij})=0$ whenever $i \leqq j$. By (4.2), the matrix $Df(T)$, therefore, has at least $n(n+1)/2$ columns identically zero. Its other columns can be constructed by the recipe described below which can be verified by direct observations. For $i>j$, the $(n\cdot(j-1)+i)$th column of $Df(T)$, given by

$$Df(T)(e_{ij})=C(e_{ij})T,$$

is the sum of two matrices $A=(a_{\alpha\beta})$ and $B=(b_{\alpha\beta})$ (identified as column vectors) where, for $\alpha, \beta = 1, \cdots, n$,

$$\begin{aligned} &a_{j\beta} = t_{i\beta}, \qquad a_{i\beta} = -t_{j\beta}, \\ &b_{\alpha j} = t_{\alpha i}, \qquad b_{\alpha i} = -t_{\alpha j}, \\ &a_{\alpha\beta} = b_{\alpha b} = 0, \quad \text{otherwise.} \end{aligned}$$

It is clear that $Df(T)$ always has $n(n+1)/2$ zero eigenvalues. To calculate the remaining $\binom{n}{2}$ eigenvalues, it is sufficient to consider the $\binom{n}{2} \times \binom{n}{2}$ submatrix obtained by deleting all elements on the $(n \cdot (j-1)+i)$th column and row of $Df(T)$ for $i \leqq j$. For example, when $n=4$, the nontrivial columns of $Df(T)$ form the following 16×6 matrix according to the above recipe.

$$\left[\begin{array}{ccc|cc|c}
t_{12} & t_{13} & t_{14} & & & \\
t_{22}-t_{11} & t_{23} & t_{24} & 0 & & 0 \\
0 & t_{33}-t_{11} & t_{34} & & & \\
0 & 0 & t_{44}-t_{11} & & & \\
\hline
t_{22}-t_{11} & 0 & 0 & t_{13} & t_{14} & \\
-t_{12} & 0 & 0 & t_{23} & t_{24} & 0 \\
0 & -t_{12} & 0 & t_{32}-t_{22} & t_{34} & \\
0 & 0 & -t_{12} & 0 & t_{44}-t_{22} & \\
\hline
t_{23} & t_{33}-t_{11} & 0 & -t_{12} & 0 & t_{14} \\
-t_{13} & 0 & 0 & t_{33}-t_{22} & 0 & t_{24} \\
0 & -t_{13} & 0 & -t_{23} & 0 & t_{34} \\
0 & 0 & -t_{13} & 0 & -t_{23} & t_{44}-t_{33} \\
\hline
t_{24} & t_{34} & t_{44}-t_{11} & 0 & -t_{12} & -t_{13} \\
-t_{14} & 0 & 0 & t_{34} & t_{44}-t_{22} & -t_{23} \\
0 & -t_{14} & 0 & -t_{24} & 0 & t_{44}-t_{33} \\
0 & 0 & -t_{14} & 0 & -t_{24} & -t_{34}
\end{array}\right].$$

By deleting the corresponding rows, we obtain the 6×6 matrix

$$\left[\begin{array}{ccc|cc|c}
t_{22}-t_{11} & t_{23} & t_{24} & & & \\
0 & t_{33}-t_{11} & t_{34} & 0 & & 0 \\
0 & 0 & t_{44}-t_{11} & & & \\
\hline
0 & -t_{12} & 0 & t_{33}-t_{22} & t_{34} & 0 \\
0 & 0 & -t_{12} & 0 & t_{44}-t_{22} & \\
\hline
0 & 0 & -t_{13} & 0 & -t_{23} & t_{44}-t_{33}
\end{array}\right].$$

In fact, it can be shown by induction that, in general, the $\binom{n}{2} \times \binom{n}{2}$ submatrix is always in block form with all blocks above the diagonal block identically zero and all diagonal blocks are upper triangular matrices of which the diagonal entries are of the form $t_{jj} - t_{ii}$, $j > i$. The assertion of this lemma thus follows.

LEMMA 4.4. *The problem of* (2.3), *with* $G(X)=X$, *can be written as*

$$\dot{Y}=Df(T)\cdot Y+C(Y)\cdot Y, \tag{4.3}$$

if $Y=X-T$.

Proof. It follows from (4.2) that

$$Df(T)\cdot Y=C(Y)\cdot T.$$

This implies that

$$\begin{aligned} f(Y+T)&=C(Y+D)\cdot(Y+T)=C(Y)\cdot(Y+T)\\ &=f(Y)+Df(T)\cdot Y. \end{aligned}$$

We are now ready to prove the major result.

THEOREM 4.1. *If* X_0 *is sufficiently close to the matrix* T, *then the flow* $X(t)$ *of the Problem* (2.3), *with* $G(X)=X$, *converges to a constant upper triangular matrix exponentially as* $t\to\infty$.

Proof. Let P^T be the permutation matrix which, when acting on Y, shifts all elements strictly under the diagonal to the bottom components when regarded as a vector in $\mathbb{R}^{n^2}$. Let

$$\hat{Y}=P^TY,$$

so that we can represent this matrix $\hat{Y}$ as

$$\hat{Y}=\begin{bmatrix}\mathcal{U}\\ \mathcal{L}\end{bmatrix},$$

where $\mathcal{U}$ is just the upper triangular part of Y and $\mathcal{L}$ is the strictly lower triangular part of Y. The equation (4.3) now becomes

$$\dot{\hat{Y}}=\left[\begin{array}{c:c} 0 & \end{array}\right]\begin{bmatrix}\mathcal{U}\\ \mathcal{L}\end{bmatrix}+P^TC(P\hat{Y})P\hat{Y}. \tag{4.4}$$

There exists a nonsingular matrix E whose columns consist of eigenvectors of the matrix in (4.4). This matrix E may be written in the form

$$E=\left[\begin{array}{c:c} I & M\\ \hdashline 0 & N \end{array}\right].$$

It is easy to see that

$$E^{-1}=\left[\begin{array}{c:c} I & -MN^{-1}\\ \hdashline O & N^{-1} \end{array}\right].$$

Define

$$Z=E^{-1}\hat{Y}=E^{-1}P^TY,$$

and denote Z as

$$Z=\begin{bmatrix}\hat{\mathcal{U}}\\ \hat{\mathcal{L}}\end{bmatrix}.$$

Then

$$Z=\begin{bmatrix}\hat{\mathcal{U}}\\ \hat{\mathcal{L}}\end{bmatrix}=\begin{bmatrix}\mathcal{U}-MN^{-1}\mathcal{L}\\ N^{-1}\mathcal{L}\end{bmatrix}. \tag{4.5}$$

Furthermore, we have the following canonical differential equation

$$\dot{\hat{\mathscr{U}}} = F(\hat{\mathscr{U}}, \hat{\mathscr{L}}), \qquad \dot{\hat{\mathscr{L}}} = B\hat{\mathscr{L}} + G(\hat{\mathscr{U}}, \hat{\mathscr{L}}), \tag{4.6}$$

where B is the diagonal matrix with elements $d_{jj} - d_{ii}$, $j > i$ and

$$\begin{bmatrix} F(\hat{\mathscr{U}}, \hat{\mathscr{L}}) \\ G(\hat{\mathscr{U}}, \hat{\mathscr{L}}) \end{bmatrix} = E^{-1}P^TC(PEZ)PEZ. \tag{4.7}$$

With the condition of T, by Lemma 3.1, there exists a center manifold $\hat{\mathscr{L}} = h(\hat{\mathscr{U}})$ for (4.6). This manifold satisfies the operator equation

$$M\phi(\hat{\mathscr{U}}) = D\phi(\hat{\mathscr{U}})F(\hat{\mathscr{U}}, \phi(\hat{\mathscr{U}})) - B\phi(\hat{\mathscr{U}}) - G(\hat{\mathscr{U}}, \phi(\hat{\mathscr{U}})). \tag{4.8}$$

Observe that $F(\hat{\mathscr{U}}, 0) = 0$ and $G(\hat{\mathscr{U}}, 0) = 0$ according to (4.7). So $h(\hat{\mathscr{U}}) \equiv 0$ is a trivial solution to (4.8). Since the flow on this center manifold remains constant according to Lemma 3.2, we know that, by Lemma 3.3, if X_0 is close enough to T, then the flow $X(t)$ converges to a constant upper triangular matrix exponentially.

Remark. The hypothesis that $t_{nn} < \cdots < t_{22} < t_{11}$ (or that they have distinct real parts) is essential in the sense that having two equal diagonal elements is not generic, i.e. a small perturbation might result in bifurcation. For a geometric interpretation, the reader may look into § 5.

Remark. Although what we get from the center manifold theory is only a much weaker "local" convergence theorem for the unshifted QR algorithm, the reader should remember that we have shown an "almost" global convergence theorem in § 2. For more detailed asymptotic analysis, see, e.g., [2] and [3].

Remark. Theorem 4.1 does not mention the condition on the moduli of eigenvalues. However, by the words "sufficiently close to the matrix T," it does imply necessarily that eigenvalues of X_0 have distinct moduli. See, e.g., the diagram in the next section.

5. Analysis of the simplest Toda flow. In this section we shall analyze qualitatively the dynamics of the Toda flow in $\mathbb{R}^{2\times 2}$. For a general treatment of Jacobi matrices in $\mathbb{R}^{3\times 3}$, refer to [4].

The flow is governed by the equation

$$\dot{X} = \begin{bmatrix} x_{21}(x_{12} + x_{21}), & x_{21}(x_{22} - x_{11}) \\ x_{21}(x_{22} - x_{11}), & -x_{21}(x_{12} + x_{21}) \end{bmatrix}. \tag{5.1}$$

Notice that $x_{21}(t)$ can never change sign by the uniqueness theorem. Since $\dot{x}_{12} = \dot{x}_{21}$ and $\dot{x}_{22} = -\dot{x}_{11}$, we know that, for all $t \in \mathbb{R}$,

$$x_{12} = x_{21} + c, \qquad x_{22} = -x_{11} + d, \tag{5.2}$$

for some constants c and d which are fixed by the initial data. It is sufficient to study the flow of $x_{21}(t)$ and $x_{11}(t)$. Let us rename

$$x(t) = x_{21}(t), \qquad y(t) = x_{11}(t), \tag{5.3}$$

then we have

$$\dot{x} = -2xy + dx, \qquad \dot{y} = 2x^2 + cx. \tag{5.4}$$

Define

$$z = y - \frac{d}{2}; \tag{5.5}$$

then the equation becomes

$$\dot{x} = -2xz, \qquad \dot{z} = 2x^2 + cx. \tag{5.6}$$

Obviously, $x(t) \equiv 0$ and $z(t) \equiv$ constant is a trivial solution to (5.6). Without loss, we shall assume

$$c < 0. \tag{5.7}$$

Figure 1 is a computer-plotted vector field of (5.6).

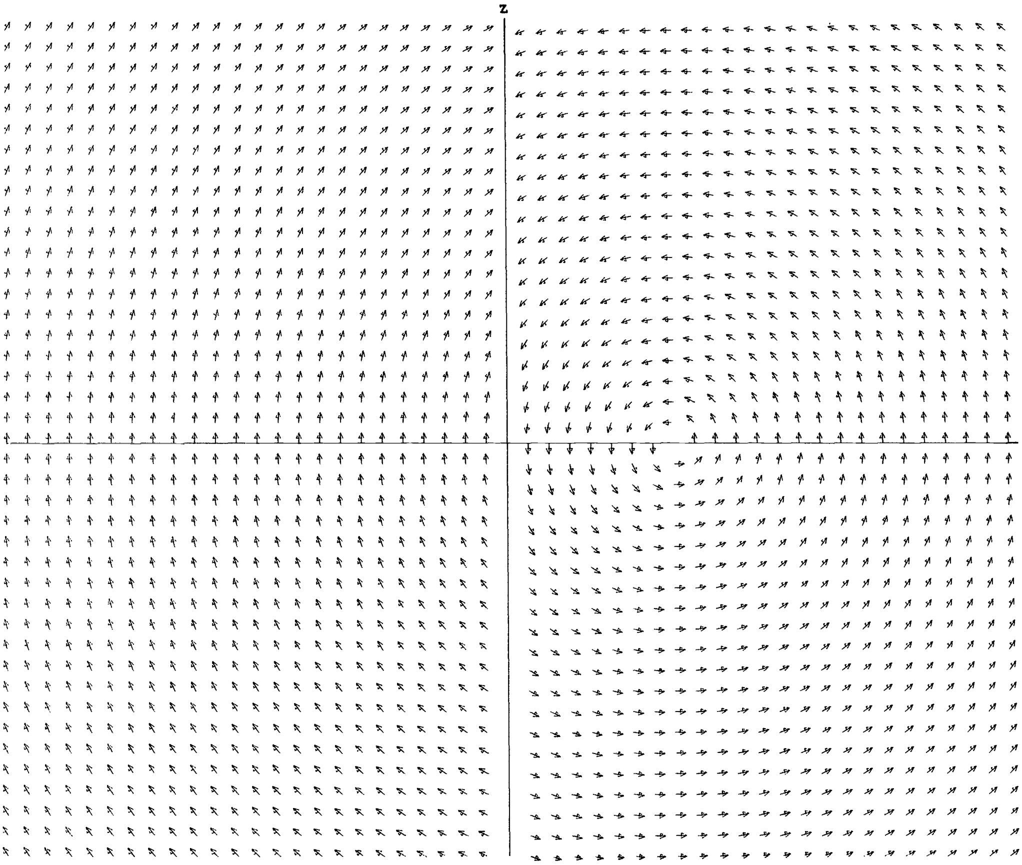


FIG. 1

It is clear that if $x(0) < 0$, then $x(t)$ converges to 0 and $z(t)$ converges to a nonnegative constant as $t \to \infty$. Note that $A = (-c/2, 0)$ is another critical point. We claim that near A there are periodic solutions. Indeed, we have the following lemma.

LEMMA 5.1. *With $(x(0), z(0))$ lying in the interior of the disc*

$$z^2 + \left(x + \frac{c}{2}\right)^2 = \frac{c^2}{4}, \tag{5.8}$$

the solution to (5.6) *is a circle in the xz-plane.*

Proof. With the following proof, we know even more about the trajectory. Let

$$w = x + \frac{c}{2};$$

then (5.6) becomes

$$\dot{w} = z(c-2w), \qquad \dot{z} = w(2w-c). \tag{5.9}$$

Changing into polar coordinates, i.e.,

$$w = r\cos\theta, \qquad z = r\sin\theta, \tag{5.10}$$

makes it easy to see that

$$\dot{r} = 0, \qquad \dot{\theta} = 2r\cos\theta - c. \tag{5.11}$$

Obviously, all solutions of (5.6) are circular sections.

Remark. This critical circle (5.8) can easily be checked to be the criterion for X_0 to have real eigenvalues, i.e. with c and d fixed; then the matrix X_0, with components outside the circle (5.8) and satisfying (5.2), has real eigenvalues and vice versa. Evidently, the upper half z-axis represents the stable center manifold which we have discussed in § 3. So we have

LEMMA 5.2. *If the matrix X_0 has real eigenvalues (not necessarily distinct), then the flow $X(t)$ converges to an upper triangular matrix with diagonal elements listed in descending order.*

We give some numerical examples below.

Example 1.

$$X_0 = \begin{bmatrix} -1, & -1 \\ 3, & 2 \end{bmatrix}.$$

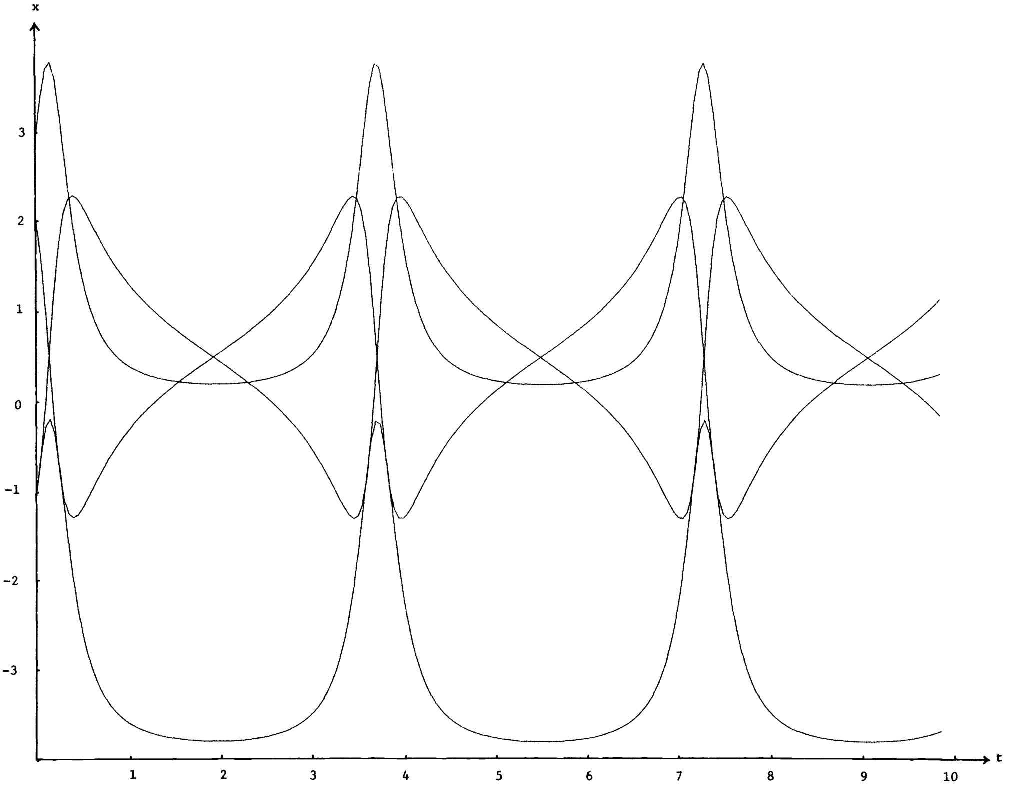


FIG. 2

Example 2.

$$X_0=\begin{bmatrix}-4, & 1\\ 4, & -2\end{bmatrix}.$$

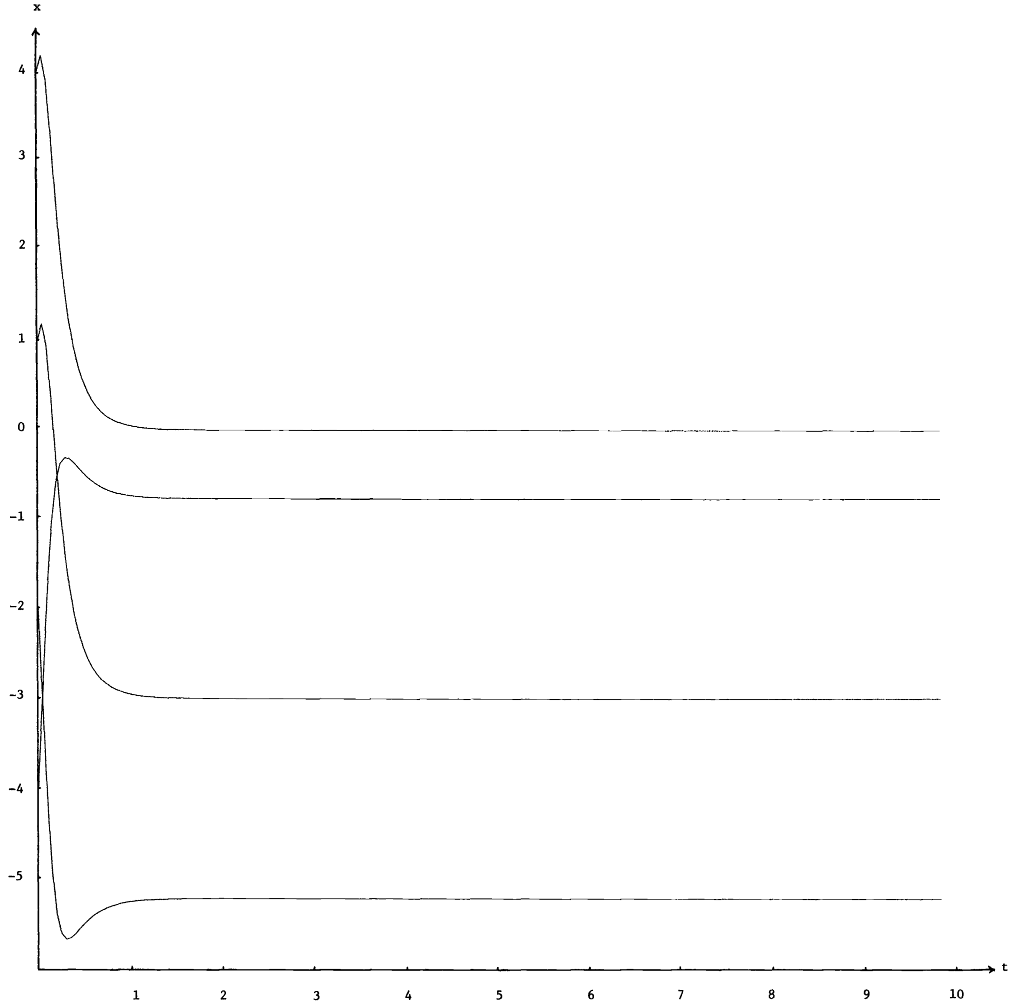


FIG. 3

Remark. All the numerical experiments are carried out with the standard IVP solver—DE and STEP. The special structure of the Toda lattice and numerical experiences indicate that a more efficient method can be designed. In particular, the error control is much easier to handle. We are still studying this implementation problem, see, e.g., [12].

Remark. If the matrix X_0 is symmetric, then $c=0$. From the diagram, it is obvious why the flow always converges to a diagonal matrix.

Remark. Observe that the critical circle (5.8) corresponds to the case where the matrix X_0 has real multiple roots. The limit point $(0,0)$ is unstable in the sense that small perturbations may result in nonconvergent periodic solutions.

Remark. The equation (5.11) can be used to calculate the time needed so that x is sufficiently close to 0. For example, if $\theta(0)=0$, then the time $T(\varepsilon)$ needed in order that the flow reaches the vertical line $x=\varepsilon$ is given by

$$T(\varepsilon)=\frac{1}{\sqrt{4r^2-c^2}}\ln\frac{\sqrt{4r^2-c^2}\sqrt{\dfrac{2r-(c+2\varepsilon)}{2r+(c+2\varepsilon)}}+(2r-c)}{\sqrt{4r^2-c^2}\sqrt{\dfrac{2r-(c+2\varepsilon)}{2r+(c+2\varepsilon)}}-(2r-c)}.$$

Apparently, $T(\varepsilon)\to\infty$ as $\varepsilon\to 0$. The case $c=0$ has been discussed in [4].

REFERENCES

[1] J. CARR, *Applications of Center Manifold Theory*, Applied Mathematical Science, 35, Springer-Verlag, Berlin, 1981.

[2] M. T. CHU, *On the global convergence of the Toda lattice for real normal matrices and its application to the eigenvalue problems*, SIAM J. Math. Anal., 15 (1984), pp. 98–104.

[3] ———, *Asymptotic analysis of the Toda lattice on diagonalizable matrices*, submitted, 1982.

[4] P. DEIFT, T. NANDA AND C. TOMEI, *Differential equations for the symmetric eigenvalue problem*, SIAM J. Numer. Anal., 20 (1983), pp. 1–22.

[5] H. FLASCHKA, *Discrete and periodic illustrations of some aspects of the inverse method*, in Dynamical System, Theory and Applications, J. Moser, ed., Lecture Notes in Physics, 38, Springer-Verlag, Berlin, 1975.

[6] J. G. F. FRANCIS, *The QR transformation, a unitary analogue to the LR transformation*, Comput. J., 4 (1961), pp. 265–271.

[7] T. NANDA, *Isospectral flows on band matrices*, Ph.D. thesis, New York University, New York, 1982.

[8] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[9] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.

[10] W. W. SYMES, *The QR algorithm and scattering for the finite nonperiodic Toda lattice*, Physica, 4D (1982), pp. 275–280.

[11] D. S. WATKKINS, *Understanding the QR algorithm*, SIAM Rev., 24 (1982), pp. 427–440.

[12] ———, *Experience with the Toda flow method of calculating eigenvalues*, preprint, 1982.

[13] J. H. WILKINSON AND C. REINSCH, *Linear Algebra*, Springer-Verlag, Berlin, 1971.

© 1984 Society for Industrial and Applied Mathematics
006

LAYOUTS FOR THE SHUFFLE-EXCHANGE GRAPH BASED ON THE COMPLEX PLANE DIAGRAM*

FRANK THOMSON LEIGHTON,† MARGARET LEPLEY† AND GARY L. MILLER†

Abstract. The shuffle-exchange graph is one of the best structures known for parallel computation. Among the things, a shuffle-exchange computer can be used to compute discrete Fourier transforms, multiply matrices, evaluate polynomials, perform permutations and sort lists. The algorithms needed for these operations are quite simple and many require no more than logarithmic time and space per processor. In this paper, we analyze the algebraic structure of the shuffle-exchange graph in order to find area-efficient embeddings of the graph in a two-dimensional grid. The results are applicable to the design of Very Large Scale Integration (VLSI) circuit layouts for a shuffle-exchange computer.

Key words. area-efficient chip layouts, complex plane diagram, graph embedding, necklace, shuffle-exchange graph, Thompson grid model, Very Large Scale Integration (VLSI)

1. Introduction. The shuffle-exchange graph has long been recognized as one of the best structures known for parallel computation. Among its many applications, a shuffle-exchange computer can be used to compute discrete Fourier transforms, multiply matrices, evaluate polynomials, perform permutations and sort lists [S71], [P80], [S80]. The algorithms needed for these operations are quite simple and many require no more than logarithmic time and space per processor.

Recent developments in Very Large Scale Integration (VLSI) circuit technology have made it possible to fabricate large numbers of very simple processors on a single chip. As most of the processors contained in a shuffle-exchange computer are very simple, the shuffle-exchange graph serves as an excellent basis upon which to design and build chip-sized microcomputers. One of the main difficulties with such an architecture, however, is the problem of routing the wires which link the processors together in a shuffle-exchange network. Current fabrication technology limits the designer to two or three layers of insulated wiring on a chip and demands that he make the chip as small in area as possible.

Abstracted, the designer's problem becomes the mathematical question of how to embed the shuffle-exchange graph in the smallest possible two-dimensional grid. Thompson was the first to formalize the question mathematically. In his thesis [T80], he showed that any *layout* (i.e., embedding in a two-dimensional grid) of the N-node shuffle-exchange graph requires at least $\Omega(N^2/\log^2 N)$ area. In addition, he described a layout requiring only $O(N^2/\log^{1/2} N)$ area. Shortly thereafter, Hoey and Leiserson [HL80] described an embedding for the shuffle-exchange graph in the complex plane (which we call the *complex plane diagram*) and showed how the diagram could be used to find an $O(N^2/\log N)$-area layout for the N-node shuffle-exchange graph.

In this paper, we investigate the algebraic properties of the complex plane diagram in order to find several $O(N^2/\log^{3/2} N)$-area layouts for the N-node shuffle-exchange graph. In addition to being *asymptotically* superior to previously discovered layouts, the layouts described in this paper are also superior for *small* values of N. In fact, one of these layouts serves as the basis for the more recent work of Leighton and

[*] Received by the editors June 28, 1982, and in revised form March 24, 1983. This research was supported in part by the National Science Foundation under grant MCS 80-07756, the Defense Advanced Research Projects Agency under grant N00014-80-C-0622, and the Bantrell Foundation.

[†] Mathematics Department and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Miller who have described *optimal* layouts for *small* shuffle-exchange graphs in [LM81].

Subsequent to the completion of the research presented in this paper, we learned that Rodeh and Steinberg independently discovered an $O(N^2/\log^{3/2} N)$-area layout for the N-node shuffle-exchange graph. Their work is also based on the complex plane diagram and appears in [SR81]. Even more recently, Kleitman, Leighton, Lepley and Miller [KLLM81] have discovered an entirely new method for laying out shuffle-exchange graphs which can be used to find *asymptotically optimal* $O(N^2/\log^2 N)$-area layouts. Although their layouts are not entirely practical, they are the only layouts known to achieve Thompson's lower bound asymptotically.

The remainder of the paper is divided into six sections. In § 2, we define the shuffle-exchange graph and the grid model of a chip. We also describe Thompson's $O(N^2/\log^{1/2} N)$-area layout for the N-node shuffle-exchange graph. In § 3, we define the complex plane diagram for the shuffle-exchange graph and mention several of its properties. In § 4, we describe several layouts for the shuffle-exchange graph which are based on the complex plane diagram. These include a straightforward $O(N^2/\log N)$-area layout and several new $O(N^2/\log^{3/2} N)$-area layouts. Section 5 contains some remarks and open questions, and §§ 6 and 7 contain the acknowledgments and references.

2. Preliminaries.

2a. The shuffle-exchange graph. The *shuffle-exchange graph* comes in various sizes. In particular, there is an N-node shuffle-exchange graph for every N which is a power of two. Each node of the $(N = 2^k)$-node shuffle-exchange graph is associated with a unique k-bit binary string $a_{k-1} \cdots a_0$. Two nodes w *and* w' are linked via a *shuffle edge* if w' is a left or right cyclic 1-shift of w (i.e., if $w = a_{k-1} \cdots a_0$ and $w' = a_{k-2} \cdots a_0 a_{k-1}$ or $w' = a_0 a_{k-1} \cdots a_1$, respectively). Two nodes w and w' are linked via an *exchange edge* if w and w' differ only in the last bit (i.e., if $w = a_{k-1} \cdots a_1 0$ and $w' = a_{k-1} \cdots a_1 1$ or vice versa). As an example, we have drawn the 8-node shuffle-exchange graph in Fig. 1. Note that the shuffle edges are drawn with solid lines while the exchange edges are drawn with dashed lines. We shall follow this convention throughout the paper.

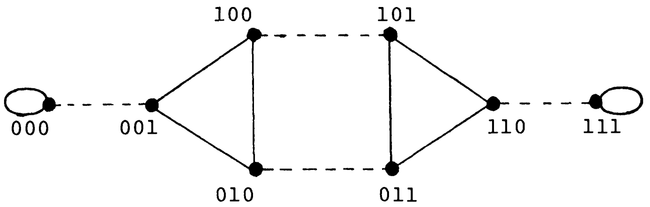


FIG. 1. *The* 8-*node shuffle-exchange graph.*

By replacing the nodes and edges of the shuffle-exchange graph by processors and wires (respectively), the shuffle-exchange graph can be transformed into a very powerful parallel computer (which we call the *shuffle-exchange computer*). The computational power of the shuffle-exchange computer is partly derived from the fact that every pair of nodes in an N-node shuffle-exchange graph is linked by a path containing at most $2 \log N$ edges and thus the communication time between any pair of processors is short.

More importantly, however, the shuffle-exchange computer is capable of performing a perfect shuffle on a set of data in a single parallel operation. For example, consider a deck of 8 cards distributed among the 8 processors of the 8-node shuffle-exchange graph so that processor 000 initially has card 0, processor 001 initially has card 1, processor 010 initially has card 2, and so forth. Next, consider a (parallel) operation of the shuffle-exchange computer in which each processor $a_2a_1a_0$ sends its card across a shuffle edge to the neighboring processor $a_1a_0a_2$. It is easily verified that, after completion of the operation, processor 000 contains card 0 (the top card in the shuffled deck), processor 001 contains card 4 (the second card in the shuffled deck), and so forth.

The power of card shuffling and its mathematical abstractions is well known to magicians and mathematicians [DGK81] as well as to computer scientists [S71, S80]. For a good survey of the computational power of the shuffle-exchange graph, we recommend Schwartz' paper on ultracomputers [S80]. In addition, Stone's paper [S71] contains a nice description of some important parallel algorithms based on the shuffle-exchange graph.

2b. The grid model. Among the many mathematical models that have been proposed for VLSI computation, the most widely accepted is due to Thompson and is known as the *Thompson grid model* [T79], [T80]. The grid model of a VLSI chip is quite simple. The chip is presumed to consist of a grid of vertical and horizontal *tracks* which are spaced apart by unit intervals. Processors are viewed as points and are located only at the intersection of grid tracks. Wires are routed through the tracks in order to connect pairs of processors. Although a wire in a horizontal track is allowed to cross a wire in a vertical track (without making an electrical connection), pairs of wires are not allowed to overlap for any distance or to overlap at corners (i.e., they cannot overlap in the same track). Further, wires are not allowed to overlap processors to which they are not linked. (The routing of wires in this fashion is also known as *layer per direction routing* and *Manhattan routing.*)

As an example, we have included a grid layout for the 8-node shuffle-exchange graph in Fig. 2. As before, the shuffle edges are drawn with solid lines while the exchange edges are drawn with dashed lines. Notice that we have omitted the self-loops in Fig. 2 since they are electrically redundant. In general, the processors need not all be placed on a single horizontal line (as they are in this example).

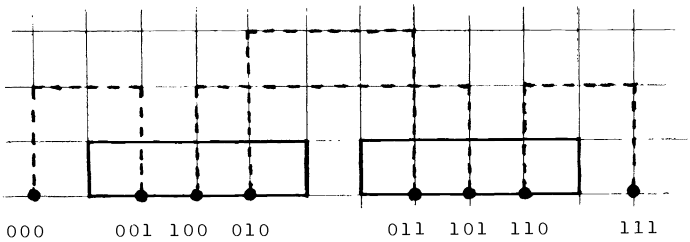


FIG. 2. *A grid model layout of the* 8-*node shuffle-exchange graph.*

Practical considerations dictate that the area of a VLSI layout be as small as possible. The *area of a layout* in the grid model is defined to be the product of the number of horizontal tracks and the number of vertical tracks which contain a processor or wire segment of the layout. For example, the layout in Fig. 2 has area 48. As can be easily observed, this is far from optimal.

2c. Thompson's layout. Given any k-bit string w, define the (Hamming) *weight* of w to be the number of 1-bits it contains. For example, the weight of 10110 is 3. Thompson's idea was to lay out the $N=2^k$ nodes of the shuffle-exchange graph on a straight line in order of nondecreasing weight. It is easily seen that shuffle edges link nodes which have the same weight and that exchange edges link nodes which have weights differing by one. Thus the edges of such a layout are relatively short. In fact, nodes connected by shuffle edges can be placed in a group, so that only 2 horizontal tracks are used for all the shuffle connections. The remaining horizontal tracks are occupied by exchange edges.

The exchange eges are inserted from left to right so that each exchange edge occupies two vertical tracks and a portion of the lowest horizontal track which is empty at the time of its insertion. (For example, Fig. 2 displays a layout for the 8-node shuffle-exchange graph designed in this way.) This well-known strategy for inserting exchange edges guarantees that the number of horizontal tracks used will be minimal, and equal to the maximum number of edges which must (at some fixed point) overlap one another. Since exchange edges link nodes which differ in weight by one, it is easily seen that the maximum overlap is at most $O(\max_{0\leqq s\leqq k} B_s)$ where B_s is the number of nodes of weight s.

It is easy to show that $B_s = C(k,s)$ for each s, where

$$C(k,s)=k!/[s!(k-s)!]$$

is the well-known function for binomial coefficients. It is also well known that $C(k,s)$ achieves its maximum value at $s=k/2$ for any k. Using standard asymptotic analysis, it is easily shown that $C(k,k/2)\sim(2/\pi)^{1/2}(2^k/k^{1/2})$ for large k. (For a good review of such techniques, see Bender and Orszag's book [BO78].) Thus Thompson's layout requires only $O(N/\log^{1/2} N)$ horizontal tracks. Since only 1 or 2 vertical tracks are needed to embed the vertical portions of the edges incident to any given node, we can conclude that Thompson's layout has area $O(N^2/\log^{1/2} N)$.

3. The complex plane diagram. In [HL80], Hoey and Leiserson observed that there is a very natural embedding of the shuffle-exchange graph in the complex plane. In what follows, we describe this embedding (which we call the *complex plane diagram*) and point out some of its more important properties.

3a. Definition. Let $\delta_k = e^{2\pi i/k}$ denote the kth primitive root of unity. Given any k-bit binary string $w=a_{k-1}\cdots a_0$, let $p(w)$ be the map which sends w to the point

$$p(w)=a_{k-1}\delta_k^{k-1}+\cdots+a_1\delta_k+a_0$$

in the complex plane. As each node of the $(N=2^k)$-node shuffle-exchange graph corresponds to a k-bit binary string, it is possible to use the map to embed the shuffle-exchange graph in the complex plane. For example, we have done this for the 32-node shuffle-exchange graph (whence $k=5$) in Fig. 3. For simplicity, each node is labeled with its value instead of its 5-bit binary string. (By the *value* of a node, we mean the numerical value of the associated k-bit binary string.)

3b. Properties. Examination of Fig. 3 indicates that the complex plane diagram has some very interesting properties. First, it is apparent that the shuffle edges occur in cycles (which we call *necklaces*) which are symmetrically placed about the origin.

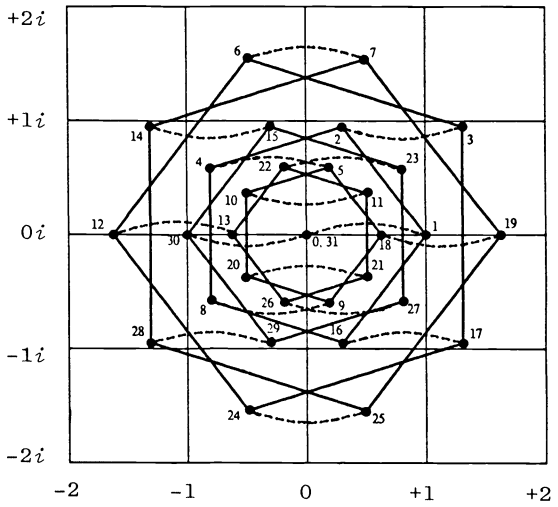


FIG. 3. *The complex plane diagram for the* 32-*node shuffle-exchange graph* (*taken from* [HL80]).

This phenomenon is easily explained by the following identity:

$$\begin{aligned}\delta_k p(a_{k-1}\cdots a_0) &= a_{k-1}\delta_k^k + a_{k-2}\delta_k^{k-1} + \cdots + a_1\delta_k^2 + a_0\delta_k\\ &= a_{k-2}\delta_k^{k-1} + \cdots + a_0\delta_k + a_{k-1}\\ &= p(a_{k-2}\cdots a_0 a_{k-1}).\end{aligned}$$

Thus traversal of a shuffle edge corresponds to a $2\pi/k$ rotation in the complex plane.

Except for degenerate cases, the preceding identity also indicates that each necklace is composed of k nodes, each a cyclic shift of the other. (Two nodes which are cyclic shifts of each other are also known as *conjugates*.) Such necklaces are called *full necklaces. Degenerate necklaces* contain fewer than k nodes and, because they must have some symmetry, are mapped entirely to the origin of the complex plane diagram. For example, {00000} and {0101, 1010} are degenerate necklaces while both {101, 011, 110} and {11100, 11001, 10011, 00111, 01110} are full. As we note in the following proposition, the number of degenerate necklaces is quite small compared to the number of full necklaces.

PROPOSITION 1. *There are* $O(N^{1/2})$ *degenerate necklaces and* $N/\log N - O(N^{1/2}/\log N)$ *full necklaces in the N-node shuffle-exchange graph.*

Proof. A node w is in a denerate necklace if its binary representation has a nontrivial symmetry with respect to cyclic shifts. Without loss of generality, such a string of bits must consist of a block of k/p bits which is repeated p times where p is some prime divisor of k. As there are $2^{k/p}$ binary strings of length k/p, this means that the number of nodes in degenerate necklaces is at most

$$\sum_{p\geqq 2}^{p|k} 2^{k/p} \leqq O(N^{1/2}).$$

The remaining $N-O(N^{1/2})$ nodes are in full necklaces. As each full necklace contains $\log N$ nodes, there are $N/\log N - O(N^{1/2}/\log N)$ full necklaces. □

It will often be convenient to refer to a necklace by one of its nodes. In particular, we will use the notation $\langle w\rangle$ to indicate the *necklace generated by w*. This is simply the collection of cyclic shifts of w. For example, the necklace generated by 101 is $\langle 101\rangle = \{101, 011, 110\}$.

Exchange edges are also embedded in a very regular fashion by the complex plane diagram. In fact, each exchange edge is embedded as a horizontal line segment of unit length. This phenomenon is explained by the identity

$$p(a_{k-1}\cdots a_1 0)+1 = a_{k-1}\delta_k^{k-1}+\cdots+a_1\delta_k+1 = p(a_{k-1}\cdots a_1 1).$$

In some cases, several exchange edges are contained in the same horizontal line of the diagram. Such lines are called *levels*. For example, there are 9 levels in the diagram of the 32-node shuffle-exchange graph shown in Fig. 3. We will use the properties of levels to find $O(N^2/\log^{3/2} N)$-area layouts for the N-node shuffle-exchange graph.

4. Layouts based on the complex plane diagram. In this section, we present several layouts of the shuffle-exchange graph which are based on the complex plane diagram. We commence with a straightforward $O(N^2/\log N)$-area layout of the N-node shuffle-exchange graph. This layout has been discovered by many researchers (including Hoey and Leierson). Later, we show how the layout can be modified so as to require only $O(N^2/\log^{3/2} N)$ area.

4a. A straightforward $O(N^2/\log N)$-area layout. In what follows, we describe a straightforward layout of the shuffle-exchange graph which requires only $O(N^2/\log N)$ area. The layout is formed from a grid of levels and necklaces which we call the *level-necklace grid*. Each row of the grid corresponds to a level of the complex plane diagram. The columns of the grid are divided into consecutive column pairs, each pair corresponding to a necklace. The leftmost column of each column pair corresponds to that part of the necklace which is contained in the left half of the complex plane. Similarly, the rightmost column of each pair corresponds to the part of the necklace contained in the right half of the complex plane.

The rows of the level-necklace grid must have the same top-to-bottom order as do the corresponding levels in the complex plane diagram. The columns, however, may be arranged arbitrarily (provided that columns corresponding to the same necklace are adjacent in the grid).

Each node of the shuffle-exchange graph is placed at the intersection of the row and column of the grid that corresponds to the level and part of the necklace (left half or right half) to which it belongs in the complex plane diagram. For example, we have done this for a random ordering of the necklaces of the 32-node shuffle-exchange graph in Fig. 4. (Notice that we have used just one column each for the degenerate necklaces $\langle 0\rangle$ and $\langle 31\rangle$ since they each contain just one node. In general two columns will be required for necklaces which are mapped to the origin of the complex plane diagram, but the nodes of each such necklace should still be lumped together at a single point of the level-necklace grid.)

Given a level-necklace grid for a shuffle-exchange graph, it is not difficult to produce a layout for the graph. The main step is to partition the exchange edges in each row of the grid into nonoverlapping subsets. Each subset can then be assigned to a horizontal track of the layout. Except for the row corresponding to the real line in the complex plane diagram, the assignment of subsets to horizontal tracks within a row is arbitrary. (The assignment of horizontal tracks containing nodes on the real line must preserve the cyclic orientation of the nodes which are in necklaces that are mapped to the origin.)

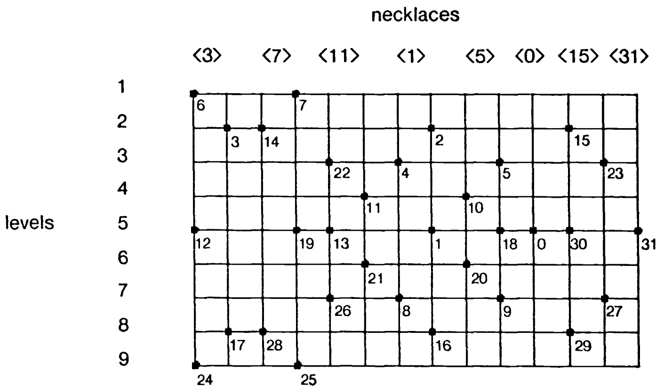


FIG. 4. *A level-necklace grid for the* 32-*node shuffle-exchange graph.*

Once this is done, the exchange edges can be inserted in the horizontal tracks and the shuffle edges can be inserted in the vertical tracks. (To be precise, some of the shuffle edges also occupy part of a horizontal track at the top or bottom of the layout.) By Proposition 1, the number of vertical tracks occupied by the necklaces is at most $2N/\log N+O(N^{1/2})$. Since there are precisely $N/2$ exchange edges, at most $N/2+2$ horizontal tracks are contained in the layout. Thus the total area of the layout of the N-node shuffle-exchange graph is at most $N^2/\log N+O(N^{3/2})$. As an example, we have displayed in Fig. 5 a layout of the 32-node shuffle-exchange graph produced from the level-necklace grid in Fig. 4.

4b. An improved $O(N^2/\log^{3/2} N)$-area layout. It is possible to improve the layout described in § 4a by reducing the number of horizontal tracks needed to embed the

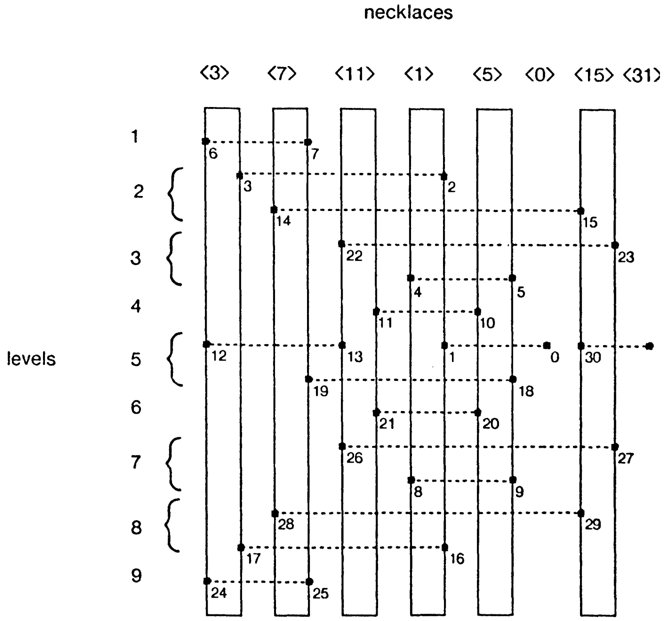


FIG 5. *Layout of the* 32-*node shuffle-exchange graph produced from the level-necklace grid shown in Fig.* 4.

exchange edges. This can be done by reordering the necklaces from left to right so as to increase the average number of exchange edges which can be inserted on each horizontal track. For example, the ordering of the necklaces shown in Fig. 6 results in far fewer horizontal tracks being used than did the ordering of necklaces shown in Fig. 5.

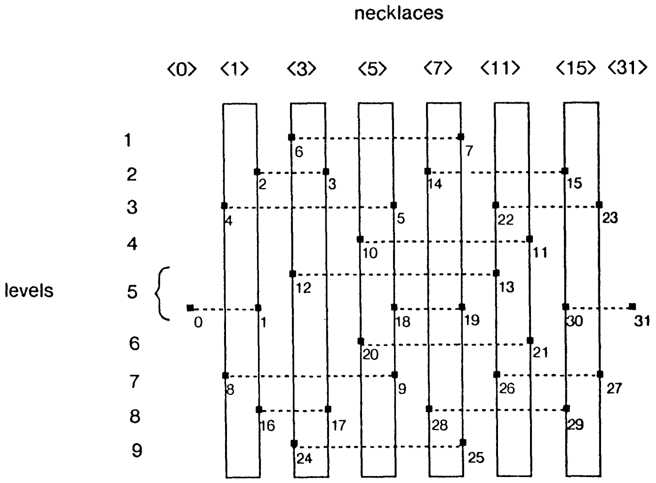


FIG. 6. *An improved layout for the* 32-*node shuffle-exchange graph.*

Although we do not know how to best order the necklaces in general, we have found several orderings which yield $O(N^2/\log^{3/2} N)$-area layouts for the N-node shuffle-exchange graph. For instance, we will show in what follows that such a layout can be constructed by arranging the necklaces from left to right in order of nondecreasing weight. (The *weight* of a necklace is simply defined to be the weight of any of its nodes.) As an example, the layout displayed in Fig. 6 is of this form. (This observation has also been made by Steinberg and Rodeh in [SR81].)

In order to bound the number of horizontal tracks needed to insert the exchange edges, we will show that the maximum overlap of exchange edges *on each level* is at most the number of nodes of size $h = \lfloor (k-1)/2 \rfloor$ on that level. Since the maximum overlap of exchange edges on each level is an upper bound on the number of horizontal tracks needed to insert the exchange edges on that level, we can thus conclude that the total number of horizontal tracks needed to insert all of the exchange edges is at most

$$B_h \leqq B_{k/2} = (2/\pi)^{1/2} N/\log^{1/2} N + O(N/\log^{3/2} N) \text{ where } N = 2^k.$$

Thus the resulting layout will have area at most

$$2(2/\pi)^{1/2} N^2/\log^{3/2} N + O(N^2/\log^{5/2} N).$$

Although it is clear the maximum *total* overlap (over all levels) of exchange edges is at most $B_{k/2}$, this is not sufficient to prove the result since any layout must also preserve the top-to-bottom partial order induced by the necklace structure on the exchange edges. It is only within individual levels that the top-to-bottom ordering of exchange edges is arbitrary. (As we noted earlier, some minor precautions are necessary for the level corresponding to the real line.) It is *not* immediately clear, however, why the maximum overlap on each level is at most the number of nodes of size $h \leqq k/2$ on that level. In what follows, we establish this result by breaking up each level into sublevels (for which the analysis is easier) and showing that the maximum overlap on each sublevel is at most the number of nodes of size h on that sublevel. The analysis requires some additional notation.

Consider a node of the form $a_{k-1} \cdots a_1 0$ for which either $a_{k-i} = 0$ or $a_i = 0$ or both for each $i \leqq k$. We will refer to such a node as a *basis node.* A node $b_{k-1} \cdots b_0$ is said to be *generated* by the basis node $a_{k-1} \cdots a_0$ if

1) $b_{k-i} = a_{k-i}$ and $b_i = a_i$ whenever $a_{k-i} \neq a_i$ for $1 \leqq i \leqq k-1$, and
2) $b_{k-i} = b_i$ whenever $a_{k-i} = a_i = 0$ for $1 \leqq i \leqq k-1$.

For example, 10000 generates 10001, 11100 and 11101 but not 11111.

It is not difficult to show that if u generates v, then both u and v are on the same level of the complex plane diagram. For example, let $u = a_{k-1} \cdots a_0$ and $v = b_{k-1} \cdots b_0$ and observe that

$$\begin{aligned} p(v) - p(u) &= (b_{k-1} - a_{k-1})\delta_k^{k-1} + \cdots + (b_1 - a_1)\delta_k + (b_0 - a_0) \\ &= c_{k-1}\delta_k^{k-1} + \cdots + c_1\delta_k + c_0, \end{aligned}$$

where $c_{k-i} = c_i$ for each i, $1 \leqq i \leqq k-1$. Since δ_k^{k-i} is the complex conjugate of δ_k^i for $1 \leqq i \leqq k-1$, we can conclude that $p(v) - p(u)$ is a real number and thus that u and v are in the same level of the complex plane diagram.

It is also easy to show that each node of the shuffle-exchange graph is generated by a unique basis node. In particular, the node which generates $b_{k-1} \cdots b_0$ can be found by

1) setting $b_0 = 0$ and (if k is even) setting $b_{k/2} = 0$, and
2) setting $b_i = b_{k-i} = 0$ for each i such that (originally) $b_i = b_{k-i} = 1$.

Since exchange edges link nodes which have the same basis node, we can conclude from the preceding arguments that it is possible to partition each level of the complex plane diagram into *sublevels* so that the nodes in each sublevel are precisely the nodes generated by some basis node. We will now show that the maximum overlap on each sublevel is at most the number of nodes of weight h on that sublevel.

Since the necklaces have been arranged from left to right in order of nondecreasing weight, the overlap of exchange edges between two nodes of weight s in any sublevel is at most $O(\max_{0 \leqq s \leqq k} B_s^*)$ where B_s^* is the number of nodes in that sublevel with weight s. In the following proposition, we compute B_s^* and show that its maximum for any sublevel occurs at $s = h$.

PROPOSITION 2. *Each basis node of weight r generates B_s^* nodes of weight s, where*

1) $B_s^* = C(h-r, i)$ *for* $s = r + 2i$ *and* $i \leqq h - r$, *and*
2) $B_s^* = C(h-r, i)$ *for* $s = r + 2i + 1$ *and* $i \leqq h - r$

when k is odd, and

1) $B_s^* = C(h-r+1, i)$ *for* $s = r + 2i$ *and* $i \leqq h - r + 1$, *and*
2) $B_s^* = 2C(h-r, i)$ *for* $s = r + 2i + 1$ *and* $i \leqq h - r$

when k is even.

Proof. When k is odd, there are precisely $h - r$ pairs $a_j = a_{k-j} = 0$ in a basis node of weight r. In order to generate a string of weight $s = r + 2i$ when k is odd, we must set $b_0 = 0$ and set i of the $h - r$ pairs so that $b_j = b_{k-j} = 1$. There are $C(h-r, i)$ such strings. To generate a string of weight $s = r + 2i + 1$ when k is odd, we must set $b_0 = 1$ and choose i of the $h - r$ pairs so that $b_j = b_{k-j} = 1$. As before, there are $C(h-r, i)$ such strings.

When k is even, there is also the degenerate pair $a_{k/2}=0$. To generate a string of weight $s=r+2i$ when k is even, we must choose i of the $h-r+1$ pairs so that $b_j=b_{k-j}=1$ (this count includes the "pair" $b_0=b_{k/2}=1$). There are $C(h-r+1, i)$ such strings. To generate a string of weight $s=r+2i+1$ when k is even, we must set either $b_0=1$ and $b_{k/2}=0$ or $b_0=0$ and $b_{k/2}=1$, and choose i of the $h-r$ pairs so that $b_j=b_{k-j}=1$ $(j\neq k/2)$. There are $2C(h-r, i)$ such strings. □

Given Proposition 2, it is easily checked that the maximum value of B_s^* for any sublevel (independent of the value of r) occurs when $s=h$. Thus the sum (over all sublevels) of the maximum overlap at each sublevel is at most the number of nodes of weight $h=\lfloor (k-1)/2\rfloor$ in the entire graph. This is at most $C(k, k/2)\sim (2/\pi)^{1/2}(2^k/k^{1/2})$. Thus the total area of the layout is no more than

$$2(2/\pi)^{1/2}N^2/\log^{3/2} N+O(N^2/\log^{5/2} N),$$

as claimed.

4c. Additional $O(N^2/\log^{3/2} N)$-area layouts. By varying the order of the necklaces in the level-necklace grid, it is possible to produce a variety of layouts for the shuffle-exchange graph which require at most $O(N^2/\log^{3/2} N)$ area. The complex plane diagram itself suggests one such ordering. For example, consider an arrangement of the necklaces from left to right in order of nondecreasing radius. (The *radius* of a necklace is defined to be the distance of its nodes from the origin in the complex plane diagram.) Such a layout corresponds to a folding of the complex plane diagram along its imaginary axis followed by a straightening of the necklaces. In what follows, we will show that, like a layout by necklace weight, a layout by necklace radius has area $O(N^2/\log^{3/2} N)$.

Because the layout by radius is so closely related to the complex plane diagram, our analysis will center on the complex plane diagram, itself. As before, we will partition the levels into sublevels and find an upper bound on the maximum overlap of exchange edges on each sublevel separately. The number of horizontal tracks needed to insert the exchange edges will then be at most the sum of these upper bounds. We will show that this sum is at most $O(N/\log^{1/2} N)$.

Notice that the maximum overlap of exchange edges on a sublevel of the level-necklace grid is at most twice the maximum overlap on that sublevel in the complex plane diagram. (The factor of two is introduced by the "folding" of the diagram along its imaginary axis. Although straightening the necklaces might affect the maximum *total* overlap of exchange edges, it does *not* affect the overlap *within* a sublevel.)

Within a sublevel, an exchange edge can be identified by the real part of its midpoint. For example, the real part of the midpoint of exchange edge $(b_{k-1}\cdots b_1 0, b_{k-1}\cdots b_1 1)$ is

$$b_{k-1}\cos[2\pi(k-1)/k]+\cdots+b_1\cos[2\pi/k]+\tfrac{1}{2}.$$

If a is a basis node of a sublevel, then a generates the other nodes in that sublevel by substitution of the appropriate pairs of ones. For instance, we may set $b_i=b_{k-i}=1$, if $a_i=a_{k-i}=0$. Let

$$T_a=\{1\leqq j\leqq h \mid a_j=a_{k-j}=0\}$$

denote those indices $1\leqq i\leqq h$ where a pair of 1-bits may be substituted for a pair of 0-bits. (As before, $h=\lfloor (k-1)/2\rfloor$ but for convenience, we shall henceforth assume that k is odd.) Notice that if b is generated by a, then the real part of the midpoint of the exchange edge incident to b is

$$\sum^{i\in T_a} 2b_i \cos(2\pi i/k) + \sum_{1\leqq i\leqq h}^{i\notin T_a} \cos(2\pi i/k) + \tfrac{1}{2}.$$

We now introduce a random variable Z_a, which has as its image, all of the real parts of the midpoints of edges in the sublevel generated by a. Since $b_i = b_{k-i}$ can be either 0 or 1 when $i \in T_a$, let B_i be a random variable representing this choice. In particular,

$$B_i = 0 \quad \text{with probability } \tfrac{1}{2}, \quad \text{and}$$

$$B_i = 1 \quad \text{with probability } \tfrac{1}{2}.$$

Then

$$\begin{aligned} Z_a &= \sum^{i\in T_a} 2\cos(2\pi i/k)B_i + \sum_{1\leqq i\leqq h}^{i\notin T_a} \cos(2\pi i/k) + \tfrac{1}{2} \\ &= \sum^{i\in T_a} 2\cos(2\pi i/k)(B_i - \tfrac{1}{2}). \end{aligned}$$

Since the exchange edges have unit length in the complex plane diagram, two edges overlap if and only if their midpoints are within unit distance of each other. Thus the number of edges which overlap at position x on the sublevel generated by a node a is given by the formula

$$2^{|T_a|} \operatorname{Prob}[x - \tfrac{1}{2} \leqq Z_a \leqq x + \tfrac{1}{2}],$$

where $|T_a|$ denotes the cardinality of T_a. (We caution the reader that the notation $|x|$ is also used to denote the *absolute value* of x.)

Although the distribution function of Z_a is difficult to analyze directly, it does behave like a normal distribution. This is because Z_a is the sum of independent random variables which have mean 0 and variance $\sigma_i^2 = \cos^2(2\pi i/k)$. The Berry–Esseen theorem states precisely how far Z_a can vary from a normal distribution. (For a proof of this theorem see [F71].)

BERRY–ESSEEN THEOREM. *Let $X_1, X_2, \cdots, X_m$ be independent random variables such that* $\mathrm{E}(X_i) = 0$, $\mathrm{E}(X_i^2) = \sigma_i^2$, *and* $\mathrm{E}(|X_i^3|) = \rho_i$ *for* $1 \leqq i \leqq m$. *Set* $s^2 = \sigma_1^2 + \cdots + \sigma_m^2$ *and* $r = \rho_1 + \cdots + \rho_m$. *In addition, let F denote the cumulative distribution function of the sum* $(X_1 + \cdots + X_m)/s$. *Then for all x,*

$$|F(x) - \Phi(x)| \leqq 6r/s^3$$

where Φ is the standard normal cumulative distribution function.

In the case of a sublevel generated by a node a, we have

$$X_i = 2\cos(2\pi i/k)(B_i - \tfrac{1}{2}) \text{ for } i \in T_a,$$

$$s_a^2 = \sum^{i\in T_a} \cos^2(2\pi i/k),$$

$$r_a = \sum^{i\in T_a} |\cos^3(2\pi i/k)|.$$

Applying the Berry–Esseen theorem, we can thus conclude that

$$\begin{aligned} \operatorname{Prob}[x - \tfrac{1}{2} \leqq Z_a \leqq x + \tfrac{1}{2}] &= \operatorname{Prob}[(x - \tfrac{1}{2})/s_a \leqq Z_a/s_a \leqq (x + \tfrac{1}{2})/s_a] \\ &\leqq \Phi[(x + \tfrac{1}{2})/s_a] - \Phi[(x - \tfrac{1}{2})/s_a] + 12r_a/s_a^3. \end{aligned}$$

Because the standard normal density function is symmetric and unimodal, we can conclude that the maximum of Prob $[x-\frac{1}{2}\leqq Z_a\leqq x+\frac{1}{2}]$ occurs at $x=0$ and is at most $O(1/s_a+r_a/s_a^3)$.

In the following proposition, we find bounds for the values of r_a and s_a.

PROPOSITION 3. *For any basis node a*

$$r_a=\sum^{i\in T_a}|\cos^3(2\pi i/k)|\leqq|T_a|,$$

$$s_a^2=\sum^{i\in T_a}\cos^2(2\pi i/k)\geqq\Omega(|T_a|^3/k^2).$$

Proof. The bound on r_a is easy to compute since $|\cos^3(2\pi i/k)|\leqq 1$. The calculation of s_a is a bit more tedious. In order to obtain a lower bound, $\cos^2(2\pi i/k)$ must be made as small as possible. The smallest values occur when T_a contains indices i which are as close to $(k-1)/4$ as possible. In this case, we can approximate $\cos^2(2\pi i/k)$ with the value $c(\pi/2-2\pi i/k)^2$, for some constant c. Direct computation reveals that the sum of these squares is at least $\Omega(|T_a|^3/k^2)$. □

Since $|T_a|<k$ for all a, we can conclude from the preceding that the maximum overlap of exchange edges on a sublevel generated by a is at most

$$O(2^{|T_a|}k^3/|T_a|^{7/2}).$$

Noting that there are precisely $C(h,j)2^{h-j}$ sublevels generated by a node for which $|T_a|=j$ and summing, we can conclude that the total number of horizontal tracks needed to insert all of the exchange edges is at most

$$\sum_{j=1}^{h}C(h,j)2^{h-j}O(2^jk^3/j^{7/2})=O\left[k^32^h\sum_{j=1}^{h}C(h,j)/j^{7/2}\right].$$

It is not difficult to check that the dominant terms in the preceding sum occur when $j=h/2\pm\Theta(h^{1/2}\log h)$. In this region, $j=\Theta(k)$ and thus the sum is bounded above by

$$O\left[2^hk^{-1/2}\sum_{j=1}^{h}C(h,j)\right]=O(2^{k-1}/k^{1/2})=O(N/\log^{1/2}N),$$

thus completing the proof that a layout by necklace radius takes at most $O(N^2/\log^{3/2}N)$ area.

5. Remarks. It is worth remarking that the $O(N^2/\log^{3/2}N)$-area layouts for the shuffle-exchange graph described in § 4 actually require $\Omega(N^2/\log^{3/2}N)$ area and thus our analysis of these layouts cannot be improved by more than a constant factor. In each case, the lower bound on area can be derived from the fact that the maximum *total* overlap of exchange edges in the layouts is at least $\Omega(N/\log^{1/2}N)$. (Remember that although the maximum *total* overlap of exchange edges is *not* an *upper* bound on the number of horizontal tracks needed to insert the exchange edges, it *is* a lower bound.)

The $\Omega(N/\log^{1/2}N)$ lower bound on maximum overlap is easily established for the layout according to necklace weight since $\Omega(N/\log^{1/2}N)$ exchange edges link nodes of weight $k/2$ to nodes of weight $k/2+1$. The lower bound on maximum overlap is somewhat more difficult to prove for the layout according to necklace radius. The first step in the proof is to show that at least $N/2$ exchange edges are contained within a square of side length $ck^{1/2}$ centered at the origin of the complex plane diagram (where c is a constant). (This can be done by using the techniques developed in § 4c). Next consider the sum (over i) of the total overlaps at points corresponding to radii of $i/2$ for $1 \leqq i \leqq ck^{1/2}$. Because the complex plane diagram is radially symmetric, it is possible to show that at least $\Omega(N)$ exchange edges are counted in this sum. Thus the overlap at one of these points must be at least $\Omega(N/k^{1/2}) = \Omega(N/\log^{1/2} N)$, as claimed.

Since Thompson [T80] has shown that any layout for the N-node shuffle-exchange graph must have area at least $\Omega(N^2/\log^2 N)$, we know that at least $\Omega(N/\log N)$ horizontal tracks are needed to insert the exchange edges for any ordering of necklaces in the level-necklace grid. However, there is no ordering of the necklaces known for which the exchange edges can be inserted using less than $o(N/\log^{1/2} N)$ horizontal tracks. This suggests an interesting open question since it would be nice to find an $O(N^2/\log^2 N)$-area layout based on the complex plane diagram. (Although an asymptotically optimal $O(N^2/\log^2 N)$-area layout for the shuffle-exchange graph has recently been found by Kleitman, Leighton, Lepley and Miller [KLLM81], it is rather complicated and of limited practical use.)

Although we do not know of necklace orderings for which the exchange edges can be inserted using less than $o(N/\log^{1/2} N)$ horizontal tracks, we *do* know of orderings for which the *maximum total overlap* of exchange edges is at most $O(N \log \log N/\log N)$. For example, an ordering of the neckalces by minimum value has a maximum total overlap of $\Theta(N \log \log N/\log N)$. (The *minimum value* of a necklace is simply the minimum of the values of the nodes contained in the necklace.)

Interestingly, an analysis of the minimum (over all orderings) of the maximum total overlap for small values of N indicates that there may always be an ordering for which the maximum total overlap is at most $O(N/\log N)$, the least possible. In fact, for $3 \leqq N \leqq 7$, this minimum maximum overlap is precisely $\lfloor (2^k - 2)/k \rfloor$. A summary of the minimum maximum overlap data for small values of N is included in Table 1.

TABLE 1
Maximum overlap of best known orderings

k	N	maximum overlap of best known ordering	optimal?
3	8	2	yes
4	16	3	yes
5	32	6	yes
6	64	10	yes
7	128	18	yes
8	256	33	yes
9	512	62	?
10	1024	115	?
11	2048	214	?
12	4096	388	?
13	8192	754	?

In addition to varying the order of the necklaces, improvements in the layout may also be made by rearranging the level assignments of the exchange edges. For example, the layout of the 32-node shuffle-exchange graph shown in Fig. 7 was constructed in this way. (The careful reader will notice that we have also manipulated the necklaces somewhat in order to produce this layout.) For a more detailed discussion of the manner in which exchange edges can be reassigned, we refer the reader to [LM81]. (Such layouts have also been used in conjunction with the Blue Chip Project at Purdue [S81].)

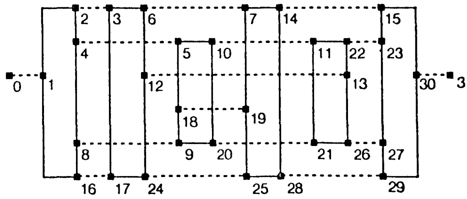


FIG. 7. *An improved layout for the* 32-*node shuffle-exchange graph.*

6. Acknowledgments. In acknowledgment, we would like to thank the following people for their helpful remarks and suggestions: Herman Chernoff, Peter Elias, Dan Hoey, Dan Kleitman, Charles Leiserson, Ron Rivest, Michael Rodeh, Larry Snyder, and Richard Zippel.

REFERENCES

[BO78] C. M. BENDER AND S. A. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[DGK81] P. DIACONIS, R. L. GRAHAM AND W. M. KANTOR, *The mathematics of perfect shuffles*, preprint, 1981.

[F71] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Volume II, second ed., John Wiley, New York, 1971.

[HL80] D. HOEY AND C. E. LEISERSON, *A layout for the shuffle-exchange network*, Proc. 1980 IEEE International Conference on Parallel Processing, August 1980.

[KLLM81] D. KLEITMAN, F. T. LEIGHTON, M. LEPLEY AND G. L. MILLER, *New layouts for the shuffle-exchange graph*, Proc. 13th Annual ACM Symposium on Theory of Computing, May 1981, pp. 278–292.

[L81] F. T. LEIGHTON, *Layouts for the shuffle-exchange graph and lower bound techniques for* VLSI, Ph.D. thesis, Mathematics Dept., Massachusetts Institute of Technology, Cambridge, September 1981.

[LM81] F. T. LEIGHTON AND G. L. MILLER, *Optimal layouts for small shuffle-exchange graphs*, VLSI 81 – Very Large Scale Integration, John P. Gray, ed., Academic Press, London, 1981, pp. 289–299.

[P80] D. S. PARKER, *Notes on shuffle/exchange-type switching networks*, IEEE Trans. Comput., C-29 (1980), pp. 213–222.

[S71] H. S. STONE, *Parallel processing with the perfect shuffle*, IEEE Trans. Comput., C-20 (1971), pp. 153–161.

[S80] J. T. SCHWARTZ, *Ultracomputers*, ACM Trans. Programming Languages and Systems, 2 (1980), pp. 484–521.

[S81] L. SNYDER, *Overview of the* CHiP *computer*, VLSI 81 – Very Large Scale Integration, J. Gray, ed., Academic Press, London, 1981, pp. 237–246.

[SR81] D. STEINBERG AND M. RODEH, *A layout for the shuffle-exchange network with* $O(N^2/\log^{3/2} N)$ *area*, IEEE Trans. Comput., C-30 (1981), pp. 977–982.

[T79] C. D. THOMPSON, *Area-time complexity for* VLSI, Proc. 11th Annual ACM Symposium on Theory of Computing, May 1979, pp. 81–88.

[T80] ———, *A complexity theory for* VLSI, Ph.D. dissertation, Dept. Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1980.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

INVERTING SIGNED GRAPHS*

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

Abstract. This paper addresses the question of determining the class of rectangular matrices having a given signed graph as a signed row or column graph. We also determine equivalent conditions on a given pair of signed graphs in order for them to be the signed row and column graphs of some rectangular matrix. In connection with these signed graph inversion problems we discuss the concept of minimality and illustrate how to invert a pair of signed graphs.

1. Introduction. In this paper we continue the systematic investigation of the structural relationships between rectangular matrices and graphs, digraphs and signed graphs started in [4], [5], and [6]. To study these relationships we make use of the following graphs.

Given an $m \times n$ matrix A, we define two sets of points $R = \{r_1, \cdots, r_m\}$ and $C = \{c_1, \cdots, c_n\}$ to represent the rows and columns of A, respectively. The three basic graphs are:

Fundamental bigraph. BG is a bipartite graph (bigraph) on R and C. The lines correspond to the nonzeros of A; i.e., $[r_i, c_j]$ is a line in BG if and only if $a_{ij} \neq 0$.

Row graph. RG has point set R. The line $[r_i, r_k]$ belongs to RG if there exists $c_j \in C$ such that $[r_i, c_j]$ and $[r_k, c_j]$ are lines of BG.

Column graph. CG has point set C. The line $[c_j, c_k]$ belongs to CG if there exists $r_i \in R$ such that $[c_j, r_i]$ and $[c_k, r_i]$ are lines of BG.

This leads naturally to the questions of determining the class of rectangular matrices having a given graph as a row or column graph and determining equivalent conditions on a given pair of graphs in order for them to be the row and column graphs of some rectangular matrix (see [7]). These graph inversion techniques are useful in characterizing the two-step graphs studied by Exoo and Harary [2] (see [8]) and in characterizing the competition graphs studied by Roberts [16], [17] (see [14]). In this paper we turn our attention to these same problems for signed graphs. First, we consider how the sign information in the matrix can be incorporated into the three graphs.

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 BG^+. The signed structure of A, i.e., the locations of the positive and negative entries of A, is immediately discernible from the signed bigraph BG^+. Thus, given a signed bigraph G^+, we can construct a unique matrix A with entries $+1$, -1 or 0 such that $BG^+(A) = G^+$.

Now it is not always possible to form signed row or column graphs. To form $RG^+(A)$, it is necessary that the scalar product of any two rows be positive, negative or zero independently of the magnitudes of the elements; i.e., all terms in the scalar product are weakly of the same sign. We can then form RG^+ where the line $[r_i, r_j]$ is positive if the corresponding row vectors have a positive scalar product, and negative if the scalar product is negative. CG^+ is defined in a similar way. In [6] it was shown that RG^+ can be formed if and only if CG^+ can be formed, and if so, we say that A is signed. Applications of these signed graphs and the importance of when they can be formed are discussed in Greenberg [3], Greenberg, Lundgren and Maybee [6],

* Received by the editors June 15, 1982, and in revised form May 3, 1983.

† Energy Information Administration, Washington, DC 20461.

‡ University of Colorado at Denver, Denver, Colorado 80202.

§ University of Colorado, Boulder, Colorado 80309.

Kydes and Provan [13] and Provan [15]. These include identifying and characterizing important components of energy economic models such as physical flows matrices and transportation matrices, and analyzing correlation and determinacy in linear systems related to networks. In [9] we show how the signed graph inversion method developed in this paper can be used to teach a computer to build models using partial information in the form of economic correlation.

In § 2 we find the class of matrices A satisfying $RG^+(A)=G^+$ (and $CG^+(A)=G^+$) for a given signed graph G^+. We use the methods developed in [7] as well as a theorem of Harary and Kabell [11] on marked graphs. We also discuss the notion of minimality. In § 3 we find necessary and sufficient conditions for a pair of signed graphs to be invertible and illustrate how to invert a pair of signed graphs.

2. One-graph inversion. In [7] we characterized the family of regular Boolean matrices A whose row (column) graph equals a specified graph G. (A Boolean matrix is regular if each row and column has a nonzero entry.) Here we investigate the same problem for signed graphs and regular signed Boolean matrices (entries are ± 1 or 0, and each row and column has a nonzero entry). As in [7], observe that if $RG^+(A)=G^+$, then $CG^+(A^T)=G^+$. Consequently, we shall consider only matrices A such that $RG^+(A)=G^+$.

Before considering the signed case, we review the situation for graphs. A k-clique, $k\geqq 1$, of a graph is a complete subgraph on k points. Given a graph G, a finite set S of cliques of G will be called a *clique cover* if every point and line of G belongs to at least one clique in S. We will use the notation $\langle X\rangle$ to denote the subgraph of G generated by the set of points X. The following result is [7, Thm. 1].

THEOREM 2.1. *Given the graph $G=(V,E)$ with $p=|V|$, the regular Boolean matrix A has the property that $RG(A)=G$ if and only if A has p rows and the columns of A correspond to a clique cover of G.*

To illustrate the difference between the two problems, we consider the following example. Let G and G^+ be as illustrated in Fig. 2.1, where we have followed the convention of using dashed lines to represent negative lines, as introduced in [12].

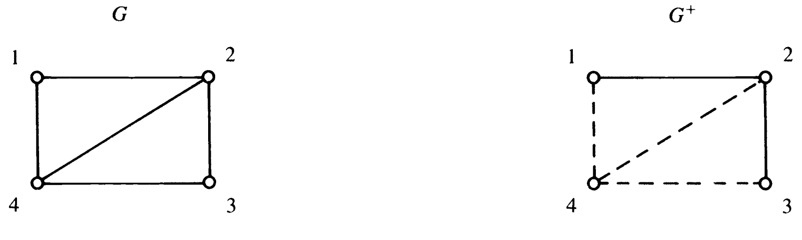


FIG. 2.1

Then $S=\{\langle 1,2,4\rangle,\langle 2,3,4\rangle\}$ is a clique cover for G. So by Theorem 2.1,

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

satisfies $RG(A)=G$. However, even though S is also a clique cover for G^+, there is no way to sign A so that $RG^+(A)=G^+$. Now, if we let $T=\{\langle 1,2,4\rangle\langle 2,3\rangle\langle 3,4\rangle\}$, then

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

also satisfies $RG(A)=G$. Here again, T is a clique cover of G^+ also, but if we sign A to get

$$A_1=\begin{bmatrix}1&0&0\\1&-1&0\\0&-1&1\\-1&0&1\end{bmatrix},$$

then $RG^+(A_1)=G^+$. The difference in the two clique covers is that each clique in T is balanced. A 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 two points in different sets is negative.

Given a signed graph G^+, a finite set S^+ of signed cliques of G^+ will be called a *balanced clique cover* of G^+ if every point and line of G^+ belongs to at least one clique in S^+, and every clique in S^+ is balanced. The next two lemmas show that every signed graph G^+ has a balanced clique cover which can be used to construct a matrix A satisfying $RG^+(A)=G^+$.

LEMMA 2.2. *Every signed graph G^+ has a balanced clique cover.*

Proof. Let S^+ be the set of all lines in G^+ together with the isolated points. Clearly S^+ is a balanced clique cover of G^+. □

LEMMA 2.3. *Let $G^+=(V,E)$ be a signed graph with $p=|V|$, $n=|E|$, and p_0 equal to the number of isolated points in G^+. Then there exists a $p\times(n+p_0)$ regular signed Boolean matrix A such that $RG^+(A)=G^+$.*

Proof. Label the points of G^+ $1,2,\cdots,p$ and the lines $1,2,\cdots,n$. We then construct A as follows. For each line of G^+ there is a corresponding column of A with 1's in rows i and j if the line $[i,j]$ is positive, and 1 in row i and -1 in row j if the line is negative and $i<j$. For each isolated point k of G^+ there is a corresponding column of A with a 1 in row k. Clearly A satisfies the conditions of the lemma. □

The columns of matrix A constructed in Lemma 2.3 corresponded to the cliques in the balanced clique cover consisting of all the lines and isolated points. However, for the graph G^+ in Fig. 2.1, we found a matrix A satisfying $RG^+(A)=G^+$ where the cliques determined by the columns of A were not all lines or points. To construct a matrix A corresponding to an arbitrary balanced clique cover of G^+, we need a method for determining the signs in each column of A. For this we use the notion of marked graphs investigated by Bieneke and Harary [1], Harary [10], and Harary and Kabell [11].

In a *marked graph*, the points are designated positive or negative. Let M be a marked graph with underlying graph $G=G(M)$ having the same points and lines as M, but without any signs on its lines or points. The signed graph of the marked graph M, written $S(M)$, is obtained from G by affixing to each line the product of the signs of its two points. The following result is [10, Thm 3.8] (also see Harary and Kabell [11]).

THEOREM 2.4. *To each marked graph M there corresponds a unique balanced signed graph $B^+ = S(M)$. To each balanced signed graph B^+ there correspond two marked graphs M and M', which are sign-reversals, such that $S(M) = S(M') = B^+$.*

Given a balanced signed graph B^+, we construct a marked graph M satisfying $S(M) = B^+$ as follows. Select an arbitrary point and mark it positive (or negative). Select a point adjacent to this point and label it with the product of the sign of the marked point and the sign of the line joining the two points. Continue in this way until the graph is marked.

We can use this procedure to see how the matrix A_1 was constructed from the graph G^+ of Fig. 2.1. Since T is a balanced clique cover of G^+, we can mark each of the cliques in T independently as follows:

$$\langle \overset{+}{1}, \overset{+}{2}, \overset{-}{4} \rangle, \qquad \langle \overset{-}{2}, \overset{-}{3} \rangle, \qquad \langle \overset{+}{3}, \overset{+}{4} \rangle.$$

A_1 is then formed using the signs for each clique. This procedure can be used for any balanced clique cover to get the following theorem.

THEOREM 2.5. *Given the graph $G^+ = (V, E)$ with $p = |V|$, the regular signed Boolean matrix A has the property that $RG^+(A) = G^+$ if and only if A has p rows and the columns of A correspond to a balanced clique cover of G^+.*

Proof. Suppose $RG^+(A) = G^+$. Then by Theorem 2.1, the columns of A correspond to a clique cover of G^+, and by [6, Lemma 3], each of these cliques is balanced.

For the converse, let A be a regular signed Boolean matrix with p rows whose columns correspond to a balanced clique cover of G^+. Observe that we can form $RG^+(A)$, since if r_i and r_j have nonzeros in columns k and q, then $[r_i, r_j]$ is in cliques C_k and C_q, and so r_i and r_j are marked with either the same signs or opposite signs in both C_k and C_q. We must show that $RG^+(A) = G^+$. Now $[r_i, r_j]$ is negative in G^+ if and only if $[r_i, r_j]$ belongs to a balanced clique, and r_i and r_j are marked with opposite signs if and only if a column of A contains nonzero entries of opposite signs in rows r_i and r_j if and only if $[r_i, r_j]$ is negative in $RG^+(A)$. Similarly, $[r_i, r_j]$ is positive in G^+ if and only if $[r_i, r_j]$ is positive in $RG^+(A)$. Hence, $RG^+(A) = G^+$, and the proof is complete. □

Given a balanced clique cover S^+ of the signed graph G^+, we can form a *signed clique cover graph*, $Q(S^+)$, as follows. Let $S^+ = \{C_1, \cdots, C_n\}$ and A be the corresponding matrix. Then $Q(S^+)$ is a signed graph on the points $1, 2, \cdots, n$, and the line $[i, j]$ is positive in $Q(S^+)$ if and only if C_i and C_j contain at least one point marked with the same sign, and negative if and only if C_i and C_j contain at least one point marked with opposite signs. Observe that if C_i and C_j contain more than one common point, they are either all marked with the same sign or all marked with opposite signs, since the determination of the signs in $Q(S^+)$ corresponds to the determination of the signs in $CG^+(A)$, which can be formed since A is signed. In fact, this construction shows that $CG^+(A) \cong Q^+(S)$. Hence, we can reformulate the above theorem as follows.

THEOREM 2.6. *Given the graph $G^+ = (V, E)$ with $p = |V|$, the regular Boolean matrix A has the property that $RG^+(A) = G$ if and only if A has p rows and there exists a balanced clique cover S^+ of G^+ such that $CG^+(A)$ is isomorphic to $Q(S^+)$.*

The above relationship between signed graphs and rectangular matrices leads to the following graph theoretic result.

THEOREM 2.7. *Let G^+ be a signed graph and S^+ a balanced clique cover of G^+. Then $Q(S^+)$ is balanced if and only if G^+ is balanced.*

Proof. Let A be the matrix constructed as described in the comments following Theorem 2.4. Then $CG^+(A)$ is balanced if and only if $RG^+(A)$ is balanced by [6, Thm. 3]. Hence, the result follows since $Q(S^+) \cong CG^+(A)$ and $RG^+(A) = G^+$. □

Now let $R(G^+)=\{A:A$ is a regular signed Boolean matrix and $RG^+(A)=G^+\}$.

As in [7], we can consider the notion of minimality for the set $R(G^+)$. That is, given a matrix $A\in R(G^+)$, it is frequently important to find a matrix $A'\in R(G^+)$ that has fewer nonzeros than A or fewer columns than A. Basically, the same results hold as in [7], so we will not develop the theory of minimality here. However, there are two significant differences that we will describe.

First, for a graph G, the minimum number of columns for a matrix A in $R(G)$ was determined by $k(G)$, the clique cover number of G. However, as is illustrated by the graphs in Fig. 2.1, a clique cover of G^+ with the smallest number of cliques may not be a balanced clique cover. To find a matrix A in $R(G^+)$ with the minimum number of columns, we must find a balanced clique cover with the smallest number of cliques. This number is denoted by $k^+(G^+)$. Clearly, $k(G^+)\leqq k^+(G^+)$.

The other difference is the number of matrices A in $R(G^+)$ corresponding to a particular balanced clique cover $S^+=\{C_1,\cdots,C_n\}$ Since to each C_i there correspond two marked graphs M_i and M_i', the above construction leads to 2^n $p\times n$ matrices A with $RG^+(A)=G^+$ and $CG^+(A)\cong Q(S^+)$.

This situation is illustrated in Fig. 2.3 for the graph G^+ and balanced clique cover S^+ in Fig. 2.2. Observe that since $CG^+(A)=CG^+(-A)$, there are at most 2^{n-1} different graphs $Q(S^+)$ corresponding to the different ways of marking the cliques. Since $RG^+(A)$ is balanced, then $CG^+(A)$ must be balanced. Since $CG(A)$ is a 3-clique, there are only two ways of signing a 3-clique so that it is balanced. Hence, in this case, there are only two nonisomorphic signed clique cover graphs but four ways of signing the particular graph so that it is balanced.

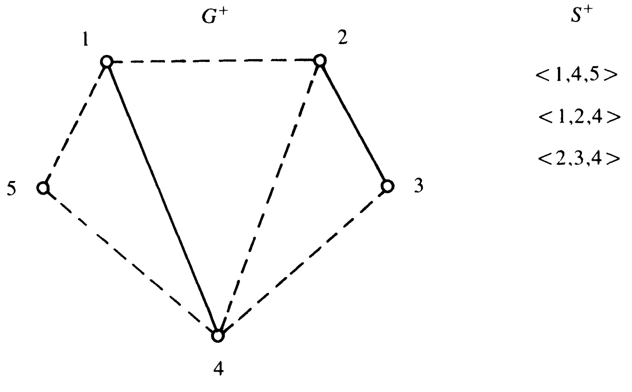
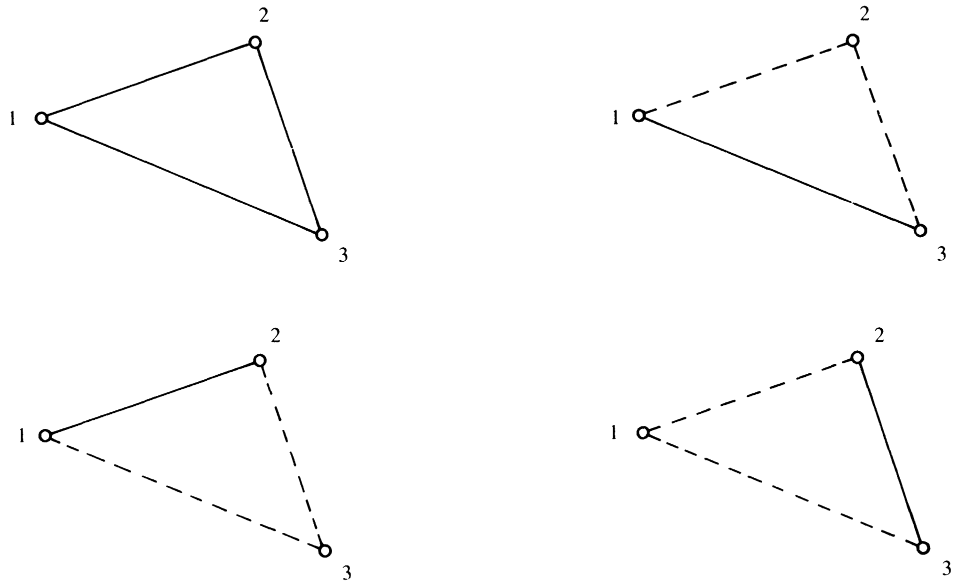


FIG. 2.2

Remark. It appears that what happens in the above example also happens in general. That is, let G^+ be balanced and S^+ be a balanced clique cover of G^+. Then all possible ways of signing $Q(S^+)$ so that it is balanced can be realized by changing the marked graphs for the various cliques in S^+.

3. Two-graph inversion. In this section we consider the problem of inverting a pair of signed graphs. That is, given signed graphs G_1^+ and G_2^+, when can a regular signed Boolean matrix A be constructed having the property that $RG^+(A)=G_1^+$ and $CG^+(A)=G_2^+$? If such a matrix A exists, as in [7], we say that G_1^+ and G_2^+ are invertible. Observe that for a regular matrix A, the signed graphs $RG^+(A)$ and $CG^+(A)$ have the same number of components by [5, Cor. 2.3]. It follows that we cannot always

$$CG^{+}(A)=CG^{+}(-A)\cong Q(S^{+})$$



Marked S^{+}

$$\begin{array}{l} \langle \overset{+}{1}, \overset{+}{4}, \bar{5}\rangle \\[1ex] \langle \overset{+}{1}, \bar{2}, \overset{+}{4}\rangle \\[1ex] \langle \bar{2}, \bar{3}, \overset{+}{4}\rangle \end{array} \qquad \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{bmatrix}$$

$$\begin{array}{l} \langle \overset{+}{1}, \overset{+}{4}, \bar{5}\rangle \\[1ex] \langle \bar{1}, \overset{+}{2}, \bar{4}\rangle \\[1ex] \langle \bar{2}, \bar{3}, \bar{4}\rangle \end{array} \qquad \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \\ 1 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}$$

$$\begin{array}{l} \langle \overset{+}{1}, \overset{+}{4}, \bar{5}\rangle \\[1ex] \langle \overset{+}{1}, \bar{2}, \overset{+}{4}\rangle \\[1ex] \langle \overset{+}{2}, \overset{+}{3}, \bar{4}\rangle \end{array} \qquad \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & -1 \\ -1 & 0 & 0 \end{bmatrix}$$

$$\begin{array}{l} \langle \overset{+}{1}, \overset{+}{4}, \bar{5}\rangle \\[1ex] \langle \bar{1}, \overset{+}{2}, \bar{4}\rangle \\[1ex] \langle \overset{+}{2}, \overset{+}{3}, \bar{4}\rangle \end{array} \qquad \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & -1 & -1 \\ -1 & 0 & 0 \end{bmatrix}$$

FIG. 2.3

invert the pair G_1^+, G_2^+. However, we can provide equivalent conditions for existence analogous to those given for a pair of graphs in [7, Thm. 2].

THEOREM 3.1. *Given two signed graphs G_1^+ and G_2^+, the following are equivalent*:

(i) *G_1^+ and G_2^+ are invertible*;

(ii) *G_1^+ is isomorphic to a signed clique cover graph of G_2^+*;

(iii) *G_2^+ is isomorphic to a signed clique cover graph of G_1^+*.

The proof is essentially the same as the proof of [7, Thm. 2], except that Theorem 2.6 is used instead of the one-graph inversion theorem of [7].

Now we will illustrate how Theorem 3.1 can be used to invert a pair of signed graphs. Consider the pair of graphs in Fig. 3.1. If we choose the balanced clique cover

$$S=\{\langle\overset{+}{1},\overset{+}{4},\bar{5}\rangle,\langle\bar{1},\overset{+}{2},\bar{4}\rangle,\langle\bar{2},\bar{3},\overset{+}{4}\rangle\}$$

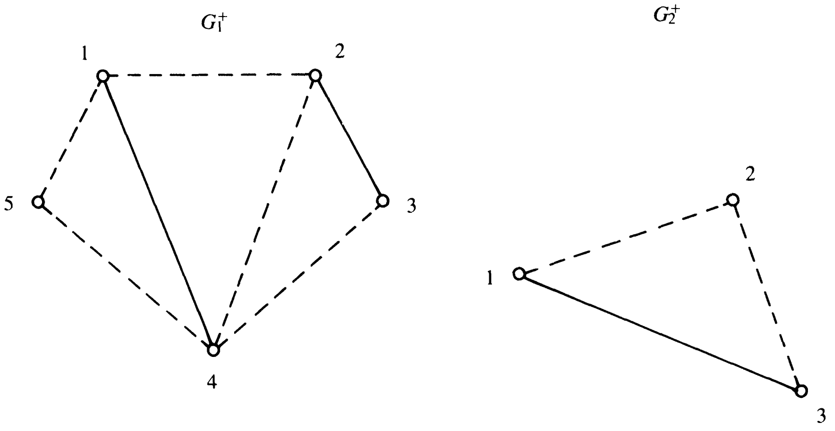


FIG. 3.1

and mark each clique in S as shown, then G_2^+ is isomorphic to $Q(S^+)$ as illustrated in Fig. 2.3. Using the signs for each clique, we can then construct

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

Clearly, $RG^+(A)=G_1^+$ and $CG^+(A)=G_2^+$. Observe that not only did we have to find a balanced clique cover S, but that we also needed an appropriate marking of the cliques in order to get $Q(S^+)\cong G_2^+$.

We close this section by observing that the concept of minimality can be developed in the same way as in [7].

REFERENCES

[1] L. W. BEINEKE AND F. HARARY, *Consistency in marked digraphs*, J. Math. Psych., 18 (1978), pp. 260–269.

[2] G. EXOO AND F. HARARY, *Step graphs*, J. Combin., Inform. System Sci., (1984), to appear.

[3] H. J. GREENBERG, *Measuring complementarity and qualitative determinacy in matricial forms*, Proceedings of the Symposium on Computer Assisted Analysis and Model Simplification, H. J. Greenberg and J. S. Maybee, eds, Academic Press, New York, 1981.

[4] H. J. GREENBERG, J. R. LUNDGREN AND J. S. MAYBEE, *Graph theoretic foundations of computer-assisted analysis*, 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.

[5] ———, *Green theoretic methods for the qualitative analysis of rectangular matrices*, this Journal, 2 (1981), pp. 227–239.

[6] ———, *Rectangular matrices and signed graphs*, this Journal, 4 (1983), pp. 50–61.

[7] ———, *Inverting graphs of rectangular matrices* Discrete Applied Math., (1984), to appear.

[8] ———, *The inversion of* 2-*step graphs.* . J. Combin., Inform. System Sci., (1984), to appear.

[9] ———, *Signed graphs of netforms*, (1984), to appear.

[10] F. HARARY, *Structural models and graph theory*, 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.

[11] F. HARARY AND J. A. KABELL, *An efficient algorithm to detect balance in signed graphs*, Mathematical Social Sciences, 1 (1980).

[12] F. HARARY, R. Z. NORMAN AND D. CARTWRIGHT, *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley, New York, 1965.

[13] A. KYDES AND J. S. PROVAN, *Correlation and determinacy in network models*, BNL Report 51243, Brookhaven National Laboratory, Upton, New York, 1980.

[14] J. R. LUNDGREN AND J. S. MAYBEE, *A characterization of graphs of competition number m*, Discrete Applied Math., 6 (1983), pp. 319–322.

[15] J. S. PROVAN, *Determinacy in linear systems and networks*, this Journal, 4 (1983), 262–278.

[16] F. S. ROBERTS, *Food webs, competition graphs, and the boxicity of ecological phase space*, Theory and Applications of Graphs—in America's Bicentennial Year, Y. Alavi and D. Lick, ed., Springer-Verlag, New York, 1978.

[17] ———, *Graph Theory and Its Applications to Problems of Society*, CBMS Regional Conference Series in Applied Mathematics 29, Society for Industrial and Applied Mathematics, Philadelphia, 1978.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

A GENERAL PRODUCT CONSTRUCTION FOR ERROR CORRECTING CODES*

K. T. PHELPS†

Abstract. A product construction for binary error correcting codes is presented. Given perfect binary single error correcting codes of length n and m, one can construct perfect binary single error correcting codes of length $nm+n+m$. Among other things, the construction is used to establish that the number of nonequivalent (perfect) binary single error correcting codes of length n is at least $2^{2^{cn}}$, for some constant $c<1$.

Key words. error correcting code, perfect code

AMS subject classification codes. 94B, 05B

1. Introduction. A code C of length n over an alphabet A can be thought of as a subset $C \subseteq A^n = A \times A \times \cdots \times A$. The (Hamming) distance between two code words $\mathbf{x}, \mathbf{y} \in C$, denoted by $d(\mathbf{x}, \mathbf{y})$, is simply the number of components in which the two vectors differ. If the alphabet $A = \{0, 1\}$ then C is called a binary code and $V^{(n)} = \{0, 1\}^n$ will be used to denote the vector space of dimension n over GF (2). The codes under consideration in this paper are almost exclusively binary codes; unless otherwise stated, any code can be assumed to be a binary code.

In this paper, we present a generalized product construction and then use it to establish lower bounds on the number of nonisomorphic and nonequivalent perfect single error correcting codes (or briefly perfect 1-codes). A binary code C of length n minimum distance d having M code words is often called an (n, M, d) code. A perfect 1-code is an $(2^k-1, 2^{2^k-k-1}, 3)$ code. Adding an overall parity check bit, gives us an extended perfect 1-code with parameters $(2^k, 2^{2^k-k-1}, 4)$. The product construction is most effective when dealing with (extended) perfect 1-code: the product of two extended perfect 1-code of lengths $(n+1)$ and $(m+1)$ is a perfect 1-code of length $(n+1)(m+1)$. For this reason, we focus our attention on extended perfect 1-codes (bounds on these are easily translated into bounds on perfect 1-codes).

The product construction can be applied to arbitrary codes; however, it appears to be more effective when the codes are single error correcting. The doubling constructions of Phelps [5] and Sloane and Whitehead [8] can be thought of, in varying degrees, as special cases of our product construction.

2. Generalized product construction. The construction is presented with regard to perfect 1-codes. It can be applied to more general classes of codes but with lessened efficiency. In what follows, let $V^{(n)}$ denote the vector space of dimension n over GF (2).

Let $C_0^0, C_1^0, \cdots, C_n^0$ be a partition of the even weight vectors of $V^{(n+1)}$ into extended perfect 1-codes of length $n+1$ (i.e., $n+1=2^k$, $|C_i^0| = 2^{n-k}$ and for any $\mathbf{x}, \mathbf{y} \in C_i^0$, $d(\mathbf{x}, \mathbf{y}) \geqq 4$ unless $\mathbf{x} = \mathbf{y}$). Similarly, let $C_0^1, C_1^1, \cdots, C_n^1$ be a partition of the odd weight vectors of $V^{(n+1)}$ into extended perfect 1-codes. Given a perfect 1-code of length n one can always find at least one such partition—a code and its translates.

Let $R \subseteq V^{(m+1)}$ be an extended perfect 1-code of length $m+1 = 2^p$. For each code word $\mathbf{r} \in R$, let $q_r(a_0, a_1, \cdots, a_{m-1}) = a_m$ be a m-ary quasigroup of order $n+1$. Alternately, one can think of $(a_0, a_1, \cdots, a_m)$ as a distance 2 code of length $m+1$ over an

[*] Received by the editors November 24, 1982, and in revised form August 16, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

alphabet of order $n+1$. The generalized direct product of codes C and R, denoted by $C\otimes_q R$ is defined as a code of length $(n+1)(m+1)=nm+n+m+1$ with:

$$\mathbf{x}=(\mathbf{x}_0,\mathbf{x}_1,\cdots,\mathbf{x}_m)\in C\otimes_q R \quad \text{iff} \quad \mathbf{x}_i\in C_{j_i}^{r_i},$$

where $\mathbf{r}=(r_0,r_1,\cdots,r_m)\in R$ and

$$q_{\mathbf{r}}(j_0,j_1,\cdots,j_{m-1})=j_m.$$

THEOREM 2.1. *The code $C\otimes_q R$ constructed above is an extended perfect 1-code of length $(n+1)(m+1)=2^{k+p}$.*

Proof. $|C_{j_i}^{r_i}|=2^{n-k}$ and $|R|=2^{m-p}$ for each $\mathbf{r}\in R$ one constructs

$$|C_{j_i}^{r_i}|^{m+1}(n+1)^m=(2^{n-k})^{m+1}(2^k)^m=2^{nm-k+n}$$

code words. This gives

$$2^{nm-k+n}2^{m-p}=2^{nm+n+m-(k+p)}=2^{2^{k+p}-(k+p)-1}$$

code words, which is the correct number of code words. All we need to do is to establish that for any $\mathbf{x},\mathbf{y}\in C\otimes_q R$, $d(\mathbf{x},\mathbf{y})\geqq 4$ unless $\mathbf{x}=\mathbf{y}$.

Let $\mathbf{x}=(\mathbf{x}_0,\mathbf{x}_1,\cdots,\mathbf{x}_m)$ and $\mathbf{y}=(\mathbf{y}_0,\mathbf{y}_1,\cdots,\mathbf{y}_m)$ be any two code words in $C\otimes_q R$. Then obviously

$$d(\mathbf{x},\mathbf{y})\geqq\sum_{i=0}^{m} d(\mathbf{x}_i,\mathbf{y}_i),$$

where $\mathbf{x}_i,\mathbf{y}_i$ are vectors of length $n+1$. If r_i denotes the parity of $\mathbf{x}_i$ and s_i denotes that of $\mathbf{y}_i$ for $i=0,1,\cdots,m$ then $\mathbf{r}=(r_0,r_1,\cdots,r_n)$ and $\mathbf{s}=(s_0,s_1,\cdots,s_m)\in R$. (Note: $r_i=0$ or 1 depending on whether $\mathbf{x}_i$ has even or odd parity—similarly for s_i.) However, if $d(\mathbf{x}_i,\mathbf{y}_i)=0$ then the parity must be the same and thus $r_i=s_i$. When $r_i\neq s_i$ then $d(\mathbf{x}_i,\mathbf{y}_i)\geqq 1$. Since $d(\mathbf{r},\mathbf{s})\geqq 4$, this means that $d(\mathbf{x}_i,\mathbf{y}_i)\geqq 1$ for at least four values of i and thus

$$\sum_{i=0}^{m} d(\mathbf{x}_i,\mathbf{y}_i)\geqq 4,$$

unless $\mathbf{r}=\mathbf{s}$.

If $\mathbf{r}=\mathbf{s}$, then the parity of $\mathbf{x}_i$ and $\mathbf{y}_i$ will be the same and $d(\mathbf{x}_i,\mathbf{y}_i)\geqq 2$ unless $\mathbf{x}_i=\mathbf{y}_i$. Assume $\mathbf{x}_i\in C_{j_i}^{r_i}$ for $i=0,1,\cdots,m$ and $\mathbf{y}_i\in C_{k_i}^{r_i}$ for $i=0,1,\cdots,m$. Then $d(\mathbf{x}_i,\mathbf{y}_i)=0$ implies that $j_i=k_i$; since $\mathbf{j}=(j_0,j_1,\cdots,j_m)$ and $\mathbf{k}=(k_0,k_1,\cdots,k_m)$ can agree in at most $m-1$ positions then $d(\mathbf{x}_i,\mathbf{y}_i)\geqq 2$ for at least two values of i and then $d(\mathbf{x},\mathbf{y})\geqq 4$—unless $\mathbf{j}=\mathbf{k}$. However, in this case if $\mathbf{x}_i\neq\mathbf{y}_i$ then $d(\mathbf{x}_i,\mathbf{y}_i)\geqq 4$, and again $d(\mathbf{x},\mathbf{y})\geqq 4$.

In conclusion, we see that the minimum distance between any two code words is four and thus $C\otimes_q R$ is an extended perfect 1-code of length $nm+n+m+1$.

If we assume that the codes C and R each contain the zero vector; that, in the previous construction, $C_0^0=C$ and $q_{\mathbf{0}}(0,0,0,\cdots,0)=0$, then $C\otimes_q R$ will contain a subcode isomorphic to C. If in addition, $q_{\mathbf{r}}(0,0,\cdots,0)=0$ for each $\mathbf{r}\in R$ and C_0^1 contains a vector of weight 1 then $C\otimes_q R$ will contain a subcode isomorphic to R.

Note that if R is a "perfect" 1-code of length 2 (i.e., R consists of a single vector) then the construction presented in Phelps [5] is a special case of the above construction, since a 1-ary quasigroup is in effect a permutation.

The main use of this construction is in establishing lower bounds on the number of nonisomorphic and nonequivalent perfect 1-code. However, the construction can be usefully applied to other less "perfect" codes. Later sections of this paper will deal with such applications.

3. Lower bounds on the number of perfect 1-codes. First, we determine the number of different codes that can be constructed from given perfect (extended) 1-codes, C, R of length $n+1$ and $m+1$, respectively. Assuming that $C=C_0^0$, then we choose some fixed partition $C_0^0, C_1^0, \cdots, C_n^0, C_0^1, C_1^1, \cdots, C_n^1$ of $V^{(n+1)}$ into extended perfect 1-codes which satisfies conditions for our construction. The remarkable fact is that from such modest assumptions one can still construct an incredible number of different extended perfect 1-codes of length $(n+1)(m+1)$.

Let $Q(m, n+1)$ denote the set of all m-ary quasigroups (or m-quasigroups) of order $n+1$. Equivalently $Q(m, n+1)$ can be thought of as the set of distance 2 codes of length $m+1$ over an alphabet of order $n+1$ having $(n+1)^m$ code words. When $m=2$, we have a quasigroup (or equivalently a latin square) of order $n+1$; the number of such quasigroups is asymptotically, $(n+1)^{(n+1)^2}$ (cf. [3]). The number of quasigroups having a left identity (e.g., $q(0, x)=x$)—which is equivalent to the number of different row-reduced latin squares—is greater than $(n+1)^{(n+1)^2-(n+1)}$ for n sufficiently large. This is the basis for the following lemma.

LEMMA 3.1. *$|Q(m, n+1)| \geqq |Q(m-1, n+1)|(n+1)^{(n+1)^2-(n+1)}$ and thus $|Q(m, n+1)| \geqq (n+1)^{[(n+1)^2-(n+1)](m-1)}$ for n sufficiently large.*

Proof. The lemma is true for $m=2$ since $|Q(2, n+1)| \approx (n+1)^{(n+1)^2}$. For any two quasigroups $q_i, q_j \in Q(2, n+1)$, $q_i(x, q_j(y, z)) \in Q(3, n+1)$. Moreover, if q_i, q_r have left identities then if $q_i(x, q_j(y, z)) = q_r(x, q_s(y, z))$ for all x, y, z, then $q_j = q_s$ and thus $q_i = q_r$. Hence $Q(3, n+1) \geqq (n^{(n+1)^2-(n+1)})^2$. More generally, if $q_j \in Q(m-1, n+1)$ and $q_i \in Q(2, n+1)$, then $q_i(x, q_j(\mathbf{y})) \in Q(m, n+1)$ and if q_i has a left identity then the choice of q_i, q_j uniquely determines the m-quasigroup. Hence

$$|Q(m, n+1)| \geqq |Q(m-1, n)|(n+1)^{(n+1)^2-(n+1)}$$

and the above lower bound follows directly.

THEOREM 3.2. *The number of nonisomorphic extended perfect 1-codes of order $(n+1)(m+1)=2^{k+p}$ is greater than*

$$2^{k2^k(2^k-1)(2^{p-1}-1)2^{2^p-1}-(k+p)2^{k+p}}.$$

Proof. Given C_j^i, $j=0, 1, \cdots, n$ and $i=0, 1$, and R satisfying the requirements of the construction presented in § 2, we can construct $|Q(m, n+1)|^{|R|}$ different codes since for each $\bar{r} \in R$ one can choose any m-quasigroup $q_{\bar{r}} \in Q(m, n+1)$. If the lengths of R and C_j^i are $(m+1)=2^p$ and $(n+1)=2^k$, respectively, then the number of nonisomorphic 1-codes of length 2^{k+p} is at least,

$$|Q(2^p-1, 2^k)|^{|R|}/(2^{k+p})! \geqq (2^k)^{(2^{2k}-2^k)(2^p-2)2^{2^p-p-1}}/2^{(k+p)2^{k+p}},$$

or at least

$$2^{k(2^{2k}-2^k)(2^p-2)2^{2^p-p-1}-(k+p)2^{k+p}},$$

which equals the lower bound of the theorem.

THEOREM 3.3. *The number of nonequivalent (extended) perfect 1-codes of length $(n+1)(m+1)=2^{k+p}$ is greater than*

$$2^{k2^k(2^k-1)(2^{p-1}-1)2^{2p-1}-(k+p)2^{k+p}-2^{k+p}-(k+p+1)}.$$

Proof. The argument is almost the same as in Theorem 3.2. Since each m-ary quasigroup constructed in Lemma 3.1 has $q(0, 0, \cdots, 0)=0$ and $C_0^0=C$ and R are both assumed to have the zero vector. We conclude that each code constructed by our method will contain the zero vector. Any code C, containing the zero vector, is equivalent to at most $(n!)|C|$ other codes each of which also contain the zero vector.

Thus, we have that there are at least

$$|Q(2^p-1,2^k)|^{|R|}/(2^{k+p})!2^{2^{k+p}-(k+p)-1}$$

nonequivalent codes constructed from fixed codes R, C of length 2^p and 2^k respectively. Dividing the lower bound of Theorem 3.2 by $2^{2^{k+p}-(k+p)-1}$ gives the above lower bound on nonequivalent perfect 1-codes.

The trivial upper bound on the number of 1-codes of length n is $2^{2^{n(1-o(1))}}$. Our lower bound is remarkably close.

COROLLARY 3.4. *The number of inequivalent extended perfect* 1*-codes of length* $n+1$ *is greater than* $2^{2^{cn}}$, *for a constant* $c<1$.

With a little reflection, it should be evident that the number of nonequivalent (extended) perfect 1-codes will be considerably larger than the lower bound of Theorem 3.3. Not only can one choose different codes C, R but one can choose different partitions of $V^{(n+1)}$.

Having achieved the primary purpose of this article—the establishment of a lower bound on the number of nonequivalent (extended) perfect 1-codes—we now consider the application of our methods of § 2 to the construction of (new) families of error correcting codes.

4. Single error correcting binary codes. Although our construction was presented in terms of (extended) perfect 1-codes, it should be clear that this limitation is unnecessary. What follows is the above construction in its most general form.

Construction 4.0. Let $C^i_{j,k}$, $j=0,1,\cdots,p-1$, $k=0,1,\cdots,m$, $i=0,1$, be codes of length n and distance d_1 where the code words of $C^0_{j,k}$ are all of even weight and those of $C^1_{j,k}$ are of odd weight, where moreover for fixed i, k the codes $C^i_{j,k}$, $j=0,1,\cdots,p-1$ are mutually disjoint. Let R be a code of length $m+1$ and distance d_2 and $Q_\mathbf{r}$ be a code of length $m+1$, distance d_3 over an alphabet of order p (i.e., $\{0,1,\cdots,p-1\}$) for each $\mathbf{r}\in R$. Then one forms a code of length $n(m+1)$:

$$C\otimes_Q R=\bigcup_{\mathbf{r}=(r_0,r_1,\cdots,r_m)\in R}\left(\bigcup_{(j_0,j_1,\cdots,j_m)\in Q_\mathbf{r}} C^{r_0}_{j_0,0}\oplus C^{r_2}_{j_1,1}\oplus\cdots\oplus C^{r_m}_{j_m,m}\right).$$

COROLLARY 4.1. *In the code* $C\otimes_Q R$ *of Construction* 4.0, *the distance between any two code words is greater than the* min $\{d_1, d_2, 2d_3\}$. *If* $|C^i_{j,k}|=c^i_{j,k}$ *then the number of code words is*:

$$\sum_{(r_0,r_1,\cdots,r_m)=\mathbf{r}\in R}\ \sum_{(j_0,j_1,\cdots,j_m)\in Q_\mathbf{r}}\left(\prod_{k=0}^{m} c^{r_k}_{j_k,k}\right).$$

Proof. The argument follows that of Theorem 2.1 almost verbatum.

For single error correcting codes, we can choose each $Q_\mathbf{r}$ so that $d_3=2$ and the number of code words, $|Q_\mathbf{r}|$, is p^m. For $C^i_{j,k}$, we have $d_1=4$. Needless to say, for fixed k, we would like the codes $C^i_{j,k}$ to partition V^n as this would clearly increase the number of code words in $C\otimes_Q R$ at no additional cost.

COROLLARY 4.2. *If the* $C^i_{j,k}$ *are distance* 4 *codes of length* n, R *is a distance* $d_2=3$ *or* 4, *code of length* $m+1$, *then* $C\otimes_Q R$ *is a single error correcting code having minimum distance* d_2. *Moreover if* $|C^i_{j,k}|\geqq M_1$ *and* $|R|=M_2$ *then* $|C\otimes_Q R|\geqq p^m M_1^{m+1}M_2$ *or in any case* $|C\otimes_Q R|\geqq n^m M_1^{m+1}M_2$.

Proof. For any (extended) single error correcting code C, we can choose $C^i_{j,k}$, $i=0,1,\cdots,p-1$ to be translates of C and thus we can assume that $p\geqq n$.

For example, starting with an extended perfect code of length 8 and a perfect code of length 7 (i.e., an $(8,2^4,4)$ and a $(7,2^4,3)$) we can construct an

$(8\cdot 7, 8^6\cdot 16^7\cdot 16, 3) = (56, 2^{50}, 3)$ code. Sloane and Whitehead [8] constructed a single error correcting code with the same parameters and in a broad sense one can consider our product construction as a generalization of their doubling construction. Certainly, if the $C^i_{j,k}$ are chosen to be extended perfect 1-codes, we can construct the same family of 1-codes [4], [8]. Our product construction will produce other families of codes as well. However, to be truly effective, not only does one need good codes C, R but more importantly good partitions, $C^i_{j,k}$ of $V^{(n)}$. Of course, with the exception of perfect codes, little is known with regard to the existence of such partitions.

5. *t*-error correcting codes. To apply construction 4.0 to t-error correcting codes, we need to be able to construct distant d codes over an alphabet of order n. This actually is relatively easy to do.

An orthogonal array of order n, depth m, and strength t (with index unity) is an $n^t \times m$ array with entries from an n-set such that any two row vectors agree in at most $t-1$ positions. Such an array is equivalent to an $(m, n^t, m-t+1)$ code over an alphabet of n symbols where the code words are row vectors. For results on the existence of orthogonal arrays, see Raghavarao [7], Bush [1], [2], or Phelps [6].

To illustrate the construction, let C be the (extended) perfect binary Golay code $(24, 2^{12}, 8)$ and R an $(8, 2, 8)$ code. The code Q is a code of length 8 distance 4 over an alphabet of order 2^{11}. Since 2^{11} is a prime power there exist an orthogonal array of order 2^{11} depth 8 and strength 5. Hence, we can assume Q exists and has $(2^{11})^5 = 2^{55}$ code words. V^{24} can be completely partitioned into extended perfect Golay codes. Thus $C\otimes_Q R$ is an $(24\cdot 8, 2^{55}(2^{12})^8 2, 8) = (192, 2^{152}, 8)$ code. Unfortunately this code is not optimal even though C, Q, R were optimal.

6. Conclusion. The product construction presented in § 2 when applied to single error correcting codes is quite effective. It appears to be less effective when applied to t-error correcting codes, $t > 1$. There are several directions for further inquiry. The first would be the construction of good partitions of $V^{(n)}$ into single error correcting codes (of length n) at least for small n (i.e., $n < 15$). This would allow for a more effective utilization of the product construction and thus one probably would be able to improve on some of the current lower bounds on the size of single error correcting codes.

The product construction was used primarily to construct nonequivalent extended perfect 1-codes. However, it should be evident that similar arguments could be used to establish lower bounds on the number of nonequivalent codes for other families of codes as well.

REFERENCES

[1] K. A. Bush, *Orthogonal arrays of index unity*, Ann. Math. Stat., 23 (1952), pp. 426–434.

[2] ———, *A generalization of a theorem due to MacNeish*, Ann. Math. Stat., 23 (1952), 293–295.

[3] J. Denes and A. Keedwell, *Latin Squares and Their Applications*, Academic Press, New York.

[4] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1978.

[5] K. T. Phelps, *A combinatorial construction of perfect codes*, this Journal, 4 (1983), pp. 398–403.

[6] ———, *Direct product of derived Steiner systems using inversive planes*, Canad. J. Math., 33 (1981), pp. 1365–1369.

[7] D. Raghavarao, *Constructions and Combinatorial Problems in the Design of Experiments*, John Wiley, New York, 1971.

[8] N. J. A. Sloane and D. S. Whitehead, *A new family of single-error-correcting codes*, IEEE Trans. Inform. Theory, IT-16 (1970), pp. 717–719.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

009

GREEDOIDS AND LINEAR OBJECTIVE FUNCTIONS*

BERNHARD KORTE† AND LÁSZLÓ LOVÁSZ‡

Abstract. Greedoids were introduced by the authors as generalizations of matroids providing a framework for the greedy algorithm. They can be characterized algorithmically via the optimality of the greedy algorithm for a class of objective functions, which are in general not linear and do not include all linear functions. It is therefore natural to ask the following questions: (1) What are those linear objective functions which can be optimized over any greedoid by the greedy algorithm; (2) what are those greedoids over which the linear objective function can be optimized by the greedy algorithm. This paper gives an answer to both questions. Moreover, it gives slimming procedures for obtaining such greedoids from matroids and it gives briefly some (negative) oracle results about greedoid optimization and greedoid recognition.

1. Introduction. In previous papers (Korte and Lovász [1981] and [1982a]) we have introduced greedoids as generalizations of matroids providing a framework for the greedy algorithm. Matroids can be characterized axiomatically as those subclusive set-systems for which the greedy solution is optimal for certain optimization problems (e.g. linear objective functions, bottleneck functions). Greedoids can also be characterized algorithmically via the optimality of the greedy algorithm for a class of objective functions, which are in general not linear and do not include all linear functions (cf. Korte and Lovász [1982a]). It is therefore natural to ask the following questions: (1) What are those linear objective functions which can be optimized over any greedoid by the greedy algorithm; (2) what are those greedoids over which any linear objective function can be optimized by the greedy algorithm. This paper gives an answer to both questions.

The algorithmic principle of greediness, i.e. of a locally myopic strategy, can be defined in different ways. The most common greedy approach is that of *best-in greedy*: starting with the empty set, the greedy solution will be built up recursively by adding the best possible element to it at each step, while remaining feasible. Another approach is that of *worst-out greedy*. Here we start with the complete ground set and eliminate from it in each step the worst-possible element as long as the remaining set is spanning. For matroids both approaches are equivalent, since the worst-out greedy is the best-in greedy for the negative objective function over the dual matroid. In the case of greedoids, it turns out that for general linear objective functions the worst-out greedy is optimal for a broader class of greedoids than the best-in-approach.

In § 2 we give some definitions and basic facts about greedoids, which will be needed in the rest of the paper. However, the interested reader is referred to Korte and Lovász [1982a] and [1982b] for a more detailed study of structural aspects of greedoids. Section 3 gives a compatibility characterization of linear objective functions which is sufficient to optimize these functions over *any* greedoid by the greedy algorithm. Section 4 characterizes those greedoids over which *any* linear objective function can be optimized by the worst-out greedy algorithm. A proper subclass of

* Received by the editors July 6, 1983. This work was supported by the joint research project "Algorithmic Aspects of Combinatorial Optimization" of the Hungarian Academy of Sciences (Magyar Tudományos Akadémia) and the German Research Association (Deutsche Forschungsgemeinschaft, SFB 21). This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Institut für Ökonometric und Operations Research, Rheinische Friedrich Wilhelms Universität, Bonn, West Germany.

‡ Eötvös Loránd University, Department of Analysis I, Budapest, Hungary.

these greedoids has the property that the best-in greedy is optimal for any linear objective function. Section 5 gives some construction principles to obtain those greedoids by slimming a given matroid. Finally, in § 6 we state some (negative) oracle results about greedoids, among which it is noteworthy that the problem of optimizing an arbitrary linear objective function over a general greedoid given by a feasibility oracle is NP-hard. There is also no polynomial feasibility oracle algorithm to distinguish a greedoid from a matroid.

2. Definitions and basic facts about greedoids. We assume that the reader is familiar with the basic facts of matroid theory (cf. Welsh [1976]) and in general our notation is in accordance with the standard matroid terminology.

A *set-system* over a finite ground set E is a pair $(E, \mathcal{F})$ with $\mathcal{F} \subseteq 2^E$. A set-system is a *matroid* if the following axioms hold:

(M1) $\varnothing \in E$;

(M2) $X \subseteq Y \in \mathcal{F}$ implies $X \in \mathcal{F}$;

(M3) if $X, Y \in \mathcal{F}$ and $|X| > |Y|$, then there exists a $x \in Y - X$ such that $Y \cup \{x\} \in \mathcal{F}$.

A set-system which satisfies only (M1) and (M2) has little structure, but very different names. It is called *independence system, simplicial complex, subclusive* or *hereditary set-system.* For an arbitrary set-system $(E, \mathcal{F})$ we define its *hereditary closure* $\mathcal{H}$ as:

$$\mathcal{H} := \{X \subseteq Y \colon Y \in \mathcal{F}\}.$$

Another but more structural way to relax matroids is to keep the exchange axiom (M3) (and the trivial axiom (M1)) but to remove subclusiveness (M2); and this is exactly one way to define *greedoids.* There is another equivalent and even more natural way to define greedoids via extending the matroidal structure to languages, i.e. systems of ordered sets or strings, but for the purpose of this paper it is sufficient to consider only the unordered version of greedoid definition. Thus, we call a set-system $(E, \mathcal{F})$ a *greedoid* if (M1) and (M3) holds. (M1) and (M3) imply a weak subclusiveness, which we call *accessibility*:

(M2′) for all $X \in \mathcal{F}$ there exists $x \in X$ such that $X - \{x\} \in \mathcal{F}$.

Analogously to hereditary set-systems we call a set system which satisfies (M1) and (M2′) an *accessible set-system.* We define the *accessible kernel* $\mathcal{K}$ of a set-system $(E, \mathcal{F})$ as

$$\mathcal{K} := \{X \in \mathcal{F} \colon X = \{x_1, \cdots, x_k\} \text{ and } \{x_1, \cdots, x_i\} \in \mathcal{F} \text{ for all } 1 \leqq i \leqq k\}.$$

(M1) and (M3) are equivalent to (M1), (M2′), and

(M3′) if $X, Y \in \mathcal{F}$ and $|X| = |Y| + 1$, then there exists a $x \in X - Y$ such that $Y \cup \{x\} \in \mathcal{F}$.

In the case of matroids (M3) and (M3′) are equivalently used, but this is only possible since (M2) holds. In analogy to matroid theory, we call sets which belong to $\mathcal{F}$ *feasible* or *independent.* Maximal independent sets are called *bases.* An element $d \in E$ is called *dummy,* if it does not occur in any feasible set. A greedoid is *normal,* if it has no dummy elements; it is called *full* if $E \in \mathcal{F}$.

For a greedoid we can define the (*independence*) *rank* of a set $X \subseteq E$ as:

$$r(X) := \max \{|A| \colon A \subseteq X, A \in \mathcal{F}\}.$$

This function has the following properties for $X, Y \subseteq E$ and $x, y \in E$

(R1) $r(\emptyset)=0$;

(R2) $r(X) \leqq |X|$;

(R3) if $X \subseteq Y$ then $r(X) \leqq r(Y)$;

(R4) if $r(X)=r(X \cup \{x\})=r(X \cup \{y\})$ then $r(X)=r(X \cup \{x\} \cup \{y\})$.

Conversely, a function $r: 2^E \to \mathbb{Z}$ satisfying (R1), (R2). (R3) and (R4) defines uniquely a greedoid (cf. Korte and Lovász [1982a]). These axioms are again direct relaxations of the rank definition of matroids, which in addition have the *unit increase property*:

$$r(X \cup \{x\}) \leqq r(X)+1 \quad \text{for } X \subseteq E, x \in E.$$

From (R1)–(R4) and the unit increase property one derives in matroid theory that the rank function is *submodular*, i.e. $r(X \cap Y)+r(X \cup Y) \leqq r(X)+r(Y)$. This fails to hold for greedoids in general; but the property (R4), which we call *local submodularity*, is often a reasonable substitute.

In contrast to matroids, the intersection of a set with a basis of a greedoid may have larger cardinality than the rank of this set. Therefore we define the *basis rank* of a set $X \subseteq E$ as

$$\beta(X) := \max\{|X \cap B|: B \in \mathcal{F}\}.$$

Clearly, $\beta(X) \geqq r(X)$. A set $X \subseteq E$ is called *rank-feasible* if $\beta(X)=r(X)$. We denote the family of all rank feasible sets by $\mathcal{R}=\mathcal{R}(E, \mathcal{F})$. Clearly, $\mathcal{F} \subseteq \mathcal{R}$ and $\mathcal{F}=\mathcal{R}$ for a full greedoid. In general $(E, \mathcal{R})$ is not a greedoid and $\mathcal{R}$ is not closed under union.

We recall here some facts about rank-feasibility (cf. Korte and Lovász [1982b]): A greedoid is a matroid iff $\mathcal{R}=2^E$. For $A, B \subseteq E$ we have

$$\beta(A \cup B)+r(A \cap B) \leqq \beta(A)+\beta(B)$$

and consequently, if $A, B \in \mathcal{R}$ then

$$r(A \cup B)+r(A \cap B) \leqq r(A)+r(B),$$

i.e. r is submodular. This can be also derived from the fact that $A \in \mathcal{R}$ iff $r(A \cup X) \leqq r(A)+|X|$ for all $X \subseteq E-A$.

A fundamental concept in matroid theory is the closure operator. Therefore we define analogously for greedoids the (*rank*) *closure* of a set $X \subseteq E$ as

$$\sigma(X) := \{x \in E: r(X \cup \{x\})=r(X)\}.$$

This operator is not monotone, but it has the following properties:

(C1) $X \subseteq \sigma(X)$ for all $X \subseteq E$;

(C2) if $X \subseteq Y \subseteq \sigma(X)$ then $\sigma(X)=\sigma(Y)$;

(C3) if $X \subseteq E$ and $x \in E-X$ such that for all $z \in X \cup x$, $z \notin \sigma(X \cup x-z)$, and $x \in \sigma(X \cup y)$, then $y \in \sigma(X \cup x)$.

It was shown in Korte and Lovász [1982a] that a mapping $\sigma: 2^E \to 2^E$ satisfying (C1), (C2), and (C3) uniquely defines a greedoid.

The closure axioms for greedoids are again relaxations of the closure for matroids. (C1) is trivial, (C2) follows from monotonicity and idempotence, and (C3) is a weakening of the Steinitz–McLane axiom for matroids. It can be shown that (C2) implies idempotence, but of course not monotonicity.

A set $X \subseteq E$ is called *closed* if $X = \sigma(X)$. An easy construction leads to a *monotone* closure operator, namely

$$\mu(X) := \bigcap \{Y: X \subseteq Y \text{ and } Y \text{ closed}\}.$$

μ does not determine the greedoid uniquely. In fact, for a full greedoid we have $\mu = \mathrm{id}$.

We call a set *closure-feasible* if $X \subseteq \sigma(A)$ implies $X \subseteq \mu(A)$, or—which is equivalent—if $X \subseteq \sigma(A)$ implies $X \subseteq \sigma(B)$ for $A \subseteq B \subseteq E$. The family of all closure feasible sets will be denoted by $\mathscr{C} = \mathscr{C}(E, \mathscr{F})$. The family $\mathscr{C}$ is closed under union and we have $\mathscr{C} \subseteq \mathscr{R}$. Further, $\mathscr{C}$ with inclusion as a partial order forms a lattice with the operation $A \vee B := A \cup B$ and $A \wedge B := \bigcup\{C \in \mathscr{C}: C \subseteq A \cap B\}$. The rank function r is submodular on this lattice. $(E, \mathscr{C})$ is not a greedoid in general, but the accessible kernel $\mathscr{K} = \mathscr{K}(\mathscr{C})$ of $\mathscr{C}$ defines trivially a greedoid. The rank function does not have the unit increase property on $\mathscr{C}$. But since $\mathscr{K} \subseteq \mathscr{C}$ is also a lattice, the rank function is also submodular on $\mathscr{K}$.

A very substantial subclass of greedoids are *interval greedoids*. We call a greedoid $(E, \mathscr{F})$ an interval greedoid if for all $A, B, C \in \mathscr{F}$ with $A \subseteq B \subseteq C$ and $x \in E - C$ such that $A \cup x \in \mathscr{F}$ and $C \cup x \in \mathscr{F}$, it follows that $B \cup x \in \mathscr{F}$. In Korte and Lovász [1982b] it was shown that a greedoid is an interval greedoid iff $\mathscr{C} = \mathscr{R}$ and iff $\mathscr{F} \subseteq \mathscr{R}$. Generally, no inclusion relation holds between $\mathscr{F}$ and $\mathscr{C}$. Furthermore, if $(E, \mathscr{F})$ is an interval greedoid, then already $(E, \mathscr{C})$ is a greedoid. We call a normal greedoid a *shelling structure* if the *interval property* mentioned above holds *without upper bounds*, i.e. if for all $A \subseteq B$ and $x \in E - B$ such that $A \cup x \in \mathscr{F}$ it follows $B \cup x \in \mathscr{F}$. Shelling structures are studied in greater detail in Korte and Lovász [1983a].

3. Special linear objective functions and general greedoids. An optimization problem over a greedoid $(E, \mathscr{F})$ can be described by introducing a linear objective function $w: E \to \mathbb{R}$ as a weighting of the elements of the ground set. This function can be extended to a modular function $w: 2^E \to \mathbb{R}$ by $w(X) := \sum_{x \in X} w(x)$ for all $X \subseteq E$. For reasons of simplicity we will consider in the following only maximization problems, i.e.

$$\max\{w(F): F \in \mathscr{F}\}.$$

We call a basis X of $(E, \mathscr{F})$ an *optimal basis* for which $w(X)$ is maximal among all bases.

The principle of the *greedy algorithm* (or more precisely: the *best-in greedy algorithm*) can be briefly described by the greedy bases, which are obtained with this algorithm. We call a basis $\{x_1, \cdots, x_r\}$ of a greedoid $(E, \mathscr{F})$ a (*best-in*) *basis for* w if it is obtained by the following recurrence: x_{i+1} is the element with the largest weight in $E - \{x_1, \cdots, x_i\}$ such that $\{x_1, \cdots, x_i, x_{i+1}\} \in \mathscr{F}$.

In the next section we refer to a *worst-out greedy algorithm* which in contrast starts with the ground set E and eliminates elements with the smallest possible weight as long as the remaining set is spanning, i.e. contains a basis. The *worst-out greedy basis for* w is then a basis $Y = E - \{x_1, \cdots, x_k\}$ which is obtained by the recurrence: x_{i+1} is the element with smallest weight in $E - \{x_1, \cdots, x_i\}$ such that $E - \{x_1, \cdots, x_i, x_{i+1}\}$ is spanning. It is an easy observation that for matroids the best-in greedy basis and the worst-out greedy basis are identical.

In general, an arbitrary linear objective function cannot be optimized over a greedoid with the greedy algorithm. Therefore, we need the following compatibility definition: Let $\mathscr{S} \subseteq 2^E$, and let $w: E \to \mathbb{R}$. We say that w is $\mathscr{S}$-*compatible* if $\{x \in E: w(x) \geqq c\} \in \mathscr{S}$ for all $c \in \mathbb{R}$, i.e. all *level sets* of w are in $\mathscr{S}$. As usual, we call a function $w: E \to \{0, 1\}$ the *characteristic function* of a set $X \subseteq E$ iff $w(x) = 1$ for all $x \in X$.

Then the definition of rank-feasibility implies immediately the following:

LEMMA 3.1. *If w is the characteristic function of a rank-feasible set, then all greedy bases are optimal.*

Our aim is to prove the following theorem:

THEOREM 3.2. *Let $(E, \mathscr{F})$ be a greedoid and $w: E \to \mathbb{R}$ be an $\mathscr{R}$-compatible weighting. Then all greedy basis for w are optimal.*

Proof. We can write w in the form

$$w = \sum_{i=1}^{t} \lambda_i w_i,$$

where $w_1 \leqq w_2 \leqq \cdots \leqq w_t$ are characteristic functions of rank-feasible sets, and $\lambda_1, \cdots, \lambda_t > 0$. In fact, let $c_1 > c_2 > \cdots > c_t$ be the different values assumed by w over 2^E, and let X_i be the level set $X_i = \{x: w(x) \geqq c_i\}$. Then we can choose w_i to be the characteristic function of X_i and $\lambda_i = c_i - c_{i+1}$, $\lambda_t = c_t$.

Let X be a greedy basis for w. Then X is, clearly, a greedy basis for each w_i. So by Lemma 3.1, X is an optimal basis for each w_i. But then, clearly, X is an optimal basis for w. □

Remark. Faigle [1979] considers certain accessible set-systems called *generating systems* and proves that the best-in greedy algorithm optimizes certain linear objective functions over them. While his systems are not necessarily greedoids, those feasible subsets of his "generating systems" which come up in a greedy basis do form a greedoid. Based on this, it is easy to derive Faigle's result from Theorem 3.2. For a more detailed discussion of the relationship between greedoids and Faigle's structures, see Korte and Lovász [1983b].

4. Special greedoids and general linear objective functions. We now invert the question of the last section and ask how much we have to restrict greedoids such that a greedy basis for any arbitrary linear objective function is optimal. The next theorem gives necessary and sufficient conditions for the worst-out greedy.

THEOREM 4.1. *For a greedoid $(E, \mathscr{F})$ the following statements are equivalent:*

(1) *Let B_1, B_2 be bases of $(E, \mathscr{F})$; for every $x \in B_1 - B_2$ there exists a $y \in B_2 - B_1$ such that $B_2 \cup x - y \in \mathscr{F}$.*

(2) *The hereditary closure $\mathscr{M}$ of $\mathscr{F}$ is a matroid $(E, \mathscr{M})$.*

(3) *β is submodular.*

(4) *For every linear objective function w a worst-out greedy basis is optimal.*

Proof. (1)⇔(2) is known from matroid theory.

(2)⇒(3): It suffices to show that β is the rank function of $(E, \mathscr{M})$. Let $X \subseteq E$; then

$$\beta(X) = \max\{|B \cap X|: B \in \mathscr{F}\} = \max\{|U|: U \subseteq X, U \in \mathscr{M}\},$$

since $\mathscr{M}$ is the hereditary closure.

(2)⇒(4): The spanning sets for $\mathscr{F}$ and $\mathscr{M}$ are the same, and so the worst-out greedy basis for $\mathscr{F}$ and $\mathscr{M}$ are the same. We know from matroid theory that the worst-out greedy bases are optimal for $\mathscr{M}$.

(3)⇒(2): Trivially, β has the unit increase property. Hence β is a matroid rank function. But $X \in \mathscr{M}$ iff $\beta(X) = |X|$. So $(E, \mathscr{M})$ is the matroid determined by β.

(4)⇒(2): Let $\mathscr{M}^* := \{X \subseteq E$: there exists a basis B with $B \cap X = \varnothing\}$. Then a worst-out greedy bases for $\mathscr{F}$ is optimal iff a best-in greedy basis for $\mathscr{M}^*$ is optimal. But this is the case iff $(E, \mathscr{M}^*)$ is a matroid which is equivalent to the fact that $(E, \mathscr{M})$ is a matroid. □

Remarks. 1. Condition (1) is not enough to guarantee the optimality of the best-in greedy: Let $E=\{a,b,c\}$ and $\mathscr{F}=\{\emptyset,\{a\},\{b\},\{a,b\},\{b,c\}\}$. The greedoid $(E,\mathscr{F})$ satisfies (1). However, with $w(a)=1$, $w(b)=0$, $w(c)=M\gg 1$, the best-in greedy basis is $\{a,b\}$ with weight 1, while the optimal basis is $\{b,c\}$ with weight M.

2. Let $\sigma_{\mathscr{M}}$ denote the matroid closure; then we have for greedoids with condition (1) that $\sigma_{\mathscr{M}}(A)=\mu(A)$ for $A\in\mathscr{F}$. In fact, $y\in\mu(A)$ iff $y\in A$ or $A\subseteq B$, $B\in\mathscr{F}$ implies $y\notin B$ which is equivalent to $y\in\sigma_{\mathscr{M}}(A)$.

The following theorem gives optimality conditions for the best-in greedy, which are of the same kind, but more restrictive.

THEOREM 4.2. *For a greedoid* $(E,\mathscr{F})$ *the following statements are equivalent*:

(1) *Let* $A\in\mathscr{F}$, $B\supseteq A$, *be a basis of* $(E,\mathscr{F})$ *and let* $x\in E-B$ *and* $A\cup\{x\}\in\mathscr{F}$. *Then there exists a* $y\in B-A$ *with* $A\cup y\in\mathscr{F}$ *such that* $B\cup x-y\in\mathscr{F}$ (*strong exchange property*).

(2) *The hereditary closure* $\mathscr{M}$ *of* $\mathscr{F}$ *is a matroid* $(E,\mathscr{M})$ *and every set which is closed in* $\mathscr{F}$ ($\mathscr{F}$*-closed*) *is also closed in* $\mathscr{M}$ ($\mathscr{M}$*-closed*).

(3) *For every linear objective function* w *a* (*best-in*) *greedy basis is optimal.*

Proof. (1)$\Rightarrow$(3). Let $w: E\to\mathbb{R}$ be any objective function, B an optimum basis, and $a_1, a_2, \cdots, a_r$ a best-in greedy basis, chosen in this order. Let $a_1,\cdots,a_k\in B$, but $a_{k+1}\notin B$ and choose B so that k is maximal. Let $A:=\{a_1,\cdots,a_k\}$. By (1), there exists a $y\in B-A$ such that $A\cup y\in\mathscr{F}$ and $B\cup a_{k+1}-y\in\mathscr{F}$. By greediness, $w(y)\leqq w(a_{k+1})$ and so $w(B\cup a_{k+1}-y)\geqq w(B)$. Since B is optimal, we have $w(B\cup a_{k+1}-y)=w(B)$ and so $B\cup a_{k+1}-y$ is also optimal, which contradicts the maximality of k.

(3)$\Rightarrow$(2). Let $X, Y\in\mathscr{M}$, $|X|<|Y|$. Let $0<t<1$ and define $w(x)=1$ if $x\in X$, $w(x)=t$ if $X\in Y-X$, and $w(x)=0$ otherwise. The set of greedy bases is independent of the value of t. If $t\approx 0$, then every optimal basis must contain X. Hence every greedy basis must contain X. But if $t>|X-Y|/|Y-X|$ then there is a basis containing Y, and hence the maximal objective value is greater or equal to $t|Y-X|+|Y\cap X|>|X|$. So a greedy basis must contain some element $y\in Y-X$ (besides X). Then $X\cup y\in\mathscr{M}$.

Thus, we know that $(E,\mathscr{M})$ is a matroid. It remains to show that every $\mathscr{F}$-closed set is also $\mathscr{M}$-closed. Let U be any $\mathscr{F}$-closed set, A an $\mathscr{F}$-basis of U, and extend A to an $\mathscr{M}$-basis A' of U. Let $v\in E-U$. Consider the objective function $w(x)=1$ if $x\in A'$ and $w(x)=0$ otherwise. Then there exists a basis containing A', and so every optimal basis contains A'. Of course, every best-in greedy basis also contains A'. But there must be also a greedy basis B starting with $A\cup v$, and so $(A\cup v)\cup A'=A'\cup v\subseteq B$. Thus $A'\cup v\in\mathscr{M}$ and so $v\notin\sigma_{\mathscr{M}}(A')$, ($\mathscr{M}$-closure of A'). This holds for all $v\in E-U$, so $\sigma_{\mathscr{M}}(A')\subseteq U$. But A' is an $\mathscr{M}$-basis of U, so $\sigma_{\mathscr{M}}(A')=U$ and so U is $\mathscr{M}$-closed.

(2)$\Rightarrow$(1). Consider $\sigma_{\mathscr{F}}(A)$, ($\mathscr{F}$-closure of A); by hypothesis $\sigma_{\mathscr{F}}(A)$ is also $\mathscr{M}$-closed. $B\cup x$ has a unique (fundamental) $\mathscr{M}$-circuit C. We have $x\in C-\sigma_{\mathscr{F}}(A)$, but since $\sigma_{\mathscr{F}}(A)$ is $\mathscr{M}$-closed, it follows that $|C-\sigma_{\mathscr{F}}(A)|\geqq 2$. Let $y\in C-\sigma_{\mathscr{F}}(A)-x$. Then $A\cup y\in\mathscr{F}$ and $B\cup x-y\in\mathscr{M}$, but $B\cup x-y$ is a basis of $\mathscr{M}$, and so a basis of $\mathscr{F}$. □

Remark. Condition (1) of Theorem 4.2 was independently observed by Goetschel [1983].

5. Slimmed matroids. It is a natural question to ask what greedoids satisfy the conditions of Theorems 4.2 and 4.1. Of course, matroids and trivially also all full greedoids do so. A nontrivial class are undirected branching greedoids. In Korte and Lovász [1982a] we have described a *search* or *directed branching greedoid* $(E,\mathscr{F})$ by a directed graph G and a root $r\in V(G)$. Let $E=E(G)$ and let $\mathscr{F}$ be the set of arc-sets of all arborescences in G rooted at r. The bases of $(E,\mathscr{F})$ are maximal branchings in G. In contrast, the *undirected branching greedoid* contains as feasible sets all cycle-free connected subgraphs of G which contain r. It is easy to see that this greedoid satisfies condition (2) of Theorem 4.2. (The directed branching greedoid does not.)

On the other hand the conditions of Theorem 4.1 give rise to general constructions of greedoids from a given matroid, whose set of bases is the same, but the feasible set is slimmed. In the following we will introduce some construction principles of slimming a matroid.

Given a matroid $(E, \mathcal{M})$ we call a greedoid $(E, \mathcal{F})$ a *slimming* of the matroid $(E, \mathcal{M})$ if $\mathcal{F} \subseteq \mathcal{M}$ and all bases of $\mathcal{M}$ remain bases of $\mathcal{F}$. The undirected branching greedoid is a slimming of the graphic matroid, actually an intersection of the graphic matroid with the *line search greedoid*, which is a shelling structure defined on the same graph G where $\mathcal{F}$ is the collection of all edge-sets which are connected and contain r (cf. Korte and Lovász [1982a]).

The next theorem describes the first slimming procedure.

THEOREM 5.1. *Let $(E, \mathcal{M})$ be a matroid with rank function $r_{\mathcal{M}}$ and $r_{\mathcal{M}}(E) = k$. Let $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_{k-1} \subseteq E$ such that $r_{\mathcal{M}}(E - A_i) \leqq k - i$. Define*

$$\mathcal{F} := \{X \in \mathcal{M}\colon |X \cap A_i| \geqq i \text{ for } 1 \leqq i \leqq |X|\}.$$

Then $(E, \mathcal{F})$ is a greedoid and a slimming of $(E, \mathcal{M})$.

Proof. We first show that $(E, \mathcal{F})$ is a greedoid. To prove (M3′) we take $X, Y \in \mathcal{F}$ with $|X| = |Y| + 1$. Then there exists an $x \in X - Y$ such that $Y \cup x \in \mathcal{M}$. But $|Y \cap A_i| \geqq i$ and hence $|(Y \cup x) \cap A_i| \geqq i$ for $1 \leqq i \leqq |Y|$ as $Y \in \mathcal{F}$. Further, $|X \cap A_{|Y|+1}| \geqq |Y| + 1 = |X|$ since $X \in \mathcal{F}$ and so $X \subseteq A_{|Y|+1}$, in particular $x \in A_{|Y|+1}$. Hence

$$|(Y \cup x) \cap A_{|Y|+1}| \geqq 1 + |Y \cap A_{|Y|+1}| \geqq 1 + |Y \cap A_{|Y|}| \geqq 1 + |Y|.$$

So $Y \cup x \in \mathcal{F}$.

Further, $\mathcal{F}$ is accessible. For, let $X \in \mathcal{F}$, and let i be the least index such that $X \subseteq A_i$. Since $X \in \mathcal{F}$, we have $i \leqq |X|$. Let $x \in X \cap (A_i - A_{i-1})$. Then $X - x \in \mathcal{M}$ and $|(X - x) \cap A_j| = |X - x| \geqq j$ if $j \geqq i$ and $|(X - z) \cap A_j| = |X \cap A_j| \geqq |X| > |X - x|$ if $j < i$. Hence $X - x \in \mathcal{F}$.

It remains to prove that $\mathcal{F}$ contains all bases of $\mathcal{M}$. Let B be a basis of $\mathcal{M}$; then $|B \cap A_i| = k - |B \cap (E - A_i)| \geqq k - r_{\mathcal{M}}(E - A_i) \geqq i$. □

Remarks. 1. The rank function $r_{\mathcal{F}}$ of $\mathcal{F}$ can be obtained by the following formula:

$$r_{\mathcal{F}}(X) := \max\{i\colon r_{\mathcal{M}}(X \cap A_j) \geqq j \text{ for } 1 \leqq j \leqq i\}.$$

2. It can be easily verified that the family

$$\mathcal{F}_0 = \{X \subseteq E\colon |X \cap A_i| \geqq i \text{ for } 1 \leqq i \leqq |X|\}$$

defines a shelling structure. Hence $\mathcal{F} := \mathcal{M} \cap \mathcal{F}_0$ is the intersection of a matroid with a shelling structure, and therefore an interval greedoid.

Another slimming procedure is given by

THEOREM 5.2. *Let $(E, \mathcal{M})$ be a matroid, $(E, \mathcal{F})$ a greedoid and suppose that the following hold:*

(1) *For $X, Y \in \mathcal{M}$ such that $\sigma_{\mathcal{M}}(X) = \sigma_{\mathcal{M}}(Y)$, we have $X \in \mathcal{F}$ iff $Y \in \mathcal{F}$.*

(2) *All (or equivalently at least one) bases of $\mathcal{M}$ are in $\mathcal{F}$.*

Then $(E, \mathcal{M} \cap \mathcal{F})$ is a greedoid which is a slimming of $\mathcal{M}$.

Proof. We show (M3). Suppose $X, Y \in \mathcal{M} \cap \mathcal{F}$, $|X| > |Y|$. Extend $\sigma_{\mathcal{M}}(Y) \cap X$ to an $\mathcal{M}$-basis X_1 of $\sigma_{\mathcal{M}}(Y)$. Then $\sigma_{\mathcal{M}}(X_1) = \sigma_{\mathcal{M}}(Y)$ and so by (1), $X_1 \in \mathcal{F}$. Since $|X_1| < |X|$, there exists a $x \in X - X_1$ such that $X_1 \cup x \in \mathcal{F}$. But $x \notin \sigma_{\mathcal{M}}(Y) = \sigma_{\mathcal{M}}(X_1)$ since $X \cap \sigma_{\mathcal{M}}(Y) \subseteq X_1$, but $x \notin X_1$. Hence $X_1 \cup x \in \mathcal{M}$ and so $X_1 \cup x \in \mathcal{M} \cap \mathcal{F}$. But $\sigma_{\mathcal{M}}(X_1 \cup x) = \sigma_{\mathcal{M}}(Y \cup x)$ and so $Y \cup x \in \mathcal{M} \cap \mathcal{F}$ by (1). □

The next theorem gives a further slimming construction.

THEOREM 5.3. (a) *Let $(E, \mathcal{M})$ be a matroid, $\mathcal{G}$ an accessible family of flats in $\mathcal{M}$, closed under union in the geometric lattice of $(E, \mathcal{M})$. Let*

$$\mathcal{F}_0 := \{X \in \mathcal{M} : \sigma_{\mathcal{M}}(X) \in \mathcal{G}\}$$

and let $\mathcal{F}$ be the accessible kernel of $\mathcal{F}_0$. Then $(E, \mathcal{F})$ is a greedoid.

(b) *Moreover, $(E, \mathcal{F})$ is a slimming of $(E, \mathcal{M})$ iff the following holds: for every $F \in \mathcal{G}$ and $F_1, \cdots, F_t \notin \mathcal{G}$ such that $F_1, \cdots, F_t$ cover F in the lattice, we have that $F_1 \cup \cdots \cup F_t$ is nonspanning in $(E, \mathcal{M})$.*

Proof. (a) We show (M3): Let $X, Y \in \mathcal{F}$, $|X| > |Y|$. By accessibility, $X = \{x_1, \cdots, x_m\}$ such that $\{x_1, \cdots, x_i\} \in \mathcal{F}$ for all $1 \leqq i \leqq m$. Let i be the first index with $x_i \notin \sigma_{\mathcal{M}}(Y)$. Then $Y \cup x_i \in \mathcal{M}$. Furthermore $\sigma_{\mathcal{M}}(Y \cup x_i) = \sigma_{\mathcal{M}}(\sigma_{\mathcal{M}}(Y) \cup \sigma_{\mathcal{M}}(x_1, \cdots, x_i)) \in \mathcal{G}$. Hence $Y \cup x_i \in \mathcal{F}$.

(b) I. By accessibility of $\mathcal{F}$, there exists a sequence of flats $B_0 \subset B_1 \subset \cdots \subset B_m \in \mathcal{F}$ such that $B_i \in \mathcal{G}$ and $r(B_i) = i$. Let $b_i \in B_i - B_{i-1}$; then $\{b_1, \cdots, b_m\} \in \mathcal{F}$. If $F_1 \cup \cdots \cup F_t$ is spanning, we can extend $\{b_1, \cdots, b_m\}$ to a basis A of $(E, \mathcal{F})$. Let $a \in A - \{b_1, \cdots, b_m\}$ such that $\{b_1, \cdots, b_m, a\} \in \mathcal{F}$. Then $a \in F_v$ for some $1 \leqq v \leqq t$, and so $\sigma_{\mathcal{M}}(\{b_1, \cdots, b_m, a\}) = F_v$. But $\{b_1, \cdots, b_m, a\} \in \mathcal{F}$ implies $\sigma_{\mathcal{M}}(\{b_1, \cdots, b_m, a\}) \in \mathcal{G}$, a contradiction.

II. Let b be any basis of $\mathcal{M}$. Consider a maximal subset $A \subseteq B$ with $A \in \mathcal{F}$. We claim that $A = B$. Suppose not, and let $F = \sigma_{\mathcal{M}}(A)$, $B - A = \{b_1, \cdots, b_t\}$ and let $F_i = \sigma_{\mathcal{M}}(A \cup b_i)$. Then $\cup F_i$ is spanning in $(E, \mathcal{M})$, because $B \subseteq \cup F_i$. Thus, there exists an $F_i \in \mathcal{G}$. But then $A \cup b_i \in \mathcal{F}$, contradiction. □

Remark. If $(E, \mathcal{M})$ is the free matroid, then the construction of Theorem 5.3 gives every shelling structure $(E, \mathcal{F}_1)$ by letting $\mathcal{G} = \mathcal{F}_1$.

6. Oracle results. In this final section we mention briefly some negative results about greedoid optimization and greedoid recognition obtained by an oracle approach. We do not go into details of oracle techniques here. The reader is referred to similar approaches for independence systems and matroids in earlier papers (cf. Hausmann and Korte [1981] and Jensen and Korte [1982]). As in the case of matroids we assume that the greedoid $(E, \mathcal{F})$ is given by a *feasibility oracle*, i.e. a mapping $O: 2^E \to \{\text{Yes}, \text{No}\}$ which is defined for $X \subseteq E$ as $O(X) = \text{Yes}$ if $X \in \mathcal{F}$, $O(X) = \text{No}$ otherwise.

It is clear that a feasibility oracle uniquely determines the greedoid. Moreover, several questions concerning greedoids can be decided in polynomial time using the feasibility oracle: e.g. computing the rank or closure of a set, as well as the problems discussed in previous chapters. However, some other important questions cannot be decided by good algorithms. To formulate these negative results, we need the following definition.

A problem concerning greedoids given by a feasibility oracle is called NP-hard, if there is a special class of greedoids, with some "name" (encoding) for each member, such that the oracle can be realized by a polynomial-time algorithm for members of this class (polynomial in the length of the "name") and the problem is NP-hard already for members of this class.

THEOREM 6.1. *The problem of optimizing a linear objective function over the bases of an arbitrary greedoid given by a feasibility oracle is* NP*-hard.*

Proof. We consider the *k-truncation* of the directed or undirected branching greedoid $(E, \mathcal{F})$, i.e. the greedoid $(E, \mathcal{F}^{(k)})$ with $\mathcal{F}^{(k)} := \{X \subseteq E : X \in \mathcal{F} \text{ and } |X| \leqq k\}$. The problem of finding a maximum weighted branching of size less or equal to k includes the *Steiner problem*, which is known to be NP-hard. □

Remark. The problem of optimizing an arbitrary linear objective function over the feasible sets of a greedoid remains NP-hard even for shelling structures. In fact, this optimization problem for line search greedoids also contains the Steiner problem.

THEOREM 6.2. *There is no polynomial-time algorithm to decide whether a greedoid given by a feasibility oracle is a matroid.*

Proof. Consider the uniform matroid $(E, \mathcal{M})$ of rank $r=|E|/2$ and the greedoid $(E, \mathcal{F})$ with $\mathcal{F}:=\mathcal{M}-\{X\}$ where $|X|=r-1$. With the usual argument (cf. Hausmann and Korte [1981]) one can show that any feasibility oracle algorithm can not distinguish between $(E, \mathcal{M})$ and $(E, \mathcal{F})$ using only polynomially many calls on the feasibility oracle. □

COROLLARY 6.3. *There is no polynomial algorithm to recognize a closure feasible set for a greedoid given by a feasibility oracle, i.e. to decide membership in $\mathcal{C}$.*

Proof. It is an easy observation that a greedoid $(E, \mathcal{F})$ is a matroid iff $\{x\}\in\mathcal{C}$ for all $x\in E$. (To prove this one needs that $\mathcal{C}$ is closed under union.) Then apply Theorem 6.2. □

THEOREM 6.4. *There is no polynomial-time algorithm to decide whether a greedoid given by a feasibility oracle is normal.*

Proof. Let $(E, \mathcal{F})$ be a uniform matroid of rank $r=|E|/2$. Let $d\notin E$ and consider the greedoid $(E\cup\{d\}, \mathcal{F})$. Let $X\subseteq E$, $|X|=r-1$ and $\mathcal{F}'=\mathcal{F}\cup\{X\cup\{d\}\}$. Then it is easy to check that $(E\cup\{d\}, \mathcal{F}')$ is also a greedoid. By the usual argument again, no feasibility oracle algorithm can distinguish between $(E\cup\{d\}, \mathcal{F})$ and $(E\cup\{d\}, \mathcal{F}')$ in polynomial time. □

COROLLARY 6.5. *There is no polynomial-time algorithm to decide whether a given element is a dummy.*

COROLLARY 6.6. *There is no polynomial-time algorithm to recognize a rank-feasible set in a greedoid given by a feasibility oracle.*

Proof. Observe that $d\in E$ is a dummy iff $\{d\}\in\mathcal{R}-\mathcal{F}$.

THEOREM 6.7. *It is* NP-*hard to recognize for a greedoid a rank-feasible* (*or closure-feasible*) *set, i.e. to decide membership in $\mathcal{R}$* (*or in C*).

Proof. Let G be a digraph, $E=E(G)$, $V(G)=\{v_1,\cdots,v_n\}$. We call an arc e a *shortcut* in G if there exists a dipath in $G-e$ from the tail of the head of e. Let

$$\mathcal{F}:=\{e_1,\cdots,e_k\colon e_i \text{ is not a shortcut in } G-\{e_1,\cdots,e_{i-1}\}\}.$$

Then $(E, \mathcal{F})$ is a shelling structure, which we call the *digraph shortcut greedoid.* This greedoid was first observed by A. Björner [1983]. It can be also represented as a *convex shelling structure* (cf. Korte and Lovász [1983a]) in $\mathbb{R}^n$ of the following set of points $\{0, e_{ij}\}$ where 0 is the 0-vector and e_{ij} is a 0, ±1 incidence vector of the arc $e=(v_i, v_j)$ which has a -1 at the ith component, a $+1$ at the jth component and 0's elsewhere. Then $\{0\}\in\mathcal{F}$ iff G is acyclic. We take the k-truncation of this greedoid. Then $\{0\}\notin\mathcal{R}$ iff the feedback number of G is $\leqq k-1$, but this is a well-known NP-hard problem. □

This shortcut greedoid is an interval greedoid, and thus $\mathcal{R}=\mathcal{C}$. So the assertion concerning closure feasibility follows in the same way.

Remark. The test for membership in $\mathcal{R}$ is of course a special case of optimizing a linear objective function over $(E, \mathcal{F})$.

REFERENCES

A. BJÖRNER [1983], *On matroids, groups and exchange languages*, preprint, Dept. Mathematics, Univ. of Stockholm, 1983, to appear in Matroid Theory and Its Applications, L. Lovász and A. Recski, eds., Conference Proceedings, Szeged, September 1982. Colloquia Mathematica Societatis Janós Bolyai, North-Holland, Amsterdam/Oxford/New York.

U. FAIGLE [1979], *The greedy algorithm for partially ordered sets*, Discrete Math., 28 (1979), pp. 153–159.

R. H. GOETSCHEL [1983], *Linear objective functions and certain classes of greedoids*, preprint, Univ. Idaho, Moscow, 1983.

D. HAUSMANN AND B. KORTE [1981], *Algorithmic versus axiomatic definitions of matroids*, Math. Programming Study, 14 (1981), pp. 98–111.

P. M. JENSEN AND B. KORTE [1982], *Complexity of matroid property algorithms*, SIAM J. Comput., 11 (1982), pp. 184–190.

B. KORTE AND L. LOVÁSZ [1981], *Mathematical structures underlying greedy algorithms*, in Fundamentals of Computation Theory, F. Gécseg, ed., Lecture Notes in Computer Sciences 117, Springer, Berlin/Heidelberg/New York, 1981, pp. 205–209.

———, [1982a], *Greedoids, a structural framework for the greedy algorithm*, Report No. 82230-OR, Institute of Operations Research, Univ. Bonn, 1982; to appear in Progress in Combinatorial Optimization, W. R. Pulleyblank, ed., Proceedings of the Silver Jubilee Conference on Combinatorics, Waterloo, June 1982, Academic Press, London/New York, San Francisco.

———, [1982b], *Structural properties of greedoids*, Report No. 82242-OR, Institute of Operations Research, Univ. Bonn, 1982; Combinatorica 3, 3–4, to appear.

———, [1983a], *Shelling structures, convexity and a happy end*, Report 83274-OR, Institute of Operations Research, Univ. Bonn, 1983.

———, [1983b], *Posets, matroids, and greedoids*, Report No. 83278-OR, Institute of Operations Research, Univ. Bonn, 1983, to appear in Matroid Theory and Its Applications, L. Lovász and A. Recski, eds., Conference Proceedings, Szeged, September 1982, Colloquia Mathematica Societatis Janós Bolyai, North-Holland, Amsterdam/Oxford/New York.

D. J. A. WELSH [1976], *Matroid Theory*, Academic Press, London, New York, San Francisco, 1976.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

010

ON SOME PROPERTIES OF THE STRUCTION OF A GRAPH*

DOMINIQUE DE WERRA†

Abstract. The struction is defined as an operation which associates with a graph G with stability number $\alpha(G)$ another graph G' with stability number $\alpha(G')=\alpha(G)-1$. Properties of the graph G' are related to those of G. Namely, one exhibits some classes of graphs which are closed with respect to the struction; i.e., if G is in class C, then so is G'.

One shows that for a fixed k, the class of graphs containing no induced P_k (path on k nodes) is closed. So is the class of graphs containing no induced P_k and no induced C_k (cycle on k nodes). One also shows that the class of graphs G with $\alpha(G)=\theta(G)$ is closed. (Here $\theta(G)$ is the minimum number of cliques covering the nodes of G.)

1. Introduction. The stability number of a graph G is the maximum number of pairwise nonadjacent nodes which can be found in G.

An approach to the problem of determining the stability under $\alpha(G)$ of a graph G might be suggested by the following: a construction has been given for associating with any graph G with (unknown) stability number $\alpha(G)$ another graph G' with stability number $\alpha(G)-1$ [2]. Such an operation which might be considered as a STability number RedUCTION has been called a *struction* ([5], [6]). Our purpose in this note is to study some properties of the struction; more precisely, we shall try to relate a few characteristic parameters of a graph G' obtained by a struction to the corresponding parameters of the original graph G.

We shall compare the clique numbers of G and G' (i.e. the smallest number of cliques covering the nodes) and also the length of the longest induced cycle.

In fact, some classes of graphs which are closed with respect to the struction will be exhibited: a class F will be *closed* (with respect to the struction) if $G\in F$ implies $G'\in F$ where G' is obtained from G by a struction. For all graph-theoretical terms not defined here, the reader is referred to [3].

We shall sometimes write $[i, j]$ to indicate that nodes i and j are linked in a graph. $N(x)$ will represent the set of neighbours of node x.

2. The struction. A construction associating with any graph G another graph G' with $\alpha(G')=\alpha(G)-1$ has been given in [2]; it was in fact first derived by using pseudo-Boolean methods.

Such a construction may be used for determining the stability number of a graph G; we repeatedly apply the struction, thereby obtaining a sequence $G, G', G'', \cdots$ of graphs; we shall stop as soon as we get a graph $G^{(k)}$, the stability number of which can be determined easily; then if $\alpha(G^{(k)})=p$, we have $\alpha(G)=k+p$.

We may stop when $G^{(k)}$ is a clique, in which case $\alpha(G)=k+1$, or for instance, when $G^{(k)}$ is a graph containing no induced P_4 (induced path on 4 nodes), since, as noted in [1], the stability number of P_4-free graphs is particularly easy to obtain. A drawback of this approach is that the number of nodes in the graphs $G, G', G'', \cdots$ may increase in the general case, although rather encouraging computational results for random graphs have been reported in [2].

* Received by the editors July 14, 1983. This research was carried out while the author was visiting the University of Waterloo in the summer of 1983. The support of the Natural Sciences and Engineering Research Council is gratefully acknowledged. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Département de Mathématiques, Ecole Polytechnique Fedérale de Lausanne, CH-105 Lausanne-Ecublens, Switzerland.

For some classes of graphs, one has derived a modified struction which has the property that the number of nodes in the graphs $G, G', G'', \cdots$ does not increase. In this case, the struction gives a polynomial algorithm for obtaining $\alpha(G)$. It is the case for the CN-free graphs, i.e. the graphs containing no induced Claw (unique graph with degree sequence $(3,1,1,1)$) and no induced Net (unique claw-free graph with degree sequence $(3,3,3,1,1,1)$) [6]. A specialized version of the struction has also been developed for a subclass of CN-free graphs [5]. An entirely different approach leading to a polynomial algorithm has been suggested for the more general class of claw-free graphs [7]. We shall now, for the sake of completeness, give the general formulation of the struction developed in [2].

The struction: $G \to G'$.

a) Let a_0 be an arbitrary node of G and let its neighbourhood be $N(a_0) = \{a_1, \cdots, a_p\}$, while the other nodes of G are $a_{p+1}, \cdots, a_n$.

b) The node set of G' will consist of $a_{p+1}, \cdots, a_n$ as well as of a set of "new" nodes a_{ij} with $i < j \leqq p$ associated to all pairs i, j of nonadjacent nodes a_i, a_j in $N(a_0)$. We shall represent this set of new nodes as being partitioned into layers $L_i = \{a_{ij_1}, a_{ij_2}, \cdots, a_{ij_k}\}$ consisting of all new nodes a_{ij} having i as first index.

c) The edge set of G' will be defined as consisting of

(c1) all the edges of the subgraph of G induced by $a_{p+1}, \cdots, a_n$;

(c2) all the edges linking new nodes $a_{i_1j_1}, a_{i_2j_2}$ with $i_1 \neq i_2$ belonging to two different layers;

(c3) edges linking two nodes a_{ij_1}, a_{ij_2} in the same layer if a_{j_1} and a_{j_2} were linked in G;

(c4) edges linking a new node a_{ij} to a node a_r $(r \geqq p+1)$ if a_r was linked to a_i or a_j in G.

We shall say that the struction is centered at node a_0; furthermore, we shall replace nodes $a_0, a_1, \cdots, a_n$ by $0, 1, 2, \cdots, n$ whenever no confusion is possible; the same will be done for nodes a_{ij} which will become (i, j).

3. The clique covering number of G'. We shall now relate some parameters of a graph G with the corresponding parameters of a graph G' obtained from G by a struction.

Let us denote by $\theta(G)$ the smallest number of cliques needed to cover the nodes of a graph G; in general we have $\theta(G) \geqq \alpha(G)$. In a perfect graph G, $\theta(G') = \alpha(G')$ for any subgraph G' of G [3].

PROPOSITION 3.1. *Let G' be obtained from G by a struction. Then $\theta(G') \leqq \theta(G) - 1$.*

Proof. Let $C = (K_1, \cdots, K_t)$ be a covering of the nodes of G by cliques with $t = \theta(G)$; we shall here for reasons of convenience identify a clique with its node set. We may assume $K_i \cap K_j = \emptyset$ if $i \neq j$.

For each K_i let $R_i = K_i \cap R$ where $R = X - (N(a_0) \cup \{a_0\})$ is the set of nodes of $G = (X, U)$ which will become old nodes in G'.

We may assume $R_i = \emptyset$ for $i \leqq r$ and $R_i \neq \emptyset$ for $i > r$. Notice that $r \geqq 1$ since in any covering C of G there is at least one clique $K \subseteq N(a_0) \cup \{a_0\}$ (namely the clique covering node a_0).

For $i = 2, 3, \cdots, t$ we define the following sets of nodes in G'.

$$K'_1 = \{(p, j) \mid p < j, p, j \in N(a_0), j \in K_i\}$$
$$\cup \{(j, p) \mid p > j, p, j \in N(a_0), j \in K_i, p \in K_1\} \cup \{R_i\}.$$

CLAIM 1. *K'_1 is a clique of G'.*

Since in G we have $[j, r]$ for all $j \in K_i \cap N(a_0)$ and all $r \in R_i$, in G' every new node (p, j) (or (j, p)) is linked to every r in R_i.

Also two new nodes (p, r), (q, s) are linked if $p \neq q$ from the construction of G'. Now if we have two nodes of the form (p, j), (p, k) with $p \in K_l$ for some $l \neq i$, then we must have $j, k \in K_i$; hence (p, j) and (p, k) are linked in G'. If we have two nodes of the form (j, k), (j, m) with $j \in K_i$, then by construction of K'_i, k and m are in K_1, hence (j, k), (j, m) are linked in G'.

CLAIM 2. *Every node of G' is covered by some K'_i.*

Since all old nodes are included in some K'_i (with $i \geqq r+1$), we only have to examine the case of new nodes.

Consider a new node (i, j) in G'; let K_u (resp. K_v) be the clique of C containing i (resp. j). If $v > 1$, then by the above construction, $(i, j) \in K'_v$; if $v = 1$, then $u > 1$ (we cannot have $u = v$ since i and j are not linked in G) and by the construction $(i, j) \in K'_u$.

This ends the proof of Claim 2.

We now have constructed a covering of the nodes of G' by cliques $K'_2, K'_3, \cdots, K'_t$. By keeping only the nonempty cliques K'_i we obtain a covering C' of the nodes of G' by cliques with $|C'| \leqq t-1 = |C|-1$. Since C was chosen to be a minimal covering, we obtain

$$\theta(G') \leqq |C'| \leqq t-1 = |C|-1 = \theta(G)-1. \qquad \square$$

We shall say that a class F of graphs is *closed* (for the struction) if $G \in F$ implies that any G' obtained from G by a struction is also in F.

PROPOSITION 3.2. *The class $C_{\alpha=\theta}$ of graphs G with $\alpha(G) = \theta(G)$ is closed for the struction.*

Proof. Let G be a graph satisfying $\alpha(G) = \theta(G)$; if G' is obtained by a struction, we have $\alpha(G)-1 = \alpha(G') \leqq \theta(G') \leqq \theta(G)-1 = \alpha(G)-1$. Hence $\alpha(G') = \theta(G')$ and G' is in $C_{\alpha=\theta}$. $\square$

Remark 3.1. We may have $\theta(G') \leqq \theta(G)-2$; consider for instance for G a pentagon; we obtain a triangle for G' and hence $\theta(G) = 3$ and $\theta(G') = 1$. Furthermore, one should observe that the class of perfect graphs is in $C_{\alpha=\theta}$; however, if one takes an arbitrary node a_0 as the centre of a struction, one may obtain a graph G' which is not perfect.

Remark 3.2. Let $\chi(G)$ be the chromatic number of G (i.e. the smallest number of colours needed to colour the nodes of G so that adjacent nodes have different colours); then $\chi(G) = \theta(\bar{G})$ where $\bar{G}$ is the complement of G.

So if we apply the struction to the complement $\bar{G}$ of graph G, we obtain a graph $\bar{G}'$, the complement of which, G^*, satisfies

$$\chi(G^*) = \theta((\bar{G})') \leqq \theta(\bar{G})-1 = \chi(G)-1.$$

This means that we may with the struction reduce the chromatic number of a graph G satisfying $\chi(G) = \omega(G)$ where $\omega(G) = \alpha(\bar{G})$ is the maximum cardinality of a clique.

According to Proposition 3.2, the new graph G^* will still satisfy $\chi(G^*) = \omega(G^*)$.

4. Some other closed classes. We shall now examine several classes of graphs which are closed with respect to the struction. These classes will be characterized by forbidden subgraphs H. Given two graphs $H = (V_H, E_H)$ and $G = (V, E)$, we shall say that G contains an induced H if there is in G an induced subgraph isomorphic to H.

In the proofs we shall need the following.

LEMMA 4.1. *Let H be a connected graph, let G be a graph and let G' be obtained from G by a struction. Assume G' contains an induced H on node set V and let N be the set of new nodes in V. If the following conditions hold*

(1) *for every v in $V-N$ there is an α in N with $[\overline{v, \alpha}]$;*

(2) *all new nodes in N are in the same layer;*

then G contains an induced H.

Proof. If $|N|=1$, then from (1) we see that no node in $V-N$ can be linked to N; this means that the connected subgraph H' isomorphic to H is induced by $V-N$ and hence it is also an induced subgraph of G.

We may now assume $|N| \geqq 2$ and let $N=\{\alpha_1, \cdots, \alpha_r\}$ with $\alpha_i=(1, i+1)$ for $i=1, \cdots, r$. From (1) we have for each v in $V-N$ an $\alpha_{i_v}=(1, i_v+1)$ in N with $[\overline{v, \alpha_{i_v}}]$; this means that $[\overline{v, 1}]$ and $[\overline{v, i_v+1}]$ in G. Hence for every $v \in V-N$ and every $\alpha_i \in N$ we have $[v, \alpha_i]$ in H' iff $[v, i+1]$ in G. Similarly $[\alpha_i, \alpha_j]$ in H' iff $[i+1, j+1]$ in G for $\alpha_i, \alpha_j \in N$. Furthermore since $[u, v]$ holds in H' for $u, v \in V-N$ iff it holds in G, the subgraph of G induced by $(V-N) \cup \{2, 3, \cdots, r+1\}$ is isomorphic to H. □

If P_k denotes an elementary chain on k nodes, we call a graph *P_k-free* if it does not contain an induced chain on k nodes.

P_3-free graphs are unions of node disjoint cliques and their stability number is trivially obtained. P_4-free graphs have been extensively studied (see for instance, [1] where these graphs are called *cographs*). Such graphs are perfect; there is also a simple way of determining their stability number when their cotree has been constructed (see [1]). We shall in fact be interested in graphs which are P_k-free for $k \geqq 5$. The following result is however valid for any fixed $k \geqq 4$.

PROPOSITION 4.1. *Let $k \geqq 4$ be a given integer. The class of P_k-free graphs is closed with respect to the struction.*

Proof of Proposition 4.1. Let us assume that G' contains an induced P_k on nodes $d_1, d_2, \cdots, d_k$; let N be the set of new nodes in $V=\{d_1, \cdots, d_k\}$. $|N| \geqq 1$; otherwise the induced P_k is in G.

Conditions (1) and (2) of Lemma 4.1 are satisfied when $N \supseteq \{d_i, d_{i+r}\}$ for $r \geqq 3$. So we are left with the following cases.

Case 1. $N=\{d_i\}$ with $d_i=(1, j)$.

Then in G we have $[1, d_{i-1}]$ and/or $[j, d_{i-1}]$ as well as $[1, d_{i+1}]$ and/or $[j, d_{i+1}]$. If we have $[d_{i-1}, r]$ and $[d_{i+1}, r]$ for some $r \in \{1, j\}$, we replace d_i by r in the P_k and we get an induced P_k in G. Otherwise, we replace d_i by $1, 0, j$ and we get an induced P_{k+2} in G.

If $i=1$, we replace d_1 by r in both cases.

Case 2. $N=\{d_i, d_{i+1}\}$ with $d_i=(1, j)$, $d_{i+1}=(2, m)$ and $j \neq m$, $j \neq 2$.

We have $[d_{i-1}, r]$ for some $r \in \{1, j\}$ and $[\overline{d_{i+2}, r}]$ for $r=1, j$ and $[d_{i+2}, s]$ for some $s \in \{2, m\}$ and $[\overline{d_{i-1}, s}]$ for $s=2, m$. If $[r, s]$, then we replace d_i, d_{i+1} by r, s and we get an induced P_k in G. If $[\overline{r, s}]$ for any r, s with $[d_{i-1}, r]$ and $[d_{i+2}, s]$, then we replace d_i, d_{i+1} by r, 0, s and we get an induced P_{k+1} in G. The case where $i=1$ is dealt with similarly by replacing d_1, d_2 by 0, s in both cases.

Case 3. $N=\{d_i, d_{i+1}\}$ with $d_i=(1, j)$, $d_{i+1}=(2, j)$.

We have $[d_{i-1}, 1]$, $[\overline{d_{i-1}, s}]$ for $s=2, j$ and $[d_{i+2}, 2]$, $[\overline{d_{i+2}, r}]$ for $r=1, j$. So we replace d_i, d_{i+1} by 1, 0, 2 if $[\overline{1, 2}]$ or by 1, 2 otherwise. The case $i=1$ is dealt with in a similar way as in the previous cases.

Case 4. $N=\{d_i, d_{i+1}\}$ and $d_i=\{1, 2\}$, $d_{i+1}=\{2, j\}$.

Since we have $[d_{i-1}, 1]$, $[\overline{d_{i-1}, j}]$, $[\overline{d_{i+2}, 1}]$, $[d_{i+2}, j]$, we replace d_i, d_{i+1} by 1, 0, j if $[\overline{1, j}]$ and we get an induced P_{k+1} in G. If $[1, j]$ we replace d_i, d_{i+1} by $1, j$ and we get an induced P_k. The case $i=1$ is similar.

Case 5. $\{d_i, d_{i+1}, d_{i+2}\} \supseteq N \supseteq \{d_i, d_{i+2}\}$ with $d_i=(1, j)$, $d_{i+2}=(1, m)$.

Now we have $[d_{i-1}, j]$, $[\overline{d_{i-1}, m}]$, $[\overline{d_{i+3}, j}]$, $[d_{i+3}, m]$ and $[\overline{j, m}]$; so by replacing d_i, d_{i+1}, d_{i+2} by j, 0, m we get an induced P_k in G.

The case $i = 1$ is quite similar.

We have now examined all cases (the other cases can be obtained by reversing the ordering of nodes d_i in the induced P_k). Hence in all cases G contains an induced P_k, so all cases are impossible and G' does not contain any induced P_k. □

A subclass of P_k-free graphs will now be shown to be closed with respect to the struction.

Let C_k denote a (chordless) cycle with k nodes; a graph G will be C_k-free if it does not contain any induced C_k. Observe that if $k < l$, then G P_k-free implies G P_l-free, but G C_k-free does not imply G C_l-free.

The class of C_4-free P_4-free graphs has been studied by Golumbic [4]; it is the class of so called trivially perfect graphs. For these the stability number of any subgraph G' is equal to the number of maximal cliques in G'.

PROPOSITION 4.2. *For any $k \geqq 5$, the class of C_k-free and P_k-free graphs is closed under the struction.*

Proof. We only have to show that any G' obtained from a P_k-free and C_k-free graph G by a struction is C_k-free.

The proof is essentially the same as the one of Proposition 4.1 (the limit cases $d_i = d_1$ do not occur); the P_k are replaced by C_k and one should recall that if one concludes that G contains a C_{k+1} or a C_{k+2}, then G also contains a P_k. □

Let us denote by $\hat{g}(G)$ the maximum length of a (chordless) cycle in G. Then we have

PROPOSITION 4.3. *Let p be a positive number; then the class $C_{\hat{g},p}$ of graphs G defined by $\hat{g}(G) \leqq p$ is closed for the struction.*

Proof. We assume G' contains an induced C_k. By examining exactly the same cases as in the proof of Proposition 4.1, we arrive in all cases to the following conclusion: if G' contains an induced C_k, then G contains an induced C_l for some l satisfying $k \leqq l \leqq k+2$. So $\hat{g}(G) \geqq \hat{g}(G')$ and hence $\hat{g}(G) \leqq p$ implies $\hat{g}(G') \leqq p$. □

As a consequence of Proposition 4.3, the class of triangulated graphs is closed under the struction.

Remark 4.1. The *girth* $g(G)$ is usually defined as the minimum length of a (chordless) cycle in G. Since our graphs here have no loops and no multiple edges, we have $g(G) \geqq 3$ (we set $g(G) = \infty$ if G has no cycles). Then it is generally not true that $g(G') \leqq g(G)$; we may in fact have $g(G') > g(G)$, as can be seen easily.

Acknowledgments. The author would like to express his gratitude to P. L. Hammer and N. V. R. Mahadev for the stimulating discussions and suggestions about the struction.

REFERENCES

[1] D. G. CORNEIL, H. LERCHS AND L. STEWART BURLINGHAM, *Complement reducible graphs*, Discr. Appl. Math., 3 (1981), pp. 163–174.

[2] CH. EBENEGGER, P. L. HAMMER AND D. DE WERRA, *Pseudo-Boolean functions and stability of graphs*, Ann. Discr. Math., to appear.

[3] M. C. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[4] ———, *Trivially perfect graphs*, Discr. Math., 24 (1978), pp. 105–107.

[5] P. L. HAMMER, N. V. R. MAHADEV AND D. DE WERRA, *Stability in* CAN-*free graphs*, CORR 83/20, Univ. Waterloo, Waterloo, Ontario, Canada.

[6] ———, *The struction of a graph: application to* CN-*free graphs*, CORR 83/21, Univ. Waterloo, Waterloo, Ontario, Canada.

[7] N. SBIHI, *Algorithme de recherche d'un stable de cardinalité maximum dans un graphe sans étoile*, Discr. Math., 29 (1980), pp. 53–76.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

011

J': A NEW TRIANGULATION OF R^n

MICHAEL J. TODD*

Abstract. This paper introduces a new triangulation of R^n and two families of related triangulations. Our interest is primarily in the use of such triangulations in piecewise-linear homotopy algorithms for solving systems of nonlinear equations, and we provide both theoretical and computational evidence of the efficiency of the new triangulations for this purpose. However, the triangulations we propose may also be of independent interest.

AMS subject classifications. 65H10, 57Q15, 65D05

1. Introduction. This paper introduces a new triangulation of R^n and two families of related triangulations. Our interest is primarily in the use of such triangulations in piecewise-linear homotopy algorithms for solving systems of nonlinear equations—see, e.g., Allgower and Georg [1], Eaves [5], and Todd [15], [18]. However, the triangulations we propose may be of independent interest.

In three dimensions the new triangulation J' is similar to the A^*K_1 triangulation of van der Laan and Talman [8] and identical to a triangulation proposed by Buneman in a different context [4]. For dimensions greater than three it is very similar to the triangulation J_1. We will assume that the reader is familiar with Tucker's triangulation J_1 and Freudenthal's triangulation K_1—see, e.g., [15]. In § 2 we define J' and prove it a triangulation of R^n, i.e., a locally finite collection of n-simplices covering R^n such that any two intersect in a common face (perhaps empty). To do this we show that it can also be viewed as a polyhedral subdivision of R^n generated by a family of hyperplanes.

Let $\mathscr{H}$ be a family of hyperplanes that is the union of a finite number of families of evenly-spaced parallel hyperplanes. Then $\mathscr{H}$ divides R^n into polyhedral sets, and it is easy to see that the result is a polyhedral subdivision of R^n, i.e., a locally finite collection of n-polyhedra such that any two intersect in a common face. We say the polyhedral subdivision is generated by the family $\mathscr{H}$ of hyperplanes.

This technique also provides an easy proof that J_1 and K_1 are triangulations.

In § 3 we consider the special case when $n = 3$. Section 4 calculates various measures of the triangulation J'. While it does not subdivide a cube (or parallelopiped) into a small number of simplices, it appears superior to K_1 and dominates J_1 according to average directional (or surface) density.

Section 5 demonstrates how two families of triangulations, $J'_k(n)$ and $J''_k(n)$, of R^n are "induced" by J'. Various members of these families are suited to various piecewise-linear homotopy methods. Finally, in § 6 we give some computational experience with the new triangulation $J'_{n+1}(n+1)$ in a restart algorithm. A consistent improvement is observed.

Our notation is as follows. Subscripts of vectors denote coordinates while superscripts are used for sequences. The jth unit vector is denoted e^j and e is the vector of ones. We use $[a, b, \cdots, z]$ to denote the convex hull of the vectors $a, b, \cdots, z$.

* Received by the editors July 26, 1983. This research was partially supported by a fellowship from the Alfred P. Sloan Foundation and by the National Science Foundation under grant ENG82-15361. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853.

2. The triangulation J'. In this section we define the new triangulation J', giving descriptions of each simplex both by its vertices and by its facets. We also prove that J' is indeed a triangulation and state its pivot rules.

First we define J' and its simplices via their vertices.

DEFINITION 2.1. The set of vertices of J' is the set of vectors $v \in R^n$ with each component an integer, such that there is not precisely one even component nor precisely one odd component. Each simplex σ of J' is of the form $\sigma = j'(v, \pi, s) = [v^0, \cdots, v^n]$, where v is a vector each of whose components is an even integer, $\pi = (\pi(1), \cdots, \pi(n))$ is a permutation of $(1, \cdots, n)$ and s is a sign vector (each $s_j = \pm 1$) with $s_{\pi(1)} = s_{\pi(n)} = 1$. To define the vertices of σ, it is convenient to let $\tilde{e}^i$ denote $s_{\pi(i)} e^{\pi(i)}$, where e^j is the jth unit vector. Thus $\tilde{e}^1, \cdots, \tilde{e}^n$ are possibly permuted and reversed unit vectors. For $n \geqq 4$, we have

$$
\begin{aligned}
&v^0 = v, && v^j = v^{j-1} + \tilde{e}^j, \qquad 3 \leqq j \leqq n-2,\\
(2.1)\qquad &v^1 = v^0 + 2\tilde{e}^1, && v^{n-1} = v^{n-2} + \tilde{e}^{n-1} - \tilde{e}^n,\\
&v^2 = v^1 - \tilde{e}^1 + \tilde{e}^2, && v^n = v^{n-1} + 2\tilde{e}^n.
\end{aligned}
$$

For $n = 3$, the formulae for $j = 2$ and $j = n-1$ are combined to give $v^2 = v^1 - \tilde{e}^1 + \tilde{e}^2 - \tilde{e}^3$.

Note that the restrictions $s_{\pi(1)} = s_{\pi(n)} = 1$ are only present to give a one-to-one correspondence between simplices and their descriptions $j'(v, \pi, s)$. Without this restriction $v = v^0$ could be replaced by $v = v^1$ with $s_{\pi(1)} = -1$, and $s_{\pi(n)} = 1$ could be replaced by $s_{\pi(n)} = -1$. There are other ways to resolve the nonuniqueness—we could insist that $v_{\pi(1)}$ be a multiple of 4 and that $v^n_{\pi(n)}$ be one more than a multiple of 4, for instance, but our choice is simpler if less symmetrical than other possibilities.

We remark that we could also insist that each component of v be odd rather than even. The same simplex σ is of the form $j'(v', \pi', -s + 2\tilde{e}^1 + 2\tilde{e}^n)$ where $v' = v^{n-1}$ and $\pi' = (\pi(n), \cdots, \pi(1))$. Thus J' is invariant under permutations of coordinates, under reflections in coordinate hyperplanes and under translations by vectors with each component even or each component odd.

Note that each simplex of J' is the union of four simplices of J_1. Indeed, $\sigma = j'(v, \pi, s) \in J'$ is the union of $j_1(v, \pi, s)$, $j_1(v, \pi, s - 2\tilde{e}^n)$, $j_i(v + 2\tilde{e}^1, \pi, s - 2\tilde{e}^1)$ and $j_1(v + 2\tilde{e}^1, \pi, s - 2\tilde{e}^1 - 2\tilde{e}^n)$.

Our first task is to obtain a facetal description of J'. To this end, let $\sigma = j'(v, \pi, s) = [v^0, \cdots, v^n]$. Let V be the matrix with ith column $\binom{1}{v^i}$ for $0 \leqq i \leqq n$, and let V_0 be the matrix whose every column is $\binom{0}{v^0}$. Let S be the diagonal matrix with diagonal entries $1, s_1, \cdots, s_n$, P the permutation matrix $\begin{pmatrix} 1 & 0 \\ 0 & \tilde{P} \end{pmatrix}$, where the ith column of $\tilde{P}$ is $e^{\pi(i)}$, and

$$
(2.2)\qquad Y = \begin{bmatrix}
1 & 1 & 1 & \cdots & & 1 \\
 & 2 & 1 & \cdots & & 1 \\
 & & 1 & \cdots & & 1 \\
 & & & \ddots & & \vdots \\
 & & 0 & & 1 & 1 \\
 & & & & -1 & 1
\end{bmatrix}.
$$

It is easy to see that

$$
(2.3)\qquad V = V_0 + SPY.
$$

Since each of S, P and Y is clearly nonsingular, this implies that V is also, since $V - V_0$ results from subtracting multiples of its zeroth row from each other row. Thus σ is indeed a simplex, i.e., its vertices are affinely independent.

To express an arbitrary vector $x \in R^n$ as an affine combination of $v^0, \cdots, v^n$, we need to solve $V\lambda = \binom{1}{x}$; since $e^T\lambda = 1$ implies $V_0\lambda = \binom{0}{v^0}$, writing $z = \binom{1}{x-v^0}$, we find $(V - V_0)\lambda = z$ or, using (2.3)

$$\lambda = Y^{-1}P^TSz. \tag{2.4}$$

It is easy to check that

$$Y^{-1} = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & & & & & \\ & \frac{1}{2} & -\frac{1}{2} & & & & & \\ & & 1 & -1 & & & & \\ & & & \ddots & \ddots & & & \\ & & & & 1 & -1 & & \\ & & & & & \frac{1}{2} & -\frac{1}{2} \\ & & & & & \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{2.5}$$

Hence we obtain

PROPOSITION 2.2. *Given $\sigma = j'(v, \pi, s)$ and $x \in R^n$, let $w_i = s_{\pi(i)}(x_{\pi(i)} - v_{\pi(i)})$, $1 \leqq i \leqq n$. Then $x \in \sigma$ iff*

$$2 - w_2 \geqq w_1 \geqq w_2 \geqq \cdots \geqq w_{n-1} \geqq w_n \geqq - w_{n-1}. \tag{2.6}$$

Proof. Note that $P^TSz = \binom{1}{w}$ and that $x \in \sigma$ iff λ in (2.4) is nonnegative. □

We call (2.6) a facetal description of σ.

THEOREM 2.3. *J' is a triangulation of R^n. In fact, J' is also the subdivision of R^n generated by all hyperplanes of the form $x_i \pm x_j \in 2\mathbb{Z}$.*

Proof. It is clear that J' is locally finite, i.e., that each point of R^n has a neighborhood meeting only finitely many simplices of R^n. We complete the proof by showing that J' is indeed the simplicial subdivision claimed. First note that, by Proposition 2.2, each simplex of J' is bounded by hyperplanes of the form $x_i \pm x_j \in 2\mathbb{Z}$. Next we show that each simplex is a single piece of the subdivision. If not, there is some hyperplane that cuts a simplex of J', and hence, that has two vertices of the simplex strictly on opposite sides. Without loss of generality, we can assume that the hyperplane is $x_1 = x_2$ and the simplex $\sigma = j'(v, \pi, s)$, where $i = \pi^{-1}(1) < \pi^{-1}(2) = j$. By considering the cases $i = 1$, $j = 2$, $i = 1$, $2 < j < n$, $i = 1$, $j = n$, $2 < i < j < n$, $2 < i < n - 1$, $j = n$ and $i = n - 1$, $j = n$, we can easily show that all vertices of σ lie on the same side of the hyperplane.

Finally we must show that there are no other pieces of the subdivision. Thus we show that each $x \in R^n$ lies in some $\sigma \in J'$. For each i let v_i be a closest even integer to x_i and choose a permutation π and a sign vector s so that

$$1 \geqq w_1 \geqq \cdots \geqq w_n \geqq 0$$

where $w_i = s_{\pi(i)}(x_{\pi(i)} - v_{\pi(i)})$. Since $w_2 \leqq 1$, we have $2 - w_2 \geqq 1 \geqq w_1$ and since $w_{n-1} \geqq 0$, we have $w_n \geqq 0 \geqq - w_{n-1}$. Thus (2.6) holds. However we may have $s_{\pi(1)}$ or $s_{\pi(n)}$ equal to -1. In the latter case, we may simply reset $s_{\pi(n)}$ to $+1$ so that w_n switches sign but $w_{n-1} \geqq w_n \geqq -w_{n-1}$ still holds. If $s_{\pi(1)} = -1$, then reset $s_{\pi(1)}$ to $+1$ and decrease $v_{\pi(1)}$ by 2. Then w_1 becomes $2 - w_1$ so that the inequalities $2 - w_2 \geqq w_1 \geqq w_2$ are still satisfied. After these changes (2.6) holds with $s_{\pi(1)} = s_{\pi(n)} = 1$, so that x belongs to $j'(v, \pi, s)$ as desired. □

To conclude this section we give the pivot rules of J'. Suppose $\bar{\sigma} = j'(\bar{v}, \bar{\pi}, \bar{s})$ contains all vertices of $\sigma = j'(v, \pi, s)$ except v^i, with $\bar{\sigma} \neq \sigma$. Then we can obtain $\bar{v}$, $\bar{\pi}$, $\bar{s}$ and the index j of the new vertex of $\bar{\sigma}$ from the table below. As in Definition 2.1, $\tilde{e}^i$ denotes $s_{\pi(i)}e^{\pi(i)}$.

TABLE 2.4

		$\bar{v}$	$\bar{\pi}$	$\bar{s}$	j
$i=0$	$s_{\pi(2)}=+1$	$v+2\tilde{e}^1$	$(\pi(2), \pi(1), \cdots, \pi(n))$	$s-2\tilde{e}^1$	1
	$s_{\pi(2)}=-1$	$v+2\tilde{e}^1+2\tilde{e}^2$	$(\pi(2), \pi(1), \cdots, \pi(n))$	$s-2\tilde{e}^1-2\tilde{e}^2$	0
$i=1$	$s_{\pi(2)}=+1$	v	$(\pi(2), \pi(1), \cdots, \pi(n))$	s	1
	$s_{\pi(2)}=-1$	$v+2\tilde{e}^2$	$(\pi(2), \pi(1), \cdots, \pi(n))$	$s-2\tilde{e}^2$	0
$1<i<n-1$		v	$(\pi(1), \cdots, \pi(i+1), \pi(i), \cdots, \pi(n))$	s	i
$i=n-1$	$s_{\pi(n-1)}=+1$	v	$(\pi(1), \cdots, \pi(n-2), \pi(n), \pi(n-1))$	s	$n-1$
	$s_{\pi(n-1)}=-1$	v	$(\pi(1), \cdots, \pi(n-2), \pi(n), \pi(n-1))$	$s-2\tilde{e}^{n-1}$	n
$i=n$	$s_{\pi(n-1)}=+1$	v	$(\pi(1), \cdots, \pi(n-2), \pi(n), \pi(n-1))$	$s-2\tilde{e}^n$	$n-1$
	$s_{\pi(n-1)}=-1$	v	$(\pi(1), \cdots, \pi(n-2), \pi(n), \pi(n-1))$	$s-2\tilde{e}^{n-1}-2\tilde{e}^n$	n

3. The 3-dimensional case. When $n=3$, each simplex of J' is of the form $\sigma = j'(v, \pi, s) = [v^0, v^1, v^2, v^3]$ where

$$v^0 = v, \qquad v^2 = v^1 - e^{\pi(1)} + s_{\pi(2)}\, e^{\pi(2)} - e^{\pi(3)},$$
$$v^1 = v^0 + 2\, e^{\pi(1)}, \qquad v^3 = v^2 + 2\, e^{\pi(3)}.$$

In this section, we show that J' is similar to $A^*_\pm K_1$, where $A^*_\pm$ is the matrix $(n+1 \pm \sqrt{n+1})I - ee^T$, the "optimal" linear transformation of the Freudenthal triangulation K_1; see van der Laan and Talman [8], Eaves [6], and [20]. Here similar means "obtainable by an orthogonal matrix and a scaling."

First note that, for any n, $A^*_\pm K_1$ are similar. Indeed, $A^*_- = ((n+1-\sqrt{n+1})/(n+1+\sqrt{n+1}))\ (I-2ee^T/n)A^*_+$, and $I-2ee^T/n$ is easily seen to be orthogonal.

Thus it suffices to show that $A^*_- K_1$ and J' are similar when $n=3$. In fact, we will show them to be identical. For $n=3$,

$$A^*_- = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}.$$

We will denote the columns of A^*_- a^1, a^2 and a^3. Note first that the set of vertices of $A^*_- K_1$ is $\{v \in \mathbb{Z}^3$: all components of v are even or all are odd$\} = \{v \in \mathbb{Z}^3$: v does not have exactly 1 nor exactly $n-1=2$ odd components$\}$. Thus the two triangulations have the same vertices.

Let $\sigma = [v^0, v^1, v^2, v^3] \in A^*_- K_1$, so that v^0 has all components odd or all even,

$$v^1 = v^0 + a^i, \quad v^2 = v^1 + a^j, \quad v^3 = v^2 + a^k$$

for some permutation (i, j, k) of $(1, 2, 3)$. Then either v^3 or v^2 has all components even. In the first case,

$$v^1 = v^3 + 2e^i, \quad v^2 = v^1 - e^i + e^j - e^k, \quad v^0 = v^2 + 2e^k,$$

so $\sigma = [v^3, v^1, v^2, v^0] \in J'$. In the second case,

$$v^0 = v^2 + 2e^k, \quad v^3 = v^0 - e^k - e^j - e^i, \quad v^1 = v^3 + 2e^i,$$

so $\sigma=[v^2, v^0, v^3, v^1]\in J'$.

Conversely, let $\sigma=[v^0, v^1, v^2, v^3]\in J'$, so that v^0 has all components even,

$$v^1=v^0+2e^i,\quad v^2=v^1-e^i+s_je^j-e^k,\quad v^3=v^2+2e^k$$

for some permutation (i, j, k) of $(1, 2, 3)$ and some $s_j\in\{+1, -1\}$. Then if $s_j=+1$,

$$v^1=v^3+a^i,\quad v^2=v^1+a^j,\quad v^0=v^2+a^k,$$

so $\sigma=[v^3, v^1, v^2, v^0]\in A^*_-K_1$. On the other hand, if $\sigma_j=-1$, we have

$$v^3=v^1+a^k,\quad v^0=v^3+a^j,\quad v^2=v^0+a^i,$$

so $\sigma=[v^1, v^3, v^0, v^2]\in A^*_-K_1$. This completes the proof that J' is identical to $A^*_-K_1$, and thus is similar to $A^*_+K_1$.

4. Measures for *J'*. In this section we compute several measures to evaluate the new triangulation J'.

The first crude measure is the number of simplices used to triangulate the unit cube $[0, 1]^n$. This is $n!$ for J_1 and K_1, but there are triangulations with far fewer simplices in the unit cube—see Lee [10] and Sallee [13]. Unfortunately, the unit cube is not triangulated by J', whose generating hyperplanes are of the form $x_i\pm x_j\in 2\mathbb{Z}$; nor is any cube for odd n. However, we can find parallelopipeds that are triangulated by J' for each n. One such is

$$\begin{aligned}\tilde{C}^n=\{x\in R^n\colon\ & 0\leqq x_{2i-1}-x_{2i}\leqq 2,\ 0\leqq x_{2i-1}+x_{2i}\leqq 2,\\ & i=1, 2,\cdots, \lfloor n/2\rfloor, \text{and } 0\leqq x_{n-1}+x_n\leqq 2 \text{ if } n \text{ is odd}\}.\end{aligned}$$

Let A_n be the $n\times n$ matrix given by

$$A_{2k}=\begin{bmatrix} 1 & 1 & & & & & \\ 1 & -1 & & & & 0 & \\ & & 1 & 1 & & & \\ & & 1 & -1 & & & \\ & & & & \ddots & & \\ & 0 & & & & 1 & 1 \\ & & & & & 1 & -1 \end{bmatrix},$$

$$A_{2k+1}=\begin{bmatrix} 1 & 1 & & & & & & \\ 1 & -1 & & & & 0 & & \\ & & 1 & 1 & & & & \\ & & 1 & -1 & & & & \\ & & & & \ddots & & & \\ & 0 & & & & 1 & 1 & \\ & & & & & 1 & -1 & \\ & & & & & & 1 & 1 \end{bmatrix};$$

then $\tilde{C}^n=\{x\in R^n\colon 0\leqq y=A_nx\leqq 2e\}$. Since A_n has determinant $2^{\lfloor n/2\rfloor}$ and $\{y\in R^n\colon 0\leqq y\leqq 2e\}$ has volume 2^n, $\tilde{C}^n$ has volume $2^{\lceil n/2\rceil}$. Since each simplex of J', as the union of four simplices of J_1, has volume $4/n!$ (this can easily be seen directly from (2.2), (2.3)), we obtain

THEOREM 4.1. *There are parallelopipeds in R^n that are triangulated by J' into $n!\,2^{\lceil n/2\rceil}/4$ simplices.*

For $n \geqq 4$, this measure is worse than J_1 and K_1. To me, this indicates the inadequacies of the measure: the problem is not the simplices in J' but the lack of small parallelopipeds. Notice also that, for $n \geqq 4$, the triangulation of $\tilde{C}^n$ has vertices that are not vertices of $\tilde{C}^n$.

While the simplices of J' have volume four times those of J_1 or K_1, they share the small mesh size of these triangulations of the unit cube. The mesh size of a triangulation is the supremum of the diameters of its simplices, or the supremum of the lengths of its 1-simplices. In fact, we have

PROPOSITION 4.2. *The mesh size of J' is* $\max\{2, \sqrt{n}\}$.

Next we compute the average directional density of J' [16]. This is, roughly, the rate at which a random straight line meets facets of J' per unit length. Alternatively, Eaves and Yorke [7] have shown that it is the surface density, i.e. the surface area of simplices per unit volume, up to a scale factor. Since x lies in a facet of a simplex of J' iff $x_i \pm x_j$ is an even integer for some i, j, we obtain by the arguments of [16]:

THEOREM 4.3. *The directional density of J' in direction $\underline{d}$ is $N(J', d) = \sum_{i<j} \frac{1}{2}\{|d_i + d_j| + |d_i - d_j|\}$ and its average directional density is $\binom{n}{2}\sqrt{2}g_n$, where $g_n = 2\Gamma(n/2)/(n-1)\sqrt{\pi}\Gamma((n-1)/2)$.*

In a companion paper [20], we show that, of all triangulations AJ', with A a nonsingular linear transformation such that AJ' has the same mesh size as J', J' itself has the smallest average directional density.

5. Triangulations induced by J'. Note that J' refines the cubical subdivision of R^n, that is, the polyhedral subdivision generated by all hyperplanes of the form $x_i \pm x_j = 0$ (whose pieces are cones with faces of a cube centered at the origin as cross sections). This follows directly from Theorem 2.3. An immediate implication is that J' can be used in the octahedral piecewise-linear homotopy algorithm of Wright [21].

However, J' does not refine the octahedral (or orthant) subdivision of R^n, since $x_i = 0$ is not among its generating hyperplanes. Thus it cannot be used in the cubical (or $2n$-) algorithm of van der Laan and Talman [9] and Reiser [12]. Even more apparently limiting is the fact that its $(n+1)$-dimensional version does not also triangulate $R^n \times [0, 1]$, and thus it cannot directly be used in a restart method such as Merrill's [11].

In this section we show that a wealth of other triangulations are induced by J', so that these objections lose their force. Indeed, it follows from, e.g., [19, Thms. 3.1, 7.1] that, if S is a triangulation of R^n generated by a family of hyperplanes, and if H is one of these hyperplanes, then $T = \{\tau\colon \tau \text{ an } (n-1)\text{-face of some } \sigma \in S, \tau \subseteq H\}$ triangulates H. We say T is induced by S. If α is an affine isomorphism between H and R^{n-1}, then αT is a triangulation of R^{n-1}, and we also say αT is induced by S. We use these ideas to construct from J' two families of triangulations. In order to specify the dimension, we write $J'(n)$ for the triangulation J' of R^n.

DEFINITION 5.1. For $1 \leqq k \leqq n+1$, let $J'_k(n)$ denote the polyhedral subdivision of R^n generated by the hyperplanes $x_i \pm x_j \in 2\mathbb{Z}$, $1 \leqq i < j \leqq n$ and $x_i \in \mathbb{Z}$, $k \leqq i \leqq n$, and let $J''_k(n)$ denote that generated by the hyperplanes $x_i \pm x_j \in 2\mathbb{Z}$, $1 \leqq i < j \leqq n$, $x_i \in 2\mathbb{Z}$, $1 \leqq i < k$, and $x_i \in \mathbb{Z}$, $k \leqq i \leqq n$. We write $J''(n)$ for $J''_{n+1}(n)$ (note that $J'(n) = J'_{n+1}(n)$). Also note that $J'_1(n) = J''_1(n) = J_1(n)$.

In this section, we shall prove

THEOREM 5.2. *$J'_k(n)$ is a triangulation of R^n for $n = 1, 2$ and $1 \leqq k \leqq n$ and for $n \geqq 3$ and $1 \leqq k \leqq n+1$. $J''_k(n)$ is a triangulation of R^n for $n \geqq 1$ and $1 \leqq k \leqq n+1$.*

Let B_{kn}, for $n \geqq 2$ and $2 \leqq k \leqq n$, denote the $n \times n$ matrix

$$\left[\begin{array}{cccccccc} 1 & & & & & & & \\ & \ddots & & & & 0 & & \\ & & 1 & & & & & \\ & & & \frac{1}{2} & -\frac{1}{2} & & & \\ & & & \frac{1}{2} & \frac{1}{2} & & & \\ & & 0 & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{array}\right] \begin{array}{l} \\ \\ \\ \leftarrow k-1 \\ \leftarrow k \\ \\ \\ \\ \end{array}$$

and note that

$$B_{kn}^{-1} = \left[\begin{array}{cccccccc} 1 & & & & & & & \\ & \ddots & & & & 0 & & \\ & & 1 & & & & & \\ & & & 1 & 1 & & & \\ & & & -1 & 1 & & & \\ & & & & & 1 & & \\ & & 0 & & & & \ddots & \\ & & & & & & & 1 \end{array}\right] \begin{array}{l} \\ \\ \\ \leftarrow k-1 \\ . \leftarrow k \\ \\ \\ \\ \end{array}$$

Thus if $T(n)$ is a subdivision of R^n generated by hyperplanes containing $x_{k-1} \pm x_k \in 2\mathbb{Z}$, then the nonsingular transformation $x \to y = B_{kn}x$ takes $T(n)$ into a subdivision $B_{kn}T(n)$, also generated by hyperplanes, and $y_{k-1} \in \mathbb{Z}$, $y_k \in \mathbb{Z}$ are among these.

Next let $S(n)$ be a hyperplane-generated subdivision of R^n with $x_k = 0$ one of its generating hyperplanes, $1 \leqq k \leqq n$. Then $S(n)$ induces a subdivision S' of $H = \{x \in R^n: x_k = 0\}$. Let α be the affine isomorphism of H and R^{n-1} defined by $\alpha(x_1, \cdots, x_{k-1}, 0, x_{k+1}, \cdots, x_n) = (x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_n)$. Then we denote the subdivision $\alpha S'$ of R^{n-1} by $P_kS(n)$. Finally, let $Q_kT(n) = P_kB_{kn}T(n)$.

Note that, from the remarks above Definition 5.1, if $T(n)$, $S(n)$ are triangulations, then so are $Q_kT(n)$ and $P_kS(n)$.

LEMMA 5.3.

(a) $Q_{k-1}J'_k(n) = J'_{k-2}(n-1)$ $(3 \leqq k \leqq n+1)$.

(b) $P_nJ'_k(n) = J''_k(n-1)$ $(1 \leqq k \leqq n)$.

(c) $Q_nJ'_k(n) = J'_k(n-1)$ $(1 \leqq k < n)$.

(d) $Q_kJ'_k(n) = J'_{k-1}(n-1)$ $(2 \leqq k \leqq n)$.

(e) $Q_{k-1}J''_k(n) = J''_{k-2}(n-1)$ $(3 \leqq k \leqq n+1)$.

(f) $P_nJ''_k(n) = J''_k(n-1)$ $(1 \leqq k \leqq n)$.

(g) $Q_nJ''_k(n) = J''_k(n-1)$ $(1 \leqq k < n)$.

(h) $Q_kJ''_k(n) = J''_{k-1}(n-1)$ $(2 \leqq k \leqq n)$.

(i) $P_{k-1}J''_k(n) = J''_{\min\{k,n\}}(n-1)$ $(2 \leqq k \leqq n+1)$.

Proof. In each case we merely need to check the generating hyperplanes. We will show (a) and (b); the reader will have no difficulty in verifying (c)–(i).

(a) Consider $y = B_{kn}x$, $x = B_{kn}^{-1}y$. Then the hyperplanes $x_i \pm x_j \in 2\mathbb{Z}$, for $\{i, j\} \cap \{k-2, k-1\} = \varnothing$ become $y_i \pm y_j \in 2\mathbb{Z}$. The hyperplanes $x_{k-2} \pm x_{k-1} \in 2\mathbb{Z}$ become $y_{k-2} \in \mathbb{Z}$, $y_{k-1} \in \mathbb{Z}$. The hyperplanes $x_i \pm x_{k-2} \in 2\mathbb{Z}$ and $x_i \pm x_{k-1} \in 2\mathbb{Z}$, $i \notin \{k-2, k-1\}$, become $y_i \pm y_{k-2} \pm y_{k-1} \in 2\mathbb{Z}$. Finally, the hyperplanes $x_i \in \mathbb{Z}$, $i \geqq k$, become $y_i \in \mathbb{Z}$, $i \geqq k$. Now intersecting these hyperplanes with $y_{k-1} = 0$ and projecting down to R^{n-1} gives the hyperplanes $x_i \pm x_j \in 2\mathbb{Z}$, all i, j and $x_i \in \mathbb{Z}$, $i \geqq k-2$, as desired.

(b) Consider the effect of intersecting hyperplanes with $x_n = 0$ and then projecting down to R^{n-1}. The hyperplanes $x_i \pm x_j \in 2\mathbb{Z}$, $i<j<n$, remain the same, as do $x_i \in \mathbb{Z}$, $k \leqq i < n$. The hyperplanes $x_i \pm x_n \in 2\mathbb{Z}$, $i<n$, become $x_i \in 2\mathbb{Z}$, $i<n$. Thus we have the generating family of $J''_k(n-1)$. □

Proof of Theorem 5.2. Applying (a) inductively implies that each $J'_k(n)$ is a triangulation. Indeed, if we have established that all $J'_k(n)$ where $n \geqq m$ and $n-k \leqq l$ are triangulations, then (a) shows that all $J'_k(n)$ for $n \geqq m-1$ and $n-k \leqq l+1$ are triangulations. The case $n=k=1$ is trivial. Similarly, (b) now shows that all $J''_k(n)$ are triangulations. □

Note that parts (f) and (g) of the lemma show that J_1 induces only copies of itself. A similar analysis demonstrates that K_1 also induces only copies of itself.

Let us tabulate the symmetry properties of these triangulations. By invariance under permutations, we mean under permutations of $\{1, \cdots, n\}$ that leave $\{1, \cdots, k-1\}$ and $\{k, \cdots, n\}$ invariant, and by invariance under reflection, we mean under transformations $x_i \to -x_i$ for any i. Even translations are translations by vectors with all components even integers. Table 5.4 also notes which triangulations refine the cubical and octahedral (orthant) subdivisions of R^n and which also triangulate $R^{n-1} \times [0,1]$.

TABLE 5.4

	Invariant under						
			Even	Translations	Refines		Triangulates
	Permutations	Reflections	translations	by e	Cubical	Octahedral	$R^{n-1}x\{0,1\}$
$J'(n)$	√	√	√	√	√	×	×
$J'_k(n)$ $1<k\leqq n$	√	√	√	√	√	×	√
$J''(n)$	√	√	√	×	√	√	×
$J''_k(n)$ $1<k\leqq n$	√	√	√	×	√	√	√
$J_1(n)$	√	√	√	√	√	√	√
$K_1(n)$	√	×	√	√	×	√	√

Among the new triangulations we have introduced, $J'(n)$ seems most suitable for the octahedral algorithm and $J''(n)$ for the cubical algorithm. We also recommend the restriction of $J'_{n+1}(n+1)$ to $R^n \times [0,1]$ for use in a restart algorithm. There is, however, another candidate; we may use $Q_{n+1,n+1}J'(n+1)$. This triangulation is generated by hyperplanes $x_i \pm x_j \in 2\mathbb{Z}$, $i<j<n$; $x_i \pm x_n \pm x_{n+1} \in 2\mathbb{Z}$, $i<n$; and $x_n \in \mathbb{Z}$, $x_{n+1} \in \mathbb{Z}$. It appears at first sight (and based on its directional density) to be inferior to $J'_{n+1}(n+1)$. However, $Q_{n+1,n+1}J'(n+1)$ has vertices with $x_{n+1} = \frac{1}{2}$. For example, consider its simplex $[0, 2e^1, e^1+\frac{1}{2}e^4+\frac{1}{2}e^5, e^1+e^5, e^1+e^5-e^2-e^3, e^1+e^5-e^2+e^3]$ when $n=4$. In a restart algorithm we may choose the image of such a vertex arbitrarily. By making this choice appropriately, we may identify two vertices of several simplices, thus obtaining a new and simplified triangulation. For example, if we round x_{n+1} to the nearest odd integer and x_n to the nearest even integer when either (and hence both) is an odd multiple of $\frac{1}{2}$, then the sample simplex above disappears. The details of this approach have not

yet been worked out, but there remains the possibility that some such modification of $Q_{n+1,n+1}J'(n+1)$ will be a reasonable choice for a restart method.

So far the various induced triangulations have been defined only by their generating hyperplanes. We now describe them by their individual simplices. We consider $J'_k(n)$ or $J''_k(n)$. Each simplex takes the following form. As in § 2, we let v be a vector of R^n with each component an even integer, π be a permutation of $\{1, \cdots, n\}$ and s a sign vector. We require $s_{\pi(1)}$ to be 1 if $\pi(1) < k$, and $s_{\pi(n)}$ to be 1 if $\pi(n) < k$ and we are considering $J'_k(n)$. We let $\tilde{e}^i = s_{\pi(i)}e^{\pi(i)}$. Also, set (α, β) to $(2, -1)$ if $\pi(1) < k$ and to $(1, 0)$ otherwise, and set (γ, δ) to $(-1, 2)$ if $\pi(n) < k$ and we are considering $J'_k(n)$ and to $(0, 1)$ otherwise. Then the vertices of the corresponding simplex are

$$\begin{aligned} v^0 &= v, & v^j &= v^{j-1} + \tilde{e}^j, \quad 3 \leqq j \leqq n-2, \\ v^1 &= v^0 + \alpha\tilde{e}^1, & v^{n-1} &= v^{n-2} + \tilde{e}^{n-1} + \gamma\tilde{e}^n, \\ v^2 &= v^1 + \beta\tilde{e}^1 + \tilde{e}^2, & v^n &= v^{n-1} + \delta\tilde{e}^n. \end{aligned}$$

We leave the reader the verification of this description and the derivation of appropriate pivot rules for a particular $J'_k(n)$ or $J''_k(n)$.

The following results are easy to derive.

PROPOSITION 5.5. *For $k \geq 1$, the mesh sizes of $J'_k(n)$ and $J''_k(n)$ are* max $\{2, \sqrt{n}\}$, *while for $k = 1$ the figure is $\sqrt{n}$.*

THEOREM 5.6. *The directional density of $J'_k(n)$ in direction d is*

$$N(J'_k(n), d) = \sum_{i \geqq k} |d_i| + \sum_{i<j} \tfrac{1}{2}\{|d_i + d_j| + |d_i - d_j|\}$$

while for $J''_k(n)$ it is

$$N(J''_k(n), d) = \sum_{i<k} \tfrac{1}{2}|d_i| + \sum_{i \geqq k} |d_i| + \sum_{i<j} \tfrac{1}{2}\{|d_i + d_j| + |d_i - d_j|\}.$$

The average directional densities are

$$N(J'_k(n)) = ((n+1-k) + \tbinom{n}{2}\sqrt{2})g_n \quad \textit{and}$$

$$N(J''_k(n)) = ((n+\tfrac{1}{2} - k/2) + \tbinom{n}{2}\sqrt{2})g_n$$

where g_n is as in Theorem 4.3.

6. Computational experience. In this section we give the results of some numerical experimentation with the use of $J'_{n+1}(n+1)$ in a restart algorithm. Note that Broadie [3] has compared J' to J_1 and K'_1 in the octahedral algorithm; his tests indicate that J' almost always requires fewer function evaluations.

In all our runs we use the homotopy $h(x, t) = tf(x) + (1-t)r(x)$ where r is the identity. Thus we employed the large pieces induced by the linearity of r as described in [17]. The code PLALGO [18] was used and required very little change from its program for the large pieces for J_1. We compared Merrill's restart algorithm with large pieces based on $J'_{n+1}(n+1)$ and $J_1(n+1)$ and van der Laan and Talman's cubical (or $2n$-) algorithm with the triangulation $K'(n)$. We distinguish these cases below by writing J', J_1 or K' respectively.

For the economic equilibrium problems we chose an initial grid size of $1/(n+1)$ for problems in dimension n. The equilibrium problems were converted to zero-finding problems as in [2]. We solved the pure trade examples of dimension 4, 7 and 9 (E1–E3) and the examples with production of dimension 5 and 13 (EP1 and EP2) in Scarf with Hansen [14]; note that in all cases the number of prices (commodities) is one larger than the dimension. For these problems, we report the results of a run as $p/q/r$, where p linear programming pivots, q function evaluations, and r demand evaluations were required.

For E1–E3, a refinement factor between restarts of .37 was used while the convergence test was $\|f(x)\|_\infty \leqq 10^{-12}$. For EP1 and EP2 and the other runs below, the default refinement factor of .5 was used. In EP1 and EP2, the convergence tolerance was relaxed to 10^{-10}. The remaining parameters in PLALGO had their default values; thus quasi-Newton acceleration was employed.

Our next test problem is Brown's almost linear function, defined by

$$f_1(x) = \prod_{j=1}^{n} x_j - 1,$$

$$f_j(x) = \sum_{i=1}^{n} x_i + x_j - n - 1, \qquad j > 2.$$

We solved this for $n = 10$, 15 and 20, with starting point the origin and initial grid size $\delta = .5$. The convergence tolerance was 10^{-10} for $n = 10$ and $n = 15$, and 10^{-8} for $n = 20$. We report the results as p/q, with p and q as above.

Finally we considered Watson's test function, defined by

$$f_j(x) = x_j - \exp\left(\cos\left(j \sum_{i=1}^{n} x_i\right)\right).$$

We solved this for $n = 1, 2, \cdots, 10$, with starting point the origin, initial grid size $\delta = .5$ and convergence tolerance 10^{-8}. The results are reported similarly.

The results demonstrate a consistent advantage of $J'_{n+1}(n+1)$ over $J_1(n+1)$ in Merrill's algorithm, and (usually) a considerable advantage over the cubical algorithm with $K'(n)$. Other experimentation has shown that the first statement holds true over a variety of test problems, while the comparative advantages of Merrill's algorithm and the cubical algorithm can depend considerably on the problem type.

TABLE 6.1

	Economic equilibrium problems				
	E1	E2	E3	EP1	EP2
J'	56/65/65	82/91/91	60/73/73	119/126/56	726/689/70
J_1	60/68/68	104/112/112	70/84/84	124/132/53	842/802/96
K'	53/67/67	88/101/101	62/77/77	124/133/57	839/817/91

	Brown's almost-linear function		
n	10	15	20
J'	93/92	290/278	933/893
J_1	117/116	335/323	1032/992
K'	146/153	505/511	784/786

	Watson's test function									
n	1	2	3	4	5	6	7	8	9	10
J'	6/9	5/10	48/48	160/147	239/217	352/322	958/844	1775/1588	3043/2701	6759/6060
J_1	6/9	7/11	55/57	184/167	451/370	575/513	1582/1405	3732/3402	7080/5945	19520/17020
K'	5/9	11/18	46/56	170/182	476/482	806/813	2625/2556	4916/4823	12505/12188	21108/20732

REFERENCES

[1] E. Allgower and K. Georg, *Simplicial and continuation methods for approximating fixed points*, SIAM Rev., 22 (1980), pp. 28–85.

[2] S. A. Awoniyi and M. J. Todd, *An efficient simplicial algorithm for computing a zero of a convex union of smooth functions*, Math. Programming, 25 (1983), pp. 83–108.

[3] M. N. Broadie, *Subdivisions and antiprisms for P2 homotopy algorithms*, Ph.D. thesis, Dept. Operations Research, Stanford Univ., Stanford, CA, 1983.

[4] O. Buneman, *Tetrahedral finite elements for interpolation*, SIAM J. Sci. Stat. Comput., 1 (1980), pp. 223–248.

[5] B. C. Eaves, *A short course in solving equations with PL homotopies*, in Nonlinear Programming, Proc. Ninth SIAM-AMS Symposium in Applied Mathematics, R. W. Cottle and C. E. Lemke, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1976, pp. 73–143.

[6] ———, *Permutation congruent transformations of the Freudenthal triangulation with minimal surface density*, Tech. Rep., Dept. Operations Research, Stanford Univ., Stanford, CA, March 1982.

[7] B. C. Eaves and J. Yorke, *Equivalence of surface density and average directional density*, Math. Oper. Res., to appear.

[8] G. van der Laan and A. J. J. Talman, *An improvement of fixed point algorithms by using a good triangulation*, Math. Programming, 18 (1980), pp. 274–285.

[9] ———, *A class of simplicial restart fixed point algorithms without an extra dimension*, Math. Programming, 20 (1981), pp. 33–48.

[10] C. W. Lee, *Triangulating the d-cube*, manuscript, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1981.

[11] O. H. Merrill, *Applications and extensions of an algorithm that computes fixed points of certain upper semi-continuous point to set mappings*, Ph.D. thesis, Dept. Industrial Engineering, Univ. Michigan, Ann Arbor, 1972.

[12] P. M. Reiser, *A modified integer labeling for complementarity algorithms*, Math. Oper. Res., 6 (1981), pp. 129–139.

[13] J. F. Sallee, *A triangulation of the n-cube*, Discrete Math., 40 (1982), pp. 81–86.

[14] H. E. Scarf with T. Hansen, *Computation of Economic Equilibria*, Yale Univ. Press, New Haven, CT, 1973.

[15] M. J. Todd, *The computation of fixed points and applications*, Lecture Notes in Economic and Mathematical Systems 124, Springer-Verlag, Berlin, 1976.

[16] ———, *On triangulations for computing fixed points*, Math. Programming, 10 (1976), pp. 322–346.

[17] ———, *Traversing large pieces of linearity in algorithms that solve equations by following piecewise-linear paths*, Math. Oper. Res., 5 (1980), pp. 242–257.

[18] ———, PLALGO: *a* FORTRAN *implementation of a piecewise-linear homotopy algorithm for solving systems of nonlinear equations*, Tech. Rep. 452, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY, 1981.

[19] ———, *An introduction to piecewise-linear homotopy algorithms for solving systems of equations*, in Topics in Numerical Analysis, P. R. Turner, ed., Lecture Notes in Mathematics 965, Springer-Verlag, Berlin–Heidelberg–New York, 1982, pp. 149–202.

[20] ———, *Solutions to certain matrix optimization problems*, Tech. Rep. 584, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY, 1983.

[21] A. H. Wright, *The octahedral algorithm, a new simplicial fixed point algorithm*, Math. Programming, 21 (1981), pp. 47–69.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

012

BOOLEAN METHODS OF OPTIMIZATION OVER INDEPENDENCE SYSTEMS*

BERNIE L. HULME†

Abstract. This paper presents both a direct and an iterative method of solving the combinatorial optimization problem associated with any independence system. The methods use Boolean algebraic computations to produce solutions. In addition, the iterative method employs a version of the greedy algorithm both to compute upper bounds on the optimum value and to produce the additional circuits needed at every stage. The methods are extensions of those used to solve a problem of fire protection at nuclear reactor power plants.

AMS subject classifications. 68C20, 90C09

1. Independence systems.

An *independence system* $S=(E,\mathcal{I})$ is a finite set E and a nonempty collection $\mathcal{I}$ of subsets of E such that (I1) and (I2) are satisfied.

(I1) $\varnothing\in\mathcal{I}$.

(I2) If $X\in\mathcal{I}$ and $Y\subseteq X$, then $Y\in\mathcal{I}$.

The subsets of E in $\mathcal{I}$ are called *independent* sets, and those not in $\mathcal{I}$ are called *dependent* sets. A maximal independent set is a *base* of S. The collection of bases is $\mathcal{B}$. A minimal dependent set is a *circuit* of S, and the collection of circuits is $\mathcal{C}$. Several properties follow immediately from these definitions.

THEOREM 1. *A subset X of E is independent if and only if X is a subset of a base.*

THEOREM 2. *A subset X of E is dependent if and only if X is a superset of a circuit.*

Restating Theorem 2 in a negative way yields

COROLLARY 1. *A subset X of E is independent if and only if X does not contain a circuit.*

From Theorems 1 and 2 it is clear that either the bases $\mathcal{B}$ or the circuits $\mathcal{C}$ uniquely determine an independence system on E. Given E and $\mathcal{B}$, $\mathcal{I}$ consists of all the subsets of the members of $\mathcal{B}$, and the minimal subsets of E not in $\mathcal{I}$ form $\mathcal{C}$. Given E and $\mathcal{C}$, the dependent sets are the supersets of the members of $\mathcal{C}$, and the maximal subsets of E not among the dependent sets form $\mathcal{B}$.

Notice that (I1) prevents $\mathcal{I}$ from being empty, and hence $\mathcal{B}$ cannot be empty. However, $\mathcal{C}$ is empty whenever $E\in\mathcal{I}$ and E is the only base. The following theorems characterize $\mathcal{B}$ and $\mathcal{C}$.

THEOREM 3. *A nonempty collection $\mathcal{B}$ of subsets of E is the set of bases of an independence system on E if and only if* (B1) *is satisfied.*

(B1) *If B_1 and B_2 are distinct members of $\mathcal{B}$, then $B_1\not\subseteq B_2$.*

Proof. If (B1) holds, then no member of $\mathcal{B}$ contains another member. In $\mathcal{I}$, the collection of all subsets of the members of $\mathcal{B}$, the maximal elements are precisely the members of $\mathcal{B}$. Conversely, if $\mathcal{B}$ is a set of bases, then no member of $\mathcal{B}$ can contain another, for otherwise some base would not be maximal. □

THEOREM 4. *A collection $\mathcal{C}$ of subsets of E is the set of circuits of an independence system on E if and only if* (C1) *is satisfied.*

(C1) If C_1 and C_2 are distinct members of $\mathcal{C}$, then $C_1\not\subseteq C_2$.

* Received by the editors July 6, 1983. This work was performed at Sandia National Laboratories supported by the U.S. Department of Energy under contract DE-AC04-76DP00789. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Sandia National Laboratories, Albuquerque, New Mexico 87185.

Proof. If (C1) holds, then no member of $\mathscr{C}$ contains another member. Among the supersets of all the members of $\mathscr{C}$, the minimal ones are precisely the members of $\mathscr{C}$. Thus, $\mathscr{C}$ is the set of circuits of an independence system whose dependent sets are the supersets of the members of $\mathscr{C}$. Conversely, if $\mathscr{C}$ is a set of circuits, then no member of $\mathscr{C}$ can contain another member, because otherwise some circuit would not be minimal. □

2. Duality. For any independence system $S=(E, \mathscr{I})$, there is a *dual* independence system $S^*=(E, \mathscr{I}^*)$ whose bases are the complements (relative to E) of the bases of S. A base of S^* is called a *cobase* of S. A circuit of S^* is called a *cocircuit* of S, but a cocircuit is not necessarily the complement of a circuit. An independence system S and its dual S^* are linked by the following theorems.

THEOREM 5. *A subset X of E is a cobase of an independence system $S=(E, \mathscr{I})$ if and only if X has a nonnull intersection with every circuit of S and is minimal with respect to this property.*

Proof. Let B^* be a cobase of S. Suppose there exists a circuit C of S such that $B^* \cap C=\varnothing$. Then the base $B=E \backslash B^*$ contains C, and therefore B is dependent in S, a contradiction. Hence, B^* has nonnull intersection with every circuit of S. Now let $X \subset B^*$ have nonnull intersection with every circuit of S. Then $E \backslash X$ contains no circuit of S and, by Corollary 1, is independent in S. But $E \backslash X \supset E \backslash B^*$, a base of S, which is a contradiction. Thus, a cobase B^* is a minimal subset having nonnull intersection with every circuit of S.

Conversely, if X has a nonnull intersection with every circuit of S, then $E \backslash X$ contains no circuit of S and is independent in S. If X is also minimal, then $E \backslash X$ is a maximal independent set in S. Thus, X is a cobase of S. □

In a similar way one can prove

THEOREM 6. *A subset Y of E is a circuit of an independence system S if and only if Y has nonnull intersection with every cobase of S and is minimal with respect to this property.*

Also, by duality one can prove two other theorems which relate the bases and cocircuits of S.

3. Combinatorial optimization. The *combinatorial optimization problem* for an independence system $S=(E, \mathscr{I})$ is as follows. Given a *weight function* $w(e) \geqq 0$, for all $e \in E$, find an independent set having the maximum total weight.

One might as well restrict attention to the bases of S. This is because a maximum weight independent set is either maximal or else differs from a maximal independent set only by elements of zero weight. Also, a base is a maximum weight base if and only if its complement is a minimum weight cobase. Therefore, we may think of optimization over independence systems as finding either a maximum weight base or a minimum weight cobase.

The *greedy algorithm* is a heuristic solution process which considers the elements e of E in order of decreasing weight. As long as $I+e$ is independent, e is added to I; but if $I+e$ is dependent, e is added to Q. When all of the elements of E have been considered, I is a base and $Q=E \backslash I$ is a cobase.

THE GREEDY ALGORITHM.

```
1. Set $I=\varnothing$, $Q=\varnothing$.
2. While $E \neq \varnothing$ do
   a. let $e$ be an element of $E$ having largest weight;
   b. set $E=E-e$;
   c. if $I+e \in \mathscr{I}$, then set $I=I+e$
          else set $Q=Q+e$.
```

This algorithm does not always solve the combinatorial optimization problem for an independence system S. However, the greedy algorithm does produce correct solutions for any nonnegative weight function if and only if S is a special case called a matroid (Theorem 9).

4. Boolean methods.

4.1. A direct method. The following Boolean procedure will solve the combinatorial optimization problem for any independence system $S=(E,\mathcal{I})$. It constructs a minimum weight cobase of S from the circuits $\mathcal{C}$ using Theorem 5.

```
THE DIRECT BOOLEAN ALGORITHM (given the circuits $\mathcal{C}$).
1. Create the Boolean conjunctive normal form
       $$F=\bigwedge_{C\in\mathcal{C}}\bigvee_{e\in C} e.$$
2. Expand $F$ by the distributive laws, $(a\vee b)(a\vee c)=a\vee bc$ and $a(b\vee c)=ab\vee ac$,
   and simplify the result by idempotence, $aa=a$ and $a\vee a=a$, and absorption,
   $a\vee ab=a$.
3. For each term in the resulting disjunctive normal form of $F$, compute the weight
   sum, and save one minimum weight term, $R$.
```

THEOREM 7. *The direct Boolean algorithm produces a minimum weight cobase R of the independence system S with circuits $\mathcal{C}$.*

Proof. By construction, the terms of the disjunctive normal form of F are all of the minimal subsets having nonnull intersection with every circuit in $\mathcal{C}$. Theorem 5 guarantees that these terms are precisely the cobases of S, so that R is a minimum weight cobase of S.

The direct algorithm has a run time which grows exponentially with $|\mathcal{C}|$. This is to be expected, of course, because the problem (minimum weight hitting set) is NP-*hard* [1], [4].

4.2. An iterative method. In an effort to economize on run time, we propose an iterative method which constructs an independence system S_i on E with circuits $\mathcal{C}_i\subseteq\mathcal{C}$ such that the minimum weight cobases of S_i include at least one minimum weight cobase of S. Although the iterative process still has a run time which grows exponentially with $|\mathcal{C}_i|$, a savings occurs when $|\mathcal{C}_i|<|\mathcal{C}|$.

The iterative method relies upon a version of the greedy algorithm to provide at each stage not only a certain cobase Q but also a collection $\mathcal{Q}$ of circuits having a one-to-one correspondence with the elements of Q.

```
THE GREEDY ALGORITHM WITH CIRCUIT FINDING IN $S=(E,\mathcal{I})$.
1. Set $I=\varnothing$, $Q=\varnothing$, $\mathcal{Q}=\varnothing$.
2. While $E\neq\varnothing$ do
   a. let $e$ be an element of $E$ having largest weight;
   b. set $E=E-e$;
   c. if $I+e\in\mathcal{I}$, then set $I=I+e$
         else find a circuit $C\subseteq I+e$ $(e\in C)$;
              set $Q=Q+e$;
              set $\mathcal{Q}=\mathcal{Q}+C$.
```

The iterative process applies the direct Boolean algorithm to a sequence of independence systems S_i on E having circuit sets $\mathcal{C}_i$ which form an increasing sequence of subsets of $\mathcal{C}$,

$$\mathcal{C}_1\subset\mathcal{C}_2\subset\cdots\subset\mathcal{C}_i\cdots\subseteq\mathcal{C}.$$

The direct algorithm produces a minimum weight cobase R_i of S_i (Theorem 7). To see if R_i is also a cobase of S, the iterative process applies the greedy algorithm with circuit finding to the independence system $S\backslash R_i$, whose ground set is $E\backslash R_i$ and whose circuits are those circuits in $\mathcal{C}$ that are disjoint from R_i. The results are Q_i, a cobase of $S\backslash R_i$, and $\mathcal{Q}_i$, a collection of certain circuits of $S\backslash R_i$. Clearly, Q_i is a minimal subset of $E\backslash R_i$ having nonnull intersection with every circuit in $\mathcal{C}$ that R_i does not intersect. Thus, $R_i \cup Q_i$ intersects every circuit in $\mathcal{C}$. If $Q_i \neq \emptyset$, then R_i is not a cobase of S because it does not intersect every circuit in $\mathcal{C}$. In this case, $\mathcal{Q}_i$ is united with $\mathcal{C}_i$ to form a larger set of circuits $\mathcal{C}_{i+1}$, and the process is repeated for $i=i+1$. If $Q_i=\emptyset$, then R_i intersects every circuit in $\mathcal{C}$. Moreover, R_i is minimal with respect to this property because, if $X \subset R_i$ intersects every circuit in $\mathcal{C}$, then X intersects every circuit in $\mathcal{C}_i \subseteq \mathcal{C}$, contradicting the fact that R_i is a *minimal* subset intersecting every circuit in $\mathcal{C}_i$. Thus, when $Q_i=\emptyset$, R_i is a cobase of S. In addition, R_i is a minimum weight cobase of S because, if R were a cobase of S with $w(R)<w(R_i)$, then R would intersect every circuit in $\mathcal{C}_i$, contradicting the fact that R_i is a *minimum weight* subset intersecting every circuit in $\mathcal{C}_i$. This proves the correctness of the following procedure.

THE ITERATIVE BOOLEAN ALGORITHM.

1. Set $i=0$, $\mathcal{C}_0=\emptyset$, $R_0=\emptyset$.
2. Apply the greedy algorithm with circuit finding to $S\backslash R_i$, the independence system on $E\backslash R_i$ whose circuits are the circuits in $\mathcal{C}$ that are disjoint from R_i, obtaining a cobase Q_i and certain circuits $\mathcal{Q}_i$ of $S\backslash R_i$.
3. If $Q_i=\emptyset$, then stop, having found a minimum weight cobase R_i of S.
4. Set $i=i+1$, $\mathcal{C}_i=\mathcal{C}_{i-1}\cup\mathcal{Q}_{i-1}$.
5. Apply the direct Boolean algorithm to the independence system S_i on E with circuits $\mathcal{C}_i$, obtaining R_i, a minimum weight cobase of S_i.
6. Go to 2.

THEOREM 8. *The iterative Boolean algorithm produces a minimum weight cobase R_i of S.*

Besides the possibility of running faster, the iterative method is to be preferred over the direct method whenever the circuits in $\mathcal{C}$ are not known explicitly. In order to use the direct method, one must first construct all of the circuits, while the iterative method constructs only enough circuits to solve the problem.

5. Modifications of the iterative Boolean algorithm. At step 3 of the direct Boolean algorithm (step 5 of the iteration) one need not produce the entire disjunctive normal form of F_i and compute all the weight sums. Instead during the expansion, one should initially discard any terms having a weight sum greater than

$$U_i = \min_{1 \leq j \leq i-1} w(R_j \cup Q_j).$$

Such terms cannot be minimum weight cobases of S because the $R_i \cup Q_i$ are supersets of cobases of S, and U_i is, thus, an *upper bound* on the minimum weight we seek. Furthermore, one should save only terms of F_i having the currently minimum weight sum, and these should be discarded whenever a term of smaller weight arises. Thus, during the expansion, a truncation value should be initialized to U_i and then be continually updated to the currently minimum weight sum. Its final value is $w(R_i)$, a *lower bound.*

One could either save exactly one term at any time during the expansion or else save all of the terms of currently minimum weight. In the latter case, one of the final terms would have to be chosen as R_i. This might be one that intersects the most circuits in $\mathscr{C}$. However, in the presence of a large number of choices for R_i, an arbitrary choice seems to be warranted.

Similarly, one can alter the algorithm to find *all* of the minimum weight cobases of S, provided that the circuits have only positively weighted elements. The direct algorithm needs to save all of the currently minimum weight terms during expansion of F_i, and step 5 of the iteration needs to choose a final one to be R_i. When R_i is known to be a minimum weight cobase of S, all of the other minimum weight cobases of S are among the minimum weight terms of F_i. These terms only need to be chosen successively as R_i and to be tested by steps 2 and 3 of the iteration. This can be seen as follows. If the circuits have only positively weighted elements, then the same is true of the cobases. Let R be any other minimum weight cobase of S. Then R intersects every circuit in $\mathscr{C}_i$, just as R_i does, and $w(R)=w(R_i)$. Suppose $X\subset R$ also intersects every circuit in $\mathscr{C}_i$. Then $w(X)<w(R)$, contradicting the fact that R_i is a *minimum weight* subset intersecting every circuit in $\mathscr{C}_i$. Hence, R is a cobase of S_i having minimum weight and must appear among the minimum weight terms of F_i.

6. Matroids. The following theorem serves as three different, but equivalent, definitions of a matroid. For a proof see [4], [5], [8].

THEOREM 9. *Let $S=(E,\mathscr{I})$ be an independence system. Then the following statements are equivalent.*

1. *S is a matroid.*
2. *If $U, V\in\mathscr{I}$ and $|U|=|V|+1$, then there exists $e\in U\backslash V$ such that $V+e\in\mathscr{I}$.*
3. *If $A\subseteq E$, then all maximal independent subsets of A have the same cardinality.*
4. *The greedy algorithm correctly solves the combinatorial optimization problem for S with any nonnegative weight function.*

Letting $A=E$ in Theorem 9, we obtain

COROLLARY 2. *All bases of a matroid have the same cardinality.*

Therefore, when an independence system S is known to be a matroid, the greedy algorithm should be used to optimize over S. Otherwise, the iterative Boolean algorithm may be used.

7. An example. The Boolean formula in Table 1 is constructed by the direct Boolean algorithm to optimize over $S=(E,\mathscr{I})$, where $E=\{X1, X2,\cdots, X40\}$, the circuits $\mathscr{C}$ are given by the 50 factors in the formula, and the weights are shown on the right. In this computer printed equation, * and + are used instead of $\wedge$ and $\vee$.

The direct Boolean algorithm, modified to truncate initially on a given value and then to steadily reduce the truncation value until *all* of the minimum weight cobases are found, was implemented in a SETS user program [6], [7]. When the SETS program was applied directly to F with an initial truncation value of 1368, it got the answer in 348 seconds of run time on the CDC 7600. However, the same SETS program used at step 5 of the iterative algorithm solved the problem in a total of 42 seconds. In six iterations the SETS program expanded products of 8, 13, 16, 18, 20, and 21 factors, truncating them initially on respective upper bound values of 1632, 1604, 1604, 1484, 1460, and 1441. The final product of only 21 out of 50 factors is shown in Table 2.

The only minimum weight term in the disjunctive normal form of F_6 is

$$R_6=X1*X4*X6*X10*X11*X16$$

with $w(R_6)=1368$. Since R_6 intersects all 50 circuits, it is the unique solution. The greedy algorithm (a FORTRAN subroutine) consumed a total run time of 0.01 seconds.

It produced the initial upper bound $w(Q_0)=1632$ for a cobase Q_0 of S. Since the minimum weight for a cobase of S is 1368, S is not a matroid.

8. Application to fire protection. Let $\mathcal{N}(V, A)$ be a *network* modeling the possibilities of fire spread in a complex facility such as a nuclear reactor power plant. The *vertices* V represent areas, and the *directed arcs* A represent fire barriers between

TABLE 1

$$
\begin{aligned}
F = {} & (X8+X9+X10+X14+X18+X19+X21+X29+X35+X38)\\
& *(X6+X8+X9+X14+X23+X30+X31+X32+X33+X40)\\
& *(X4+X9+X10+X11+X12+X14+X18+X24+X33+X35)\\
& *(X3+X10+X12+X13+X20+X29+X34+X35+X37+X38)\\
& *(X10+X11+X14+X15+X20+X21+X22+X29+X35+X36)\\
& *(X4+X5+X11+X17+X18+X26+X33+X34+X36+X40)\\
& *(X4+X5+X9+X15+X16+X23+X30+X36+X38+X39)\\
& *(X3+X4+X6+X7+X11+X13+X16+X30+X32+X37)\\
& *(X2+X6+X10+X13+X14+X19+X24+X27+X33+X36)\\
& *(X1+X4+X5+X9+X10+X12+X18+X19+X36+X37)\\
& *(X4+X8+X10+X13+X14+X17+X19+X25+X30+X37)\\
& *(X3+X4+X6+X17+X21+X28+X29+X30+X37+X39)\\
& *(X1+X6+X7+X18+X19+X21+X23+X24+X32+X35)\\
& *(X16+X17+X18+X19+X21+X24+X27+X30+X31+X32)\\
& *(X1+X3+X12+X13+X15+X19+X23+X24+X28+X36)\\
& *(X14+X15+X16+X18+X20+X21+X22+X23+X24+X28)\\
& *(X8+X11+X14+X20+X22+X23+X29+X38+X39+X40)\\
& *(X4+X5+X10+X17+X18+X20+X22+X28+X33+X36)\\
& *(X5+X6+X16+X19+X25+X26+X28+X31+X34+X40)\\
& *(X1+X4+X8+X12+X14+X16+X17+X29+X31+X40)\\
& *(X3+X10+X13+X16+X17+X24+X27+X30+X31+X35)\\
& *(X1+X3+X4+X8+X9+X14+X19+X23+X25+X33)\\
& *(X3+X4+X10+X11+X12+X14+X22+X23+X35+X37)\\
& *(X7+X14+X16+X17+X26+X27+X32+X33+X34+X37)\\
& *(X2+X6+X20+X21+X23+X25+X30+X31+X36+X39)\\
& *(X4+X7+X9+X15+X17+X19+X20+X23+X24+X35)\\
& *(X3+X4+X6+X8+X11+X18+X21+X28+X33+X39)\\
& *(X2+X8+X9+X10+X12+X19+X20+X21+X22+X25)\\
& *(X5+X6+X22+X25+X27+X30+X33+X38+X39+X40)\\
& *(X2+X5+X6+X11+X15+X17+X18+X21+X24+X29)\\
& *(X1+X3+X7+X10+X11+X18+X23+X34+X36+X37)\\
& *(X3+X13+X16+X17+X19+X27+X29+X30+X34+X39)\\
& *(X5+X6+X10+X12+X14+X22+X26+X27+X33+X37)\\
& *(X6+X14+X16+X17+X18+X20+X21+X26+X29+X32)\\
& *(X2+X13+X14+X16+X19+X24+X27+X34+X36+X37)\\
& *(X1+X3+X8+X17+X18+X21+X33+X36+X38+X39)\\
& *(X3+X5+X10+X11+X12+X17+X21+X22+X24+X29)\\
& *(X2+X4+X7+X22+X23+X27+X33+X36+X37+X39)\\
& *(X3+X11+X13+X15+X17+X32+X34+X35+X36+X39)\\
& *(X3+X4+X7+X10+X11+X12+X21+X28+X29+X37)\\
& *(X7+X8+X15+X16+X17+X21+X29+X34+X38+X39)\\
& *(X6+X7+X12+X13+X17+X21+X22+X27+X29+X38)\\
& *(X2+X5+X14+X16+X17+X20+X25+X31+X36+X38)\\
& *(X2+X3+X4+X6+X12+X23+X28+X34+X38+X39)\\
& *(X6+X9+X10+X17+X20+X26+X30+X31+X35+X36)\\
& *(X1+X5+X7+X9+X17+X30+X32+X33+X34+X39)\\
& *(X3+X6+X10+X12+X13+X19+X21+X27+X35+X39)\\
& *(X9+X11+X15+X21+X24+X28+X32+X35+X36+X40)\\
& *(X1+X11+X12+X13+X15+X17+X20+X22+X24+X28)\\
& *(X3+X6+X14+X18+X22+X23+X24+X30+X34+X38)
\end{aligned}
$$

XI	$w(XI)$
$X1$	1
$X2$	73
$X3$	82
$X4$	113
$X5$	184
$X6$	203
$X7$	205
$X8$	212
$X9$	252
$X10$	262
$X11$	282
$X12$	401
$X13$	409
$X14$	460
$X15$	480
$X16$	507
$X17$	522
$X18$	526
$X19$	547
$X20$	581
$X21$	585
$X22$	606
$X23$	623
$X24$	672
$X25$	675
$X26$	695
$X27$	704
$X28$	704
$X29$	739
$X30$	745
$X31$	768
$X32$	799
$X33$	837
$X34$	870
$X35$	887
$X36$	893
$X37$	909
$X38$	938
$X39$	939
$X40$	995

TABLE 2

$$\begin{aligned}
F_6 = {} & (X10+X11+X14+X15+X20+X21+X22+X29+X35+X36)\\
& *(X\ 4+X\ 5+X11+X17+X18+X26+X33+X34+X36+X40)\\
& *(X16+X17+X18+X19+X21+X24+X27+X30+X31+X32)\\
& *(X\ 8+X11+X14+X20+X22+X23+X29+X38+X39+X40)\\
& *(X\ 3+X11+X13+X15+X17+X32+X34+X35+X36+X39)\\
& *(X\ 6+X\ 7+X12+X13+X17+X21+X22+X27+X29+X38)\\
& *(X\ 9+X11+X15+X21+X24+X28+X32+X35+X36+X40)\\
& *(X\ 3+X10+X12+X13+X20+X29+X34+X35+X37+X38)\\
& *(X\ 1+X\ 6+X\ 7+X18+X19+X21+X23+X24+X32+X35)\\
& *(X14+X15+X16+X18+X20+X21+X22+X23+X24+X28)\\
& *(X\ 5+X\ 6+X16+X19+X25+X26+X28+X31+X34+X40)\\
& *(X\ 2+X\ 4+X\ 7+X22+X23+X27+X33+X36+X37+X39)\\
& *(X\ 8+X\ 9+X10+X14+X18+X19+X21+X29+X35+X38)\\
& *(X\ 4+X\ 5+X10+X17+X18+X20+X22+X28+X33+X36)\\
& *(X\ 6+X\ 9+X10+X17+X20+X26+X30+X31+X35+X36)\\
& *(X\ 4+X\ 8+X10+X13+X14+X17+X19+X25+X30+X37)\\
& *(X\ 7+X14+X16+X17+X26+X27+X32+X33+X34+X37)\\
& *(X\ 6+X\ 8+X\ 9+X14+X23+X30+X31+X32+X33+X40)\\
& *(X\ 1+X\ 3+X\ 8+X17+X18+X21+X33+X36+X38+X39)\\
& *(X\ 4+X\ 7+X\ 9+X15+X17+X19+X20+X23+X24+X35)\\
& *(X\ 1+X11+X12+X13+X15+X17+X20+X22+X24+X28)
\end{aligned}$$

the areas. An arc joins one vertex of V to another but is disjoint from the vertices. Both the arcs and vertices are weighted with *fire protection costs* $w(e) \geqq 0$, $e \in E = V \cup A$. For example, a vertex weight could be the cost of an adequate sprinkler system for the area, and an arc weight could be the cost of increasing the fire rating of the barrier so as to prevent fire spread along the arc. Since fire cannot spread through a zero weighted arc or vertex, such arcs and vertices could be deleted from the model. When a vertex is deleted, all its incident arcs are deleted also. Vital components of safety systems are located at certain vertices called *target vertices*, and a minimal set of target vertices whose destruction could have unacceptable consequences is called a *target set*. Let $\mathcal{T} = \{T_1, T_2, \cdots, T_k\}$ be the *collection of target sets* to be protected.

Single-source fires are the most likely accidental fires. Such a fire could start at any vertex and spread along positively weighted arcs and vertices. Let an *$s-t$ net* for $(\mathcal{N}, \mathcal{T})$ be defined as a minimal, positively weighted, subnetwork $\mathcal{N}'$ in $\mathcal{N}$ consisting of a source vertex s, a target set $T_i \in \mathcal{T}$, and other vertices and arcs such that, for every $t \in T_i$, there is a path in $\mathcal{N}'$ from s to t. An $s-t$ net, then, is a particular way that a single-source fire could start, spread to, and destroy all the vertices in a target set. To prevent this, we seek a *fire protection set* $R \subseteq E$ whose deletion (protection) yields a subnetwork $\mathcal{N} \backslash R$ not containing any $s-t$ nets. The *cost of a fire protection set* is the sum of the weights of its elements, and the *minimum cost fire protection problem* is to find all of the minimum cost fire protection sets for $(\mathcal{N}, \mathcal{T})$.

This problem can be viewed as one of optimization over an independence system $S = (E, \mathcal{I})$, where $E = V \cup A$ and the circuits $\mathcal{C}$ are the $s-t$ nets. From Corollary 1 the independent subsets of E are the subnetworks not containing an $s-t$ net. The cobases of S are the minimal fire protection sets. The iterative Boolean algorithm, modified to find all the solutions, needs a greedy algorithm which produces $s-t$ nets by a polynomial-run-time pathfinding method. Then, the iterative procedure will produce all of the minimum weight cobases R, which are the minimum cost fire protection sets. This application is presented in greater detail in [2], [3] where computational results are given for a very similar iterative procedure which uses a maxflow-mincut calculation in place of the greedy algorithm.

REFERENCES

[1] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractibility: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[2] B. L. HULME, A. W. SHIVER AND P. J. SLATER, *Computing minimum cost fire protection*, SAND82-0809, Sandia National Laboratories, Albuquerque, NM, June 1982.

[3] ———, *A Boolean algebraic analysis of fire protection*, Workshop on Algebraic Structures in Operations Research, R. E. Burkard, R. A. Cuninghame-Green and U. Zimmermann, eds., Annals of Discrete Mathematics, North-Holland, Amsterdam, to appear.

[4] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[5] D. J. A. WELSH, *Matroid Theory*, Academic Press, London, 1976.

[6] R. B. WORRELL, *Set equation transformation system* (SETS), SLA-73-0028A, Sandia National Laboratories, Albuquerque, NM, January 1975.

[7] ——, *Notes on changes to the* SETS *program*, unpublished manuscript available from the author.

[8] U. ZIMMERMANN, *Linear and Combinatorial Optimization in Ordered Algebraic Structures*, Vol. 10, Annals of Discrete Mathematics, North-Holland, Amsterdam, 1981.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

013

DISCRETE MATHEMATICS IN VOTING AND GROUP CHOICE

PETER C. FISHBURN†

Abstract. The diversity of mathematics used in research on voting procedures and group decision processes is illustrated through discussions of representative systems, cyclic majorities, ranking paradoxes, impossibility theorems, approval voting, and proportional representation.

AMS subject classifications. 90A08

1. Paradoxes and profiles.

The aim of this paper is to illustrate through selected topics the diversity of mathematics that has been brought to bear on the analysis and design of voting procedures and group decision processes. In the course of our discussion we shall encounter discrete ranking structures, nested hierarchies of sign functions, finite topologies, combinatorial impossibility theorems, integer optimization problems, linear separation lemmas, and matters of probability, all of which are motivated by fundamental concerns of group decision making. We shall also meet paradoxes inherent in the subject and mention a few challenging open problems. I shall say more about the topics of ensuing sections after a brief historical note.

Our story begins with the Enlightenment figures Jean-Charles de Borda and the Marquis de Condorcet. Borda [3] proposed a now familiar method of electing one candidate from three or more nominees based on ranked voting. Condorcet [7] counterproposed that a candidate who can defeat each of the others in pairwise majority comparisons should be elected.

To be precise, suppose each of n voters, indexed by i from 1 to n, has a best-to-worst order $>_i$ on the $m \geqq 3$ candidates in the nominee set X. The list of these orders $(>_1, >_2, \cdots, >_n)$ is a *voter preference profile.* For now, assume that each $>_i$ is a total order, with no ties or individual indifference. Borda proposed that the elected candidate should be an $x \in X$ that maximizes

$$\sum_{i=1}^{n} |\{y\colon x >_i y\}|,$$

which equals the total points awarded to x when $m-1, m-2, \cdots, 0$ points are awarded respectively to each voter's first choice, second choice, $\cdots$, last choice. On the other hand, Condorcet argued that x ought to be elected if

$$|\{i\colon x >_i y\}| > |\{i\colon y >_i x\}| \quad \text{for each } y \neq x \text{ in } X,$$

i.e., if x is a *majority candidate.*

Condorcet was quick to point out that some profiles have no majority candidate. Thus, with $n = m = 3$, if the voters' orders on $X = \{a, b, c\}$ are abc, cab, and bca, then every candidate loses to one of the other two under majority comparison. This is the preeminent paradox of voting: it is known also as *Condorcet's phenomenon* and as the phenomenon of *cyclic majorities.* Extensions of Condorcet's majority method to profiles that have no majority candidate are reviewed and compared in Fishburn [15]: see also Young and Levenglick [48].

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Bell Laboratories, Murray Hill, New Jersey 07974.

Condorcet also noted a second paradox that established a fundamental incompatibility between his majority principle and Borda's positional-scoring procedure. Namely, there are profiles with a majority candidate x who will not be elected by *any* Borda-type point-count method that awards more points to a first choice than a second choice, more points to a second choice than a third choice, and so forth. Suppose $n=7$, $X=\{a, b, x\}$, and

3 voters have *xab*,
2 voters have *abx*,
1 voter has *axb*,
1 voter has *bxa*.

Then x has a 4-to-3 majority over each of a and b. However, when $w_1 > w_2 > w_3$ points are awarded for first, second, and third choices, a always gets more points $(3w_1+3w_2+w_3)$ than x $(3w_1+2w_2+2w_3)$.

Many other voting paradoxes [11], [19], [21], [36] have been uncovered in recent years, but the original phenomenon of cyclic majorities has remained as the principal impetus behind research in the theory of voting. One branch of this research, which addresses the likelihood of cyclic majorities and gets deeply involved in combinatorial probability, is discussed in § 3. Another branch, discussed in § 4, stems from Arrow's impossibility theorem for transitive social rankings based on transitive individual preference orders [1]. The latter section also considers questions motivated by Borda's scoring procedure.

Section 5 examines a simplified voting method, called approval voting [4],]5], that only asks voters to vote for the candidates they find acceptable, without ranking these choices. This method has been proposed as a practicable alternative to procedures in popular use such as the plurality ("vote for one") procedure. Despite its simplicity, approval voting involves interesting mathematical as well as practical issues.

The three sections of the paper just mentioned focus on the choice from—or the ranking of—a set X of three or more candidates. These sections are bracketed by two others that are more concerned with institutional configurations. The first of these, § 2, discusses hierarchical structures for voting in which the outcomes of votes in lower councils act like votes in higher councils. The point of this section is that even the most elementary matter of group decision, say to enact or defeat a legislative proposal, can give rise to interesting mathematics. The second institutional section, § 6, considers the problem of proportional representation in a legislature based primarily on elections in single-member districts. It proposes a variable-sized-legislature model in which seats are added to those won in the districts in an attempt to get closer to proportional representation for the parties in the legislature.

2. Nested hierarchies and linear separation. We shall suppose throughout this section that a vote is to be taken on a single issue and let v_i equal 1 if voter i votes *for* the issue, 0 if i abstains, and -1 if i votes *against* the issue. Each $v=(v_1, v_2, \cdots, v_n)$ in $V=\{1,0,-1\}^n$ is a *ballot response profile*, and a function

$$f: V \to \{1,0,-1\}$$

is a *group decision* (social choice) *function*. We interpret $f(v)=1$ as *passage* of the issue, $f(v)=-1$ as *defeat*, and $f(v)=0$ as a tie vote that requires further action.

Two facts about f make it especially attractive from a mathematical viewpoint. First, its codomain is the same set $\{1,0,-1\}$ used for each dimension in the domain V, and each of its three elements has a similar interpretation in the two places. Second,

V is a set of integral lattice points embedded in $\mathbb{R}^n$. Imagine a hyperplane through $\mathbb{R}^n$ that separates $f^*(1)=\{v: f(v)=1\}$ from $f^*(-1)$, and perhaps contains $f^*(0)$. I shall return to this shortly.

What conditions might we impose on f so that it is responsive to individuals' votes and fair to the issue at hand? Here are a few that have been suggested:

unanimity: $f(1,\cdots,1)=1$, $f(-1,\cdots,-1)=-1$;
duality: $f(-v)=-f(v)$;
anonymity: $f(v_1,\cdots,v_n)=f(v_{\sigma(1)},\cdots,v_{\sigma(n)})$ if σ is a permutation on $\{1,\cdots,n\}$;
monotonicity: $v \geqq v'$ (componentwise)$\Rightarrow f(v)\geqq f(v')$;
strong monotonicity: monotonicity, plus: $v>v'(v_i>v_i'$ for at least one i in addition to $v\geqq v')$ and $f(v')=0 \Rightarrow f(v)=1$.

You should have no trouble interpreting these. For example, duality says that if all votes are reversed (abstentions remaining unchanged) then the group decision is reversed; the second part of strong monotonicity says that a tie can be broken in favor of the issue if some voter votes more in favor of the issue.

Many specific and generic group decision functions can be characterized by such conditions and efficiently expressed with the use of sign functions [12]. For every real vector $(a_1,\cdots,a_K)$ let

$$\mathbf{s}(a_1,\cdots,a_K)=\text{sign}\,(\textstyle\sum a_k),$$

and let $(a_1,\cdots,a_K)\cdot(b_1,\cdots,b_K)=\sum a_k b_k$. The *simple majority function* is defined by

$$f(v)=\mathbf{s}(v) \quad \text{for all } v\in V,$$

and characterized [32] as the unique function that is dual, anonymous and strongly monotonic.

We define f as a *weighted majority function* if there is a $\rho>0_n$ (whose components, without loss of generality, can be presumed to be integers) such that

$$f(v)=\mathbf{s}(\rho\cdot v) \quad \text{for all } v\in V.$$

Here ρ_i is the "weight" of the ith voter or voting unit. The family of weighted majority functions is characterized as those f that are unanimous, monotonic, and satisfy

strong duality: if $K\in\{1,2,\cdots\}$, $v^k\in V$ for all k, and $\sum v^k=0_n$, then $\min\{f(v^k): k\leqq K\}+\max\{f(v^k): k\leqq K\}=0$.

When $\sum_n v^k=0_n$, each voter has the same number of *for* as *against* votes in the K ballot response profiles. Strong duality says that the issue passes for at least one v^k if and only if it is defeated for another v^k.

The proof that a unanimous, monotonic, and strongly dual f is a weighted majority function is facilitated by a standard separating-hyperplane lemma or theorem of the alternative.

LEMMA 1. *Exactly one of* (A) *and* (B) *is true when* $1\leqq J\leqq K$ *and* $a^1,\cdots,a^K$ *are rational vectors in* $\mathbb{R}^N$:

(A) *There is an integral* $\rho\in\mathbb{R}^n$ *such that*

$$\rho\cdot a^k>0 \quad \textit{for } 1\leqq k\leqq J,$$

$$\rho\cdot a^k=0 \quad \textit{for } J<k\leqq K.$$

(B) *There are nonnegative integers $r_1, \cdots, r_J$ at least one of which is positive, and integers $r_{J+1}, \cdots, r_K$ such that*

$$\sum_{k=1}^{K} r_k a^k = 0_n.$$

Assume f is unanimous, monotonic and strongly dual. By unanimity, $f^*(1)$ is not empty. Suppose no integral ρ gives $\rho\cdot v > 0$ on $f^*(1)$ and $\rho \cdot v = 0$ on $f^*(0)$, as required by weighted majority. Then Lemma 1 (B)—reverse negative r_j under duality in $f^*(0)$; replicate v in $f^*(1)$ and $f^*(0)$ by their r_k—implies that there are $v^1, \cdots, v^H \in V$ with $\sum v^h = 0_n$, $f(v^h) \geqq 0$ for all h, and $f(v^h) = 1$ for some h. But this contradicts strong duality. Hence our supposition is false, and duality, monotonicity and unanimity then imply that $\rho > 0_n$.

Murakami [35] developed the idea of representing other majority-like functions by nested hierarchies of weighted majority functions. These can be constructed recursively by defining $\mathbf{s}(f_1, \cdots, f_K): V \to \{1, 0, -1\}$ for every list $(f_1, \cdots, f_K)$ of $f_k: V \to \{1, 0, -1\}^n$ by

$$\mathbf{s}(f_1, \cdots, f_K)(v) = \mathbf{s}(f_1(v), \cdots, f_K(v)).$$

To illustrate, suppose body k ($k = 1, 2, 3$) uses the simple-majority function f_k within its body to vote on an issue, and the three outcomes are then aggregated by simple majority to decide the fate of the issue. The overall group decision function of this process is $f = \mathbf{s}(f_1, f_2, f_3)$.

More generally, let D_i be the dictatorial projection for voter i defined by $D_i(v) = v_i$, let $\mathcal{R}_0 = \{D_1, \cdots, D_n\}$, and for each positive integer t let

$$\mathcal{R}_t = \{\mathbf{s}(f_1, \cdots, f_K): K \in \{1, 2, \cdots\} \text{ and } f_1, \cdots, f_K \in \mathcal{R}_{t-1}\}.$$

It is easily seen that $\mathcal{R}_1$ is the set of weighted majority functions and, since $\mathbf{s}(f) = f$, that $\mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq \mathcal{R}_2 \subseteq \cdots$. We refer to $\mathcal{R} = \bigcup \mathcal{R}_t$ as the set of *representative systems* for n voters.

Representative systems have been studied extensively by Murakami [35] and Fishburn [9], [10], [14], [17]. Their first complete characterization in terms of conditions on f appeared in [9]. Four conditions on f suffice for $f \in \mathcal{R}$, namely unanimity, duality, monotonicity, and a "dual partition" condition that generalizes strong duality. Proofs of the sufficiency of these conditions, which use Lemma 1 or a related separation theorem, appear in [9], [10].

Further investigations of representative systems have focused on the number of hierarchical levels that are needed to express all $f \in \mathcal{R}$ for n voters. Since $\mathcal{R}$ is finite for each n, there is a smallest t such that $\mathcal{R}_t = \mathcal{R}$ for n. Let $\mu(n)$ denote this smallest t for n. Fishburn [14], [17] proved that $\mu(n) = n - 1$ for $1 \leqq n \leqq 4$, $\mu(5) = \mu(6) = 4$, $\mu(n) \leqq n - 2$ for all $n \geqq 6$, and μ is unbounded. He also conjectured that $\mu(n)/n \to 0$ as n gets large. This conjecture was recently confirmed by Keiding [29] in a remarkable paper which shows that $\mu(n) \leqq \log_2(n(n-1)) + 5$.

3. Cyclic majorities: the root of the problem. We now return to the phenomenon of cyclic majorities introduced in § 1. As before, n is the number of voters, $m \geqq 3$ the number of candidates, and $>_i$ is voter i's preference order on the candidate set X. We shall assume that every $>_i$ is a total order (no ties). For each voter preference profile $(>_1, \cdots, >_n)$, define the nonstrict simple majority relation on X by

$$x \geqq_M y \quad \text{if } |\{i: x >_i y\}| \geqq |\{i: y >_i x\}|,$$

and let $>_M$ and $=_M$ be respectively the asymmetric and symmetric parts of $\geqq_M$.

McGarvey [34] showed that if n is large enough relative to m, then for every asymmetric relation on X there is a profile that has this relation as its $>_M$. For each $m \geqq 3$ let $\sigma(m)$ be the smallest n for which this is true. Stearns [45] proved that $\sigma(m) \leqq m+1$ for odd m, $\sigma(m) \leqq m+2$ for even m, and these bounds are tight when $m \leqq 5$. He also showed that $\sigma(m) > [(\log 3)/2]m/(\log m)$. Erdös and Moser [8] then noted that $\sigma(m) \leqq c_1 m/(\log m)$ for a fixed constant c_1. Precise values of $\sigma(m)$ beyond the first few m are unknown, and the question of whether $\sigma(m)(\log m)/m$ tends to a limit is open.

How common are cyclic majorities? This has been intensely investigated; its present status is surveyed by Gehrlein [23]. One approach considers the proportion $p(m, n)$ of the $(m!)^n$ profiles that *have* majority candidates. If each voter independently chooses one of the $m!$ orders according to the uniform distribution, then $p(m, n)$ is the probability that the chosen profile has a majority candidate.

A quick check (hold one order fixed) shows that $p(3, 3) = 17/18$. Exact computations for $m > 3$ or $n > 3$ get complex very quickly. The most efficient method known for three candidates [24] uses

$$p(3, n) = 3^{-n+1} \sum \frac{n!}{n_1! n_2! n_3! n_4!} 2^{-(n_2+n_3)}$$

where $\sum$ is a triple sum with limits $\{0 \leqq n_1 \leqq (n-1)/2,\ 0 \leqq n_2 \leqq (n-1)/2 - n_1,\ 0 \leqq n_3 \leqq (n-1)/2 - n_1\}$ and $n_4 = n - n_1 - n_2 - n_3$, and the most efficient method known for three voters [25] uses

$$p(m, 3) = m \sum_{m_1=0}^{m-1} \sum_{m_2=0}^{m-1-m_1} \frac{(m-1-m_1)!(m-1-m_2)!}{m!(m-1-m_1-m_2)!(m_1+m_2+1)}.$$

When $m \geqq 4$ is even and n is odd, there is a nice recursion relationship for $p(m, n)$. The simplest case [33] is

$$p(4, n) = 2p(3, n) - 1.$$

The proof is instructive. Let $q(m, n) = 1 - p(m, n)$, the probability of *no* majority candidate. Given $|X| = 4$, X has four triples (three-element subsets), at most two of which can have cyclic majorities. Let a, b be respectively the number of the $K = (4!)^n$ profiles with 1, 2 cyclic triples. There are $a + 2b$ instances of cyclic triples in the K profiles, which divide equally among the four triples. Therefore $q(3, n) = (a+2b)/4K$. Now each case of b has no majority candidate in X, and exactly half the a cases have no majority candidate in X. Thus $q(4, n) = (a/2 + b)/K = (a+2b)/2K$, so $q(4, n) = 2q(3, n)$, which is the same as $p(4, n) = 2p(3, n) - 1$.

The recursion for $m = 6$ and n odd is

$$p(6, n) = 3p(5, n) - 5p(3, n) + 3,$$

and the general case [24] has

$$p(m, n) = \sum_{j=1}^{m/2} c(j, m)p(2j-1, n),$$

where c is independent of n. Unfortunately, there is no similar relationship for m odd.

The situation for a large number of voters is interesting. Guilbaud [28] observed that, with $p(m) = \lim_{n\to\infty} p(m, n)$,

$$p(3) = \frac{3}{4} + \frac{3}{2\pi} \sin^{-1}\left(\frac{1}{3}\right) \doteq 0.91226,$$

and Niemi and Weisberg [37] noted that $p(m)$ equals m times the $(m-1)$-dimensional normal positive orthant probability with all correlations equal to $\frac{1}{3}$. The best approximation for $p(m)$ presently known [25] is

$$p(m) \doteq \frac{9.33}{m+9.53} + (0.63)^{(m-3)/2},$$

which is accurate within one-half of one per cent for odd $m < 50$.

Several comparative results have been obtained along with the exact and approximate numerical results. Among other things, Kelly [30] proved that

$$p^*(m, n+1) > p^*(m, n) \quad \text{for } m \geqq 3 \text{ and odd } n \geqq 3$$
$$p^*(m, n+1) < p^*(m, n) \quad \text{for } m \geqq 3 \text{ and even } n \geqq 3,$$

where $p^*(m, n)$ is the probability (proportion of profiles) that a profile has a *nonstrict* majority candidate, i.e., an $x \in X$ such that $x \geqq_M y$ for all $y \in X$. The inequality reversal is caused by majority ties when n is even. Fishburn, Gehrlein, and Maskin [22] showed that

$$p(3, n) > p(3, n+2) \quad \text{for odd } n \geqq 1,$$
$$p(3, n) < p(3, n+2) \quad \text{for even } n \geqq 2,$$
$$p^*(3, n) > p^*(3, n+2) \quad \text{for all large even } n.$$

Again, ties cause the reversal in the last two lines. They also showed that

$$p(m, 3) > p(m+1, 3) \quad \text{for all } m \geqq 2.$$

The extent to which the preceding hold for other m and n remains open.

Another open problem, that involves correlations among orders, considers the special case of $m = n$. Let $p'(m, m)$ be the probability (under uniform choices) that there is a majority candidate, *given* that the ith voter has the ith candidate from an ordered set of m candidates ranked first $(i = 1, \cdots, m)$. Note that $p(2, 2) = 1/2$, $p'(2, 2) = 0$, $p(3, 3) = 17/18 = 0.944 \cdots$, and $p'(3, 3) = 3/4$. I conjecture that

$$p(m, m) > p'(m, m) \quad \text{for all } m \geqq 2.$$

This seems plausible because the conditioning event for p' would appear to inhibit the likelihood of a majority candidate.

4. Ranking paradoxes and impossibility theorems. The majority relation $\geqq_M$ provides one definition of social preference on X. However, as we know from Condorcet's phenomenon, $\geqq_M$ need not be an ordering when $m \geqq 3$, and $>_M$ could be any asymmetric relation or partial tournament on X.

Question. Is there any reasonable way of constructing a complete, transitive social preference relation $\succsim$ on X for every profile $(>_1, \cdots, >_n)$ of individual preference orders on X on the basis of binary comparisons between candidates when $m \geqq 3$?

According to Arrow's famous impossibility theorem [1], the answer is "no" if by "reasonable way" we mean that the relation $\succsim$ (with asymmetric part $>$) for each profile must satisfy the following conditions for all $x, y \in X$ and all profiles:

Pareto unanimity: $x > y$ if $x >_i y$ for all i;
binary independence: $x > y \Leftrightarrow x >' y$ whenever $\{i: x >_i y\} = \{i: x >'_i y\}$ and $\{i: y >_i x\} = \{i: y >'_i x\}$;
no dictator: there is no i such that $x >_i y \Rightarrow x > y$.

To appreciate the combinatorial insights of Arrow's theorem, I sketch the proof for $X=\{a, b, c\}$. Assume that $\succsim$ is a complete and transitive relation for every $(>_1, \cdots, >_n)$ that obeys Pareto unanimity and binary independence. Call nonempty coalition $I \subseteq \{1, \cdots, n\}$ *decisive* for x over y if $x > y$ whenever $x >_i y$ for all $i \in I$ and $y >_i x$ for all $i \in \{1, \cdots, n\} \backslash I$. We observe that some $\{i\}$ is decisive for some x over y, then note that this i is a dictator in the sense prohibited by the no-dictator condition.

By Pareto unanimity, $\{1, \cdots, n\}$ is decisive for x over y for all distinct x and y in X. It follows that there is a smallest I, say I^*, that is decisive on some pair. Assume for definiteness that I^* is decisive for a over b. Fix $i \in I^*$. We claim $I^*=\{i\}$. Otherwise, the profile

$$(c >_i a >_i b;\ a >_j b >_j c \text{ for } j \in I^* \backslash \{i\};\ b >_j c >_j a \text{ otherwise})$$

has $a > b$ (I^* decisive, binary independence), $c \succsim a$ (else $I^* \backslash \{i\}$ is decisive for a over c, contrary to I^* as the smallest decisive I), and $b \succsim c$ (else $\{i\}$ decisive for c over b); but $\{b \succsim c,\ c \succsim a,\ a > b\}$ violates the ordering assumptions on $\succsim$.

Hence $\{i\}$ is decisive for a over b. Consider a profile that has $c >_i a >_i b$ and $\{c >_j a,\ b >_j a\}$ for all $j \neq i$. Then $a > b$ (decisiveness) and $c > a$ (Pareto), hence $c > b$ (transitivity). Since the relation between c and b for $j \neq i$ is arbitrary, it follows from binary independence that $c > b$ whenever $c >_i b$. The argument just used can be reapplied, first with $a >_i b >_i c$ and $(b >_j c,\ b >_j a)$ for $j \neq i$, then for other permutations on $\{a, b, c\}$, to conclude that, for all distinct x and y in X, $x > y$ whenever $x >_i y$. Hence i is a dictator.

Arrow's monograph introduced a level of mathematics into the theory of elections that was well beyond anything seen previously in this area. His work has had a profound effect on subsequent research. One closely-related contribution is the Gibbard [27]-Satterthwaite [40] impossibility theorem. This says that every "reasonable" method for electing one of $m \geqq 3$ candidates is manipulable, i.e., there are situations in which a voter can ensure a personally-preferred outcome by voting contrary to his true preferences. Many other contributions are discussed in [10], [31], [38], [41], [42].

Arrow's condition that is violated by positional-scoring procedures and other methods for constructing social rankings is binary independence. If we insist on binary independence, reasonable social rankings on X cannot be obtained for all profiles; if we drop it, curious things can happen.

Here are two paradoxes for the Borda method. Let $X=\{a_1, a_2, \cdots, a_m\}$ and suppose that $(>_1, \cdots, >_n)$ yields the Borda point-total ranking $a_1 > a_2 > \cdots > a_m$ with a_1 first and a_m last. Now suppose that a_m withdraws, leaving $Y = X \backslash \{a_m\}$, with no change in the $>_i$ on Y. Recompute Borda point-totals as if a_m were never present. Under this recomputation we can get $a_{m-1} > \cdots > a_2 > a_1$, completely reversing the part of the original social ranking on Y. If you don't believe it, consider $(m, n)=(4, 7)$ with

$$(>_1, \cdots, >_7)=(cbax,\ cbax,\ cbax,\ baxc,\ baxc,\ axcb,\ axcb).$$

The second Borda paradox says that there are profiles in which a_1 is the original Borda winner, but a_1 is a Borda loser (recomputed after deletions from the original $>_i$) for *every* $Y \subset X$ that contains a_1 and at least one other candidate except for one such Y that has $|Y|=2$. The construction is explained in [13].

Related anomalies for more general positional-scoring rules are discussed in [18], [39]. Among the questions these results have motivated is: What types of positional-scoring rules are most likely to retain the winner when one or more losers are removed? A partial answer is given in [26]. Given any $m \geqq 3$, let $w=(w_1, \cdots, w_m)$ and $v=(v_1, \cdots, v_{m-1})$ be nonincreasing real vectors with $w_1 > w_m$ and $v_1 > v_{m-1}$. When

$|X| = m$, use w to compute a winner when w_j points are awarded to voter i's jth-ranked candidate, for all i and j. When $|Y| = m-1$, use v in a similar manner to compute a winner.

Suppose each of n voters chooses one of the $m!$ total orders on X as his ranking according to the uniform distribution. Given the winner from X as determined from w, choose *one* of the other candidates at random, delete this candidate from the rankings, and compute a new winner on the basis of v. Let $P_m(w, v)$ be the limit in n of the probability that the (new) v winner is identical to the (original) w winner.

It is shown in [26] that $P_m(w, v)$ is uniquely maximized by linear w and v, i.e., by Borda vectors or positive affine transformations of Borda vectors. The proof uses Slepian's theorem [43] that the positive orthant probability for a multivariate normal distribution with positive correlations increases as the correlations increase, and shows that the relevant correlation matrix is uniquely maximized when w and v are linear.

A new twist was introduced into the voting literature by Smith [44] and Young [47] that enabled them to give nice axiomatizations of positional-scoring rules. Their idea was to fix m but let n vary over all positive integers. Young's formulation goes as follows. Fix $m \geqq 2$, let T be the set of $m!$ total orders on X, and let

$$\Pi = \{\pi\colon T \to \{0, 1, 2, \cdots\}\colon \pi(\tau) > 0 \text{ for some } \tau \in T\}.$$

Each π is a summary profile that tells how many voters, $\pi(\tau)$, have each order τ in T, so $n(\pi) = \sum \pi(\tau) \geqq 1$. Given Π, a *social choice function* is a mapping $C\colon \Pi \to 2^X \backslash \{\varnothing\}$ that assigns a nonempty subset $C(\pi)$ of X, called the *choice set*, to each summary profile π. The crucial condition on C that makes use of the variable-n feature is

$$\textit{consistency}\colon C(\pi) \cap C(\pi') \neq \varnothing \Rightarrow C(\pi + \pi') = C(\pi) \cap C(\pi').$$

This says that if two disjoint voting groups of sizes $n(\pi)$ and $n(\pi')$ have at least one candidate in common in their choice sets, then the choice set of the combined group of size $n(\pi+\pi') = n(\pi) + n(\pi')$ shall consist of the common choices of the initial groups. Another condition used by Young is the following generalization of duality:

neutrality: if σ is a permutation on $\{1, \cdots, m\}$ that maps π into π_σ in the natural way, then $C(\pi_\sigma) = \sigma(C(\pi))$.

These two conditions, along with versions of continuity and monotonicity, imply that there is a positional-scoring vector $w = (w_1, \cdots, w_m)$ of the type noted above that determines $C(\pi)$ for all $\pi \in \Pi$ by maximum point totals. However, consistency and neutrality alone have interesting consequences. In particular, let W be the set of *all* $w \in \mathbb{R}^{m-1}$ ($w_m = 0$ is understood), define the lexicographic order $>_L$ on real vectors in the usual way as $(a_1, \cdots, a_K) >_L (b_1, \cdots, b_K)$ if $a \neq b$ and $a_k > b_k$ for the smallest k where $a_k \neq b_k$, and let $\pi[x]$ for each $\pi \in \Pi$ and $x \in X$ be the vector in $\mathbb{R}^{m-1}$ whose jth component is the number of orders (voters) in π that rank x in jth place. Then consistency and neutrality imply that there are $w^1, \cdots, w^K \in W$ with $K \leqq m-1$ such that, for all $x \in X$ and all $\pi \in \Pi$,

$$x \in C(\pi) \Leftrightarrow (w^1 \cdot \pi[y], \cdots, w^K \cdot \pi[y]) >_L (w^1 \cdot \pi[x], \cdots, w^K \cdot \pi[x]) \quad \text{for no } y \in X.$$

Young's proof of this lexicographic maximum characterization involves a complex separation argument with embedded convex cones that raises the question of whether a simpler proof is possible. Such a proof would appear to hinge on the question of whether there is a relatively simple combinatorial proof of

$$(*) \qquad \pi[x] = \pi[y] \Rightarrow \{x, y \in C(\pi) \text{ or } x, y \in X \backslash C(\pi)\},$$

which is an obvious consequence of Young's theorem. Given (*), the lexicographic representation follows easily from the separation lemma which says that if $A, B \in \mathbb{R}^n$, $A \neq \varnothing$, and the convex cone in $\mathbb{R}^n$ generated by $A \cup B$ does not intersect the negative of the convex cone generated by A, then there is a $w \in \mathbb{R}^n$ such that $w \cdot a \geqq 0$ for all $a \in A \cup B$ and $w \cdot a > 0$ for some $a \in A$. Our attempts to find a simple proof of (*) have failed.

5. Approval voting: power in simplicity. Approval voting, a topic of comparatively recent research [4], [5], [46], offers a practicable alternative to the widespread plurality and plurality-with-runoff methods. Each voter votes for a nonranked subset of candidates; the candidate with the most votes wins. Although this restricts voters' expressions of preference or approval to a simple form (unlike Borda's method, which he proposed as an alternative to plurality voting), it offers considerably more leeway than the vote-for-one method.

There are severe problems with the plurality methods that are not shared by approval voting. The most acute for ordinary plurality are the wasted-vote phenomenon (when a voter's favorite has little chance of winning) and division of votes between ideologically similar candidates that allows an overall weaker candidate to win. These are less severe when a runoff election is held between the top two candidates, but the runoff provision introduces problems of its own that involve strategic maneuvers on the first ballot and violations of monotonicity. To illustrate the latter, suppose 27 voters divide as follows when $X = \{a, b, c\}$:

preference order:	*abc*	*cab*	*bca*	*bac*	*cba*	*acb*
number of voters:	6	6	6	4	2	3

Under plurality-with-runoff, a and b go on to the runoff, where a beats b by 15 to 12. Now suppose that some of the voters change their orders in favor of the winner a by moving a toward the first position. In particular, suppose three of the four *bac* voters change to *abc* and the two *cba* voters change to *cab*. Then a and c go onto the runoff, where c beats a by 14 to 13. Hence plurality-with-runoff makes it possible for a potential winner to become a loser as he gains the approval of more voters.

An axiomatic characterization of approval voting [16] is quite similar to Young's axioms discussed in the preceding section. Here, let Π be the set of all $\pi: 2^X \to \{0, 1, 2, \cdots\}$ that have $\pi(A) > 0$ for some $A \subseteq X$, and let $\pi[x] = \sum \{\pi(A): x \in A\}$, the number of voters whose ballots contain x. A choice function $C: \Pi \to 2^X \backslash \{\varnothing\}$ is the *approval voting function* if

$$C(\pi) = \{x \in X: \pi[x] \geqq \pi[y] \text{ for all } y \in X\}.$$

Unlike the situation for (*), a fairly straightforward combinatorial proof shows that C is the approval voting function if and only if it satisfies the following conditions for all $\pi, \pi' \in \Pi$, all permutations σ on X, and all $A, B \subseteq X$:

consistency: $C(\pi) \cap C(\pi') \neq \varnothing \Rightarrow C(\pi + \pi') = C(\pi) \cap C(\pi')$;
neutrality: $C(\sigma(\pi)) = \sigma(C(\pi))$;
disjoint equality: $C(\pi) = A \cup B$ if $A \cap B = \varnothing$, $\pi(A) = \pi(B) = 1$, and $\pi(C) = 0$ otherwise.

The final condition distinguishes approval voting from other scoring procedures based on the present definition of Π.

Much of the research in approval voting has centered on questions of voter strategy. A *strategy* is any subset of X. Given a weak (ties allowed) preference order on X for a voter, strategy A *dominates* strategy B for this voter if for every possible profile of votes by other voters he prefers the outcome (C set) that obtains when he uses A as much as the outcome that obtains when he uses B, and strictly prefers the A-outcome to the B-outcome in at least one case. Call A *admissible* if it is not dominated by another strategy.

Although the definition of dominance involves preference between *subsets* (C sets), it turns out that it has a simple characterization if two basic assumptions are made about the relationship between preference between candidates and preference between subsets of candidates. First, if $x >_i y$, then $\{x\} >_i \{x, y\}$ and $\{x, y\} >_i \{y\}$. Second, if $A \cup B$ and $B \cup C$ are not empty and $x \succsim_i y$ for all $(x, y) \in (A \times B) \cup (A \times C) \cup (B \times C)$, then $A \cup B \succsim_i B \cup C$. Order topologies can then be used to talk about dominance under these assumptions. Given the voter's weak preference order $\succsim_i$ on X with symmetric part $>_i$, the *high topology* is

$$\mathscr{H} = \{A \subseteq X \colon y \in A \text{ and } x >_i y \Rightarrow x \in A\},$$

and its complement, the *low topology*, is

$$\mathscr{L} = \{A \subseteq X \colon x \in A \text{ and } x >_i y \Rightarrow y \in A\}.$$

Brams and Fishburn [4] prove that if $>_i \neq \varnothing$, then

$$A \text{ dominates } B \Leftrightarrow A \backslash B \in \mathscr{H} \backslash \{X\},\ B \backslash A \in \mathscr{L} \backslash \{X\} \quad \text{and} \quad A \neq B.$$

Given $H = \{X \colon y >_i x \text{ for no } y\}$ and $L = \{x \colon x >_i y \text{ for no } y\}$, they also show that A is admissible under approval voting if and only if $H \subseteq A$ and $A \cap L = \varnothing$. Thus if the voter's weak (parentheses denote ties) preference order on five candidates is $(ab)(cd)e$, then his admissible strategies are $\{a, b\}$, $\{a, b, c\}$, $\{a, b, d\}$, and $\{a, b, c, d\}$. Under plurality voting, similar definitions and analyses show that x is admissible if and only if $x \notin L$.

One consequence of this analysis is that if approval voting is used and if every voter has dichotomous preferences (divides X into two subsets, within each of which he is indifferent) and votes his (unique) admissible strategy, then the choice set C will equal the set of all nonstrict Condorcet candidates. Another consequence is that if a voter has dichotomous or trichotomous (divides X into three subsets . . .) preferences and uses an admissible strategy, then he votes sincerely in the sense that if he votes for x then he also votes for all candidates he prefers to x. Not only is approval voting superior to plurality voting in these respects, but it has advantages over every other nonranked single-ballot method—such as "vote for exactly two" and "vote for no more than three"—that counts votes in a similar way.

Because approval voting allows voters to vote for different numbers of candidates, it might seem less equitable than plurality voting. This is true in one sense but not in another. If there are *four or more* candidates, then voters who vote for about half of X have somewhat greater chances of influencing the outcome than voters who vote for only one or all but one candidate. On the other hand, as shown in [20], [46], plurality voting is significantly worse than approval voting in terms of expected utility gains to voters, and it causes greater disparities among voters' expected gains than does approval voting.

Concerns have been raised about approval voting that call for further research. In particular, what effects will approval voting have on:

(a) voter interest and turnout;
(b) campaign strategies and positions taken by candidates on important issues;
(c) the numbers of candidates who run for election;
(d) party structures?

Each concern has positive and negative possibilities. Well thought out models may help to evaluate the likelihoods of different effects, but in the final analysis only the use of approval voting in actual elections will tell the tale.

6. How to slice an expanding cake fairly. Proportional representation in legislatures has been promoted by several means, including multiple-member districts, at-large seats, and the addition of a fixed number of seats to those won in districts that are allocated on the basis of party vote proportions. My aim in this section is to discuss a proposal for moving toward proportional representation by the addition of a variable number of seats to those won in single-member districts that depends on the numbers of districts won by the parties and their overall vote proportions. Like Balinski and Young's study [2] of fair apportionment—e.g., allocation of seats in the United States House of Representatives to states based on population proportions—integer numbers of seats must be taken into account, but unlike their approach the size of the legislature is not fixed. I shall suggest a complete solution only for the two-party case. Good solutions for the N-party case remain an open issue.

Let M be the number of legislative districts, and let N be the number of political parties, indexed by i from 1 to N. In district elections, a voter votes for candidates and for a party. One candidate in each district is elected to the legislature. The district elections data are summarized by

$$s_i = \text{number of districts won by party } i,$$

$$p_i = \text{proportion of party vote for party } i,$$

with $s_1 + \cdots + s_N = M$ and $p_1 + \cdots + p_N = 1$.

The *base* of the legislature is $(s_1, \cdots, s_N)$. Party i is *underrepresented* in the base if $s_i/M < p_i$ and is *overrepresented* if $s_i/M > p_i$. Seats are added to the base (e.g., to losers in districts with the most votes among losers of their party) to form an *augmented legislature* $(s_1^*, \cdots, s_N^*)$ with

$$s_i \leqq s_i^* \leqq M \quad \text{for } i = 1, \cdots, N.$$

Let $M^* = s_1^* + \cdots + s_N^*$ be the size of the augmented legislature. Then s_i^*/M^* is the proportion of seats held by party i. The aim of proportional representation is to make s_i^*/M^* approximately equal to p_i for each i.

The basic question is how best to determine the s_i^*. Since many considerations impinge on the answer, there may be no best way, but clearly some methods can be judged to be better than others.

Brams and Fishburn [6] form one answer for $N = 2$ that is governed by rules for augmentation designed to keep the increase $M^* - M$ manageable, to honor the single-member district concept, to be fair to parties who win in the districts, to discourage strategic maneuvers by parties and voters, and to encourage parties to do as well as they can in every district. Their proposal is to add as many seats to the underrepresented party as possible without violating the rules for augmentation, except perhaps when the underrepresented party wins a majority of the party vote. This exceptional (and unlikely) case raises delicate political issues that are addressed in the paper but will not be discussed here.

We use three augmentation rules for $N = 2$:

RULE 1. *Seats are added only to the underrepresented party.*

RULE 2. *The overrepresented party's seat proportion in the augmented legislature shall not be less than its* p_i.

RULE 3. *The underrepresented party cannot achieve a greater seat proportion in the augmented legislature by losing in more districts, given the same* (p_1, p_2) *in both cases.*

To illustrate the effects of these rules, suppose $M=8$ and $(p_1, p_2)=(0.2, 0.8)$.

Situation 1. $(s_1, s_2)=(1, 7)$. Since $s_1/M<p_1$, party 1 qualifies for additions under Rule 1. However, since $(s_1+1)/(M+1)=2/9>p_1$, Rule 2 prevents any additions.

Situation 2. $(s_1, s_2)=(0, 8)$. Now Rules 1 and 2 allow two seats to be added to party 1 since $(s_1+2)/(M+2)=p_1$, or $s_2/(M+2)=0.8=p_2$. However, this would give party 1 a greater seat proportion than if it had won one district (Situation 1), violating Rule 3. Therefore, Rule 3 limits the addition to one seat when party 1 wins no district.

The effect of Rule 3 just illustrated accumulates as we work backwards from the most to the fewest districts that the underrepresented party could win with a fixed p_i, and severely limits the ability to achieve approximate proportional representation in some cases. To be more precise, let p denote the p_i of the underrepresented party, and assume that $0<p<1/2$. Also let

$$t = \text{integer part of } pM,$$

$$\alpha_0 = \text{integer part of } (pM-t)/(1-p) = 0 \text{ or } 1.$$

Suppose the maximum additions are made to the underrepresented party under Rules 1, 2 and 3. *Then, regardless of how many districts it wins, it gets* $t+\alpha_0$ *seats in the augmented legislature.* Note that this does not mean that it will have no incentive to win districts as well as party votes since its seat *proportion* in the augmented legislature increases if it wins more districts.

To illustrate this result, suppose $M=100$ and $p=0.493$. Then $t=40$, $\alpha_0=0$, and therefore the underrepresented party gets 40 seats altogether. If it wins in no district, its seat proportion is $40/140=0.286$; if it wins 20 districts, the proportion increases to $40/120=0.333$; with 40 wins, its proportion is 0.40. In general, we estimate that the increase to M in the two-party case will seldom exceed 20%.

Although the N-party case for $N \geqq 3$ is not ignored in [6], it is dealt with constructively only for situations in which there is one overrepresented party and a number of smaller, underrepresented parties. Extensions of Rules 1 through 3 are discussed along with other rules that are relevant only when $N \geqq 3$. The impression is given that these rules, taken together, are too restrictive to allow reasonable moves toward greater proportional representation. The most suitable relaxations to accommodate the latter goal are open to further investigation.

REFERENCES

[1] K. J. ARROW, *Social Choice and Individual Values*, Second ed., John Wiley, New York, 1963.

[2] M. L. BALINSKI AND H. P. YOUNG, *Fair Representation*, Yale Univ. Press, New Haven, CT, 1982.

[3] J.-C. DE BORDA, *Mémoire sur les élections au scrutin*, Histoire de l'Académie Royale des Sciences, 1781.

[4] S. J. BRAMS AND P. C. FISHBURN, *Approval voting*, Amer. Pol. Sci. Rev., 72 (1978), pp. 831–847.

[5] ———, *Approval Voting*, Birkhaüser, Boston, 1983.

[6] ———, *Proportional representation in variable-size legislatures*, Bell Laboratories, Murray Hill, NJ, 1983.

[7] MARQUIS DE CONCORCET, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris, 1785.

[8] P. ERDÖS AND L. MOSER, *On the representation of directed graphs as unions of orderings*, Publ. Math. Inst. Hung. Acad. Sci., 9 (1964), pp. 125–132.

[9] P. C. FISHBURN, *The theory of representative majority decision*, Econometrica, 39 (1971), pp. 273–284.

[10] ———, *The Theory of Social Choice*, Princeton Univ. Press, Princeton, NJ, 1973.

[11] ———, *Paradoxes of voting*, Amer. Pol. Sci. Rev., 68 (1974), pp. 537–546.

[12] ———, *Social choice functions*, SIAM Rev., 16 (1974), pp. 63–90.

[13] ———, *On the sum-of-ranks winner when losers are removed*, Discrete Math., 8 (1974), pp. 25–30.

[14] ———, *Three-valued representation systems*, Math. Systems Theory, 9 (1975), pp. 265–280.

[15] ———, *Condorcet social choice functions*, SIAM J. Appl. Math., 33 (1977), pp. 469–489.

[16] ———, *Axioms for approval voting: direct proof*, J. Econ. Theory, 19 (1978), pp. 180–185.

[17] ———, *Heights of representative systems*, Discrete Appl. Math., 1 (1979), pp. 181–199.

[18] ———, *Inverted orders for monotone scoring rules*, Discrete Appl. Math., 3 (1981), pp. 27–36.

[19] ———, *Monotonicity paradoxes in the theory of elections*, Discrete Appl. Math., 4 (1982), pp. 119–134.

[20] P. C. FISHBURN AND S. J. BRAMS, *Expected utility and approval voting*, Behav. Sci., 26 (1981), pp. 136–142.

[21] ———, *Paradoxes of preferential voting*, Math. Mag., 56 (1983), pp. 207–214.

[22] P. C. FISHBURN, W. V. GEHRLEIN AND E. MASKIN, *Condorcet proportions and Kelly's conjectures*, Discrete Appl. Math., 1 (1979), pp. 229–252.

[23] W. V. GEHRLEIN, *Condorcet's paradox*, Theory and Decision, 15 (1983), pp. 161–197.

[24] W. V. GEHRLEIN AND P. C. FISHBURN, *The probability of the paradox of voting: a computable solution*, J. Econ. Theory, 13 (1976), pp. 14–25.

[25] ———, *Proportions of profiles with a majority candidate*, Comp. Math. Appl., 5 (1979), pp. 117–124.

[26] W. V. GEHRLEIN, B. GOPINATH, J. C. LAGARIAS AND P. C. FISHBURN, *Optimal pairs of score vectors for positional scoring rules*, Appl. Math. Optim., 8 (1982), pp. 309–324.

[27] A. GIBBARD, *Manipulation of voting schemes: a general result*, Econometrica, 41 (1973), pp. 587–601.

[28] G. T. GUILBAUD, *Les théories de l'intérêt général et la problème logique de l'agrégation*, Écon. Appl., 5 (1952), pp. 501–584.

[29] H. KEIDING, *Heights of simple games* III: *the connection with representative systems*, Institute of Economics, Univ. Copenhagen, 1982.

[30] J. S. KELLY, *Voting anomalies, the number of voters, and the number of alternatives*, Econometrica, 42 (1974), pp. 239–251.

[31] ———, *Arrow Impossibility Theorems*, Academic Press, New York, 1978.

[32] K. O. MAY, *A set of independent necessary and sufficient conditions for simple majority decisions*, Econometrica, 20 (1952), pp. 680–684.

[33] R. M. MAY, *Some mathematical remarks on the paradox of voting*, Behav. Sci., 16 (1971), pp. 143–151.

[34] D. C. MCGARVEY, *A theorem on the construction of voting paradoxes*, Econometrica, 21 (1953), pp. 608–610.

[35] Y. MURAKAMI, *Formal structure of majority decision*, Econometrica, 34 (1966), pp. 709–718.

[36] R. G. NIEMI AND W. H. RIKER, *The choice of voting systems*, Sci. Amer., 234 (1976), pp. 21–27.

[37] R. G. NIEMI AND H. F. WEISBERG, *A mathematical solution to the problem of the paradox of voting*, Behav. Sci., 13 (1968), pp. 317–323.

[38] P. K. PATTANAIK, *Strategy and Group Choice*, North-Holland, Amsterdam, 1978.

[39] D. G. SAARI, *The ultimate of chaos resulting from weighted voting systems*, Discussion paper 505, Center for Mathematical Studies in Economics and Management Science, Northwestern Univ., Evanston, IL, 1981.

[40] M. A. SATTERTHWAITE, *Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions*, J. Econ. Theory, 10 (1975), pp. 187–217.

[41] A. K. SEN, *Collective Choice and Social Welfare*, Holden-Day, San Francisco, 1970.

[42] ———, *Social choice theory: a re-examination*, Econometrica, 45 (1977), pp. 53–89.

[43] D. SLEPIAN, *The one sided barrier problem for Gaussian noise*, Bell Sys. Tech. J., 41 (1962), pp. 463–501.

[44] J. H. SMITH, *Aggregation of preferences with variable electorate*, Econometrica, 41 (1973), pp. 1027–1041.

[45] R. STEARNS, *The voting problem*, Amer. Math. Monthly, 66 (1959), pp. 761–763.

[46] R. J. WEBER, *Comparison of voting systems*, mimeographed, 1978.

[47] H. P. YOUNG, *Social choice scoring functions*, SIAM J. Appl. Math., 28 (1975), pp. 824–838.

[48] H. P. YOUNG AND A. LEVENGLICK, *A consistent extension of Condorcet's election principle*, SIAM J. Appl. Math., 35 (1978), pp. 285–300.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 2, June 1984

014

COMPUTING LOGARITHMS IN FINITE FIELDS OF CHARACTERISTIC TWO*

I. F. BLAKE†, R. FUJI-HARA§, R. C. MULLIN‡ AND S. A. VANSTONE‡

Abstract. A simple algorithm to find logarithms in a finite field of characteristic two is described. It uses the Euclidean algorithm for polynomials in attempting to reduce an element to a product of factors all of whose logarithms are stored in a database. The algorithm, which is similar to one of Adleman, has a random runtime and constant storage requirements. It is analyzed and problems associated with the construction of the database are considered. The aim of the work is to show that the algorithm is feasible for the field with 2^{127} elements on which several proposed public key distribution systems have been based. For such application it is felt that the discrete logarithm is still a viable technique for sufficiently large fields.

AMS subject classifications. 12C05, 12C10, 94A99

1. Introduction. Let α be a primitive element in a finite field with q elements, $GF(q)$. For a positive integer j, $0 \leqq j \leqq q-2$, the determination of α^j is a simple computation requiring on the order of $\log_2(j)$ operations. The inverse problem, where the element $\beta = \alpha^j$ is given and the objective is to determine the exponent j appears to be much harder and is referred to as the discrete logarithm problem to the base α. Two cases of particular interest have arisen in the literature and applications, where the size of the field is either p, a large prime, or of the form 2^n. Only the latter case is of interest in this paper and it will be assumed that elements are represented as binary n-tuples with respect to the basis $\{1, \alpha, \alpha^2, \cdots, \alpha^{n-1}\}$ where α is a root of the primitive polynomial $f(x)$.

The problem of finding logarithms in finite fields finds application in several areas. For example, linear feedback shift registers are widely used in communications and the problem of determining the number of clock cycles between two given states of the shift register is easily shown to be equivalent to finding logarithms in the appropriate finite field. The application of interest here is to a public key distribution system, proposed by Diffie and Hellman [1], which can be described as follows. Two parties A and B use a cryptographic system which requires a common key. A chooses a random integer a and transmits α^a while B chooses a random integer b and transmits α^b. Both parties compute α^{ab} which is then used as the common key for some encipherment system. A tapper on the channel has α^a and α^b available and it appears that it must be able to compute a or b, i.e. find logarithms, in order to compromise the security of the system. While the most popular use of discrete logarithms and exponentiation appears to be in the key distribution problem for encipherment systems, it finds many other uses. It can be used as an encipherment system itself, although it is often regarded as being too slow for such an application. It also finds use in authentication and verification schemes.

Another public key distribution system which has been widely discussed in the literature is the RSA scheme [2] which depends for its cryptographic strength on the difficulty in factoring large integers. Some comments will be given on the utility of the two systems later in the paper.

* Received by the editors July 6, 1983. This work was supported by the Department of Communication and the Department of National Defence, Ottawa. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Electrical Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

§ Institute of Socio-Economic Planning, University of Tsukuba, Sakura, Ibaraki, Japan, 305.

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

From an implementation point of view it is felt there may be advantages in using fields of characteristic two and this will be the case considered in this paper. It has been observed [3] that, for an arbitrary finite field $GF(q)$, when $q-1$ is highly composite and the size of the largest factor is "small", the complexity of finding logarithms is greatly reduced. Some comments on this event will be given later in the paper. To avoid this possibility, our concern will be focussed on those values of n for which 2^n-1 is prime (i.e. a Mersenne prime), and the first few interesting values of n for which this occurs are 31, 61, 89, 107, 127, 521, 607, 1279, 2203 and 2281.

The algorithm for determining discrete logarithms is described in the next section and techniques used for constructing a certain database which the algorithm requires are given in § 3. These techniques have been successfully applied to the cases $n=31$ and 61 which correspond to primes 2^n-1 of the order 2.1×10^9 and 2.3×10^{18} respectively. It is currently being applied to $n=127$ which gives a prime of the order 1.7×10^{38}. Our purpose in this paper is to establish the feasibility of the algorithm for $n=127$ and to examine the problems associated with implementing the algorithm for this value and the properties and characteristics of the implementation. Section 4 considers briefly some associated problems.

2. The algorithm. The algorithm described here is, except for one important step, similar to one proposed by Adleman [4]. It is very simple in concept and most of the paper is concerned with the analysis and implementation details of it. A comparison with the Adleman algorithm will be given later in the section. Unless specified otherwise the field in use will be $GF(2^n)$ where 2^n-1 is prime. Elements of the field will be viewed equivalently as powers of a primitive element α (a root of a primitive polynomial $f(x)$), binary n-tuples with respect to the basis $\{1, \alpha, \alpha^2, \cdots, \alpha^{n-1}\}$ and binary polynomials of degree at most $(n-1)$.

The algorithm makes use of the Euclidean algorithm for polynomials and some preliminary properties of it are first noted. A convenient reference for the material needed is McEliece [5] who presents a continued fraction version of the algorithm.

Let $a(x)$ and $b(x)$ be two binary polynomials and $(a(x), b(x))$ their greatest common divisor. Then there exist two polynomials $s(x)$ and $t(x)$ such that

$$s(x)a(x)+t(x)b(x)=(a(x), b(x)).$$

The polynomials $s(x)$ and $t(x)$ can be determined by the following recursion relationships: for the initial conditions

$$s_{-1}(x)=1, \quad t_{-1}(x)=0, \quad r_{-1}(x)=a(x),$$
$$s_0(x)=0, \quad t_0(x)=1, \quad r_0(x)=b(x),$$

define

$$\begin{aligned} r_i(x)&=r_{i-2}(x)-q_i(x)r_{i-1}(x), \\ s_i(x)&=s_{i-2}(x)-q_i(x)s_{i-1}(x), \qquad i\geqq 1, \\ t_i(x)&=t_{i-2}(x)-q_i(x)t_{i-1}(x). \end{aligned}$$

The following properties are easily established by induction:

$$\begin{aligned} &s_i(x)a(x)+t_i(x)b(x)=r_i(x), && i\geqq -1, \\ &\deg s_i(x)+\deg r_{i-1}(x)=\deg b(x), && i\geqq 1, \\ &\deg t_i(x)+\deg r_{i-1}(x)=\deg a(x), && i\geqq 0, \\ &\deg t_i(x)+\deg r_i(x)<\deg a(x). \end{aligned}$$

The degrees of the remainder polynomials $r_i(x)$ are strictly decreasing and if $r_n(x)$ is the last nonzero one it is $(a(x), b(x))$. From the first relationship above it is observed that

$$t_i(x)b(x) \equiv r_i(x) \bmod (a(x)).$$

Using the above properties and observations the following lemma is readily established.

LEMMA [5]. *Let l and k be nonnegative integers with $k \geqq \deg(a(x), b(x))$, $l+k = \deg(a(x))-1$. Then there exists a unique integer j such that $\deg(t_j(x)) \leqq l$ and $\deg(r_j(x)) \leqq k$.*

This result is applied to elements of $GF(2^n)$, generated by a root of $f(x)$, as follows. Let $g(x)$ be an arbitrary element of $GF(2^n)$, $\deg g(x) > (n-1)/2$, whose logarithm is required. By the lemma there exists two polynomials $r(x)$, $t(x)$, $\deg(r(x)) \leqq (n-1)/2$, $\deg(t(x)) \leqq (n-1)/2$ such that

$$t(x)g(x) \equiv r(x) \bmod f(x),$$

where $f(x)$ and $g(x)$ are identified respectively with $a(x)$ and $b(x)$. It follows that if the logarithms of the lower degree polynomials $r(x)$ and $t(x)$ can be found then

$$\log g(x) = \log r(x) - \log t(x).$$

Since the probability is high that a randomly chosen element of $GF(2^n)$ will be of large degree, this step, which reduces the problem to finding logarithms of two polynomials, each of degree at most $(n-1)/2$, turns out to be crucial in the implementation of the algorithm. At another step of the algorithm the polynomials have to be factored and a further saving is realized by factoring two relatively low degree polynomials rather than one high degree one.

It will be assumed that a database D is available which contains the logarithms of all irreducible polynomials of degree at most b. These correspond to the "smooth" elements in [4]. The choice of the integer b will be discussed later in the section and techniques useful in constructing D will be given in the next section. The aim of the sequel will be to establish the feasibility of constructing the database of $GF(2^{127})$ and the performance of the algorithm given that D is available.

THE ALGORITHM. The logarithm of $g(x)$ is required.

0. Set A to 0.
1. If $\deg(g(x)) \leqq b$ then find $\log g(x)$ in D and go to 4.
2. If $\deg(g(x)) > b$ then apply the Euclidean algorithm to $g(x)$ and $f(x)$ to obtain $t(x)$, $r(x)$, $t(x)g(x) \equiv r(x) \bmod (f(x))$, $\deg t(x)$, $\deg r(x) \leqq (n-1)/2$.
3. Factor $t(x) = \prod_i p_i^{d_i}(x)$, $r(x) = \prod_j p_j^{e_j}(x)$. If $\deg p_i(x) \leqq b$, $\deg p_j(x) \leqq b$, $\forall i, j$ then compute from the database $\log g(x) = \sum_j e_j \log p_j(x) - \sum_i d_i \log p_i(x) - A$ and go to 4. Otherwise generate a random integer a, set $g(x)$ to $x^a g(x)$, A to $A+a$ and go to 1.
4. End.

To begin the analysis of the algorithm we give a recursive technique for computing the probability that a randomly chosen polynomial of degree at most m has all of its irreducible factors of degree at most k. Let $N_e(m, k)$ denote the number of monic polynomials over $GF(q)$ of degree at most m whose largest (highest degree) irreducible factors are of degree exactly k, for $m \geqq 0$, $k \geqq 1$. Similarly let $N_l(m, k)$ denote the number of monic polynomials over $GF(q)$ of degree at most m whose largest irreducible factors have degree at most k, including the zero polynomial. By convention let

$$N_e(m, 0) = N_l(m, 0) = 1, \qquad m > 0,$$

and notice that

$$N_l(m, k) = q^{m+1}, \qquad k \geqq m > 0$$

and

$$N_e(m, k) = 0, \qquad k > m > 0,$$

while $N_e(m, m)$ denotes the number of irreducible polynomials of degree m over $GF(q)$ for $m \geqq 1$ (see [6])

$$N_e(m, m) = \frac{1}{m} \sum_{d/m} \mu(m/d) q^d$$

where $\mu(\cdot)$ is the Möbius inversion function. For $m \geqq k+1$ we have the recursion relation

$$N_l(m, k) = \sum_{r=0}^{k} N_e(m, r) = N_e(m, k) + N_l(m, k-1),$$

and for $m \geqq k+1 \geqq 1$

$$N_e(m, k) = \sum_{r=1}^{\lfloor m/k \rfloor} \binom{N_e(k, k) + r - 1}{r} N_l(m - rk, k-1),$$

and for $m \geqq 1$,

$$N_e(m, m) = N_l(m, m) - N_l(m, m-1).$$

Taken together, we can use these recursion relations to compute $N_l(m, k)$. Dividing this quantity by the total number of polynomials of degree at most m yields the probability that a randomly chosen polynomial of degree at most m has all of its irreducible factors of degree at most k, a quantity we denote by $p(m, k)$.

In $GF(2^n)$ the probability that a randomly chosen binary polynomial of degree at most $n-1$ yields, by the Euclidean algorithm, two polynomials of degree at most $(n-1)/2$, each of which factors into irreducible polynomials of degrees at most k, is approximately $p^2(((n-1)/2), k)$. It is an approximation since it has been assumed the two polynomials resulting from the Euclidean algorithm are independently chosen, which appears to be a reasonable assumption for this analysis. The expected number of iterations of the algorithm to obtain a pair of polynomials, all of whose irreducible factors are of degree at most k, is $(p^2((n-1)/2, k))^{-1}$. To construct the database one might argue as follows. If a polynomial is generated in some random manner but in a way such that its logarithm is known, then with probability $p^2((n-1)/2, k)$ all of the irreducible factors of the two polynomials resulting from applying the Euclidean algorithm to this polynomial will be of degree at most k. If the factors are indeed of degree at most k then an equation results relating the logarithm of a known quantity with those of irreducible polynomials of degree at most k. If it is desired to find the logarithms of all N_b irreducible polynomials of degree at most b, the expected number of equations one would have to examine in order to obtain a sufficient number of equations to permit construction of the database, assuming all such equations are linearly independent, would be $N_b/p^2((n-1)/2, b)$. It is this rough measure of complexity, together with the magnitude of $p^2((n-1)/2, b)$, that will be used to determine the size of the database. The quantities N_b, $p^2((n-1)/2, b)$ and $N_b/p^2((n-1)/2, b)$ are shown in the appropriate columns of Tables 1 and 2 for $n = 61$ and 127 respectively. The algorithm described in this paper is referred to as the new algorithm.

TABLE 1

Probability and expected number of runs for the new and Adleman algorithm, $n = 61$.

		New algorithm		Adleman algorithm	
Degree	Total no. of irred. polys	Probability	Expected no. of runs	Probability	Expected no. of runs
1	1	5.33462163E−14	1.87454719E+13	8.20090524E−16	1.21937758E+15
2	2	1.76871018E−12	1.13076751E+12	8.81933415E−15	2.2677449E+14
3	4	1.42920165E−10	2.79876531E+10	2.4088587E−13	1.66053741E+13
4	7	6.35613994E−09	1.10129734E+09	6.16697535E−12	1.13507832E+12
5	13	3.22381391E−07	40324908.2	2.87534365E−10	4.52119871E+10
6	22	5.33811982E−06	4121301.27	6.49386181E−09	3.38781462E+09
7	40	6.74428582E−05	593094.674	1.45140186E−07	275595623
8	70	4.2333643E−04	165353.121	1.6240426E−06	43102317.6
9	126	1.86019133E−03	67734.9679	1.2624515E−05	9980581.43
10	225	5.77168417E−03	38983.4221	6.39169698E−05	3520191.91
11	411	.0144081649	28525.4925	2.48450225E−04	1654254.89
12	746	.0293335145	25431.6611	7.38423665E−04	1010260.15
13	1376	.0528409512	26040.4094	1.84778808E−03	744674.141
14	2537	.085727823	29593.6595	3.94471792E−03	643138.509
15	4719	.128359867	36763.8274	7.51093409E−03	628284.038
16	8799	.176843304	49755.9127	.0129724298	678284.649
17	16509	.229773022	71849.166	.0207338898	796232.649
18	31041	.285984158	108540.977	.0310329059	1000260.82
19	58635	.345031001	169941.251	.0441310341	1328656.83
20	111012	.40617628	273309.904	.0601228451	1846419.61
21	210870	.469073093	449546.143	.0789028551	2672526.87
22	401427	.533239803	752807.645	.100132689	4008950.56
23	766149	.598365862	1280402.26	.12355915	6200665.83
24	1465019	.664015691	2206301.78	.148935705	9836586.87
25	2807195	.729737155	3846857.71	.176080582	15942672.2
26	5387990	.794775116	6779263.58	.204816723	26306396.9
27	10358998	.857890338	12074967.6	.235001922	44080482
28	19945393	.916752227	21756579.8	.266478225	74848115.6
29	38458183	.966945476	39772855.8	.299022554	128612984
30	74248450	.999999999	74248450.1	.332077077	223588002

Choosing the database by the criterion discussed above, then for $n = 61$ the 746 logarithms of all irreducible polynomials of degree at most 12 should be stored. Of course, storing more logarithms will improve the efficiency of the algorithm but will require more effort to construct the database. Independent equations relating the logarithms of the 225 irreducible polynomials of degree at most 10 were found, by techniques described in the next section, and solved. The algorithm itself was then run with $b = 10$ to build the database up to $b = 12$, i.e. to find the logarithms of the irreducible polynomials of degrees 11 and 12. Five hundred randomly chosen polynomials of degree at most 60 were run with the algorithm with this database and the distribution of the number of polynomials requiring a given number of iterations is shown in Table 3. The average number of iterations of the algorithm was 56.0 with an average runtime on a VAX 11/780 of under 5 seconds to determine a logarithm.

For $n = 127$ the optimal value of b, for the database, would be 20. This would require the determination and storing of 111,012 logarithms. It was felt that $b = 17$, requiring 16,509 logarithms would not significantly degrade performance while making the task considerably simpler. As with $n = 61$, the initial approach taken to construct this database was to find the 746 logarithms of polynomials of degree at most 12.

TABLE 2

Probability and expected number of runs for the new and Adleman algorithm, $n = 127$.

		New algorithm		Adleman algorithm	
Degree	Total no. of irred. polys	Probability	Expected no. of runs	Probability	Expected no. of runs
1	1	1.27141469E−32	7.8652544E+31	0.47772090E−34	0.20932720E+35
2	2	1.61023467E−30	1.24205499E+30	0.10391370E−32	0.19246730E+34
3	4	1.42003032E−27	2.81684126E+27	0.10686600E−30	0.37430020E+32
4	7	1.15313844E−24	6.07038995E+24	0.16022050E−28	0.43689770E+30
5	13	3.46321796E−21	3.75373429E+21	0.12806100E−25	0.10151400E+28
6	22	2.43083741E−18	9.05037906E+18	0.60489980E−23	0.36369650E+25
7	40	1.75009226E−15	2.28559379E+16	0.58661830E−20	0.68187380E+22
8	70	2.99279106E−13	2.33895379E+14	0.20402690E−17	0.34309190E+20
9	126	2.39477444E−11	5.26145586E+12	0.40463370E−15	0.31139250E+18
10	225	7.73424039E−10	2.90914154E+11	0.31781350E−13	0.70796220E+16
11	411	1.42517804E−08	2.88385022E+10	0.13612270E−11	0.30193290E+15
12	746	1.48157461E−07	5.03518348E+09	0.29555160E−10	0.25240930E+14
13	1376	1.0660927E−06	1.29069451E+09	0.41129290E−09	0.33455440E+13
14	2537	5.46899677E−06	463887639	0.37461550E−08	0.67722710E+12
15	4719	2.19028276E−05	215451634	0.24991280E−07	0.18882570E+12
16	8799	7.12191038E−05	123548311	0.12723630E−06	0.69154690E+11
17	16509	1.96662481E−04	83945854.4	0.52428530E−06	0.31488570E+11
18	31041	4.71015488E−04	65902291.5	0.17992970E−05	0.17251720E+11
19	58635	1.00764675E−03	58190035.4	0.53284450E−05	0.11004140E+11
20	111012	1.96260912E−03	56563479.1	0.13895770E−04	0.79888990E+10
21	210870	3.54177251E−03	59537985.4	0.32587350E−04	0.64709140E+10
22	401427	5.96621694E−03	67283339.5	0.69740460E−04	0.57560100E+10
23	766149	9.45338803E−03	81044911.9	0.13806940E−03	0.55490060E+10
24	1465019	.0142135998	103071637	2.25535770E−03	0.57373450E+10
25	2807195	.0204564492	137227872	0.44523690E−03	0.63048900E+10
26	5387990	.0283794933	189855046	0.73751900E−03	0.73054240E+10
27	10358998	.0381757153	271350462	0.11680410E−02	0.88686300E+10
28	19945393	.0500248252	398709899	0.17773340E−02	0.11222780E+11
29	38458183	.0640949758	600018684	0.26095230E−02	0.14737250E+11
30	74248450	.0805164802	922152208	0.33185230E−01	—

TABLE 3

Number of iterations of the algorithm for 500 randomly chosen polynomials, $n = 61$.

Number of trials	1–20	21–40	41–60	61–80	81–100	101–120	121–140	141–160	161–180	181–200	>200
Number of polynomials	140	109	81	56	48	22	15	9	7	4	9

Although the probabilities were such that our main algorithm would not contribute much to this database a combination of various heuristic techniques, including those discussed in § 3, were used to create this database, and using a variety of techniques this base was extended to cover all polynomials of degree up to and including 17.

To compare this algorithm with the Adleman algorithm [4], a brief description of that algorithm is given, modified appropriately for fields of characteristic two, the case of interest here. Let $p_1(x), p_2(x), \cdots, p_{N_b}(x)$ be all the irreducible polynomials of degree at most b over $GF(2)$. Define a polynomial $a(x)$ to be smooth if

$$a(x)=\prod_{i=1}^{N_b} p_i^{e_i}(x)$$

and with each such polynomial associate a vector $\overline{a(x)}=(e_1, e_2, \cdots, e_{N_b})$. Let $g(x)$ be an arbitrary nonzero element of $GF(2^n)$.

Determine positive integers $s, s_1, \cdots, s_i$ such that $g(x)\alpha^s, \alpha^{s_1}, \cdots, \alpha^{s_i}$ are smooth elements and $\overline{g(x)\alpha^s}$ is linearly dependent on $\overline{\alpha^{s_1}}, \overline{\alpha^{s_2}}, \cdots, \overline{\alpha^{s_l}}$. Thus

$$\overline{g(x)\alpha^s}=\sum_{i=1}^{l} \lambda_i \overline{\alpha^{s_i}}$$

for some positive integers λ_i and

$$\log_\alpha (g(x)) = -s + \sum_{i=1}^{l} \lambda_i s_i.$$

The Adleman algorithm, as with the one presented in this paper, requires factorization of polynomials and solutions of linear equations over $GF(2^n-1)$. For a systematic attack on an encipherment it is likely better to compute the logarithms of all irreducible polynomials of degree at most b, as advocated in our algorithm, rather than attempt to find an appropriate set of equations to solve for each element. This is an operational difference between the two algorithms rather than a philosophical one.

For the Adleman algorithm, a randomly chosen element of $GF(2^n)$ will be smooth with probability $p(n-1, k)$ and the expected number of iterations of the algorithm is $p(n-1, k)^{-1}$. Thus the difference in the two algorithms means for the Adleman algorithm a polynomial of degree at most $n-1$ will have to be factored for each iteration while for the new algorithm each iteration requires a use of the Euclidean algorithm and the factorization of two polynomials, each of degree at most $(n-1)/2$.

For $GF(2^{61})$ the optimal size of the database for the Adleman algorithm in terms of minimizing the number of random equations which must be considered, is found in Table 1 and given by $b=15$, requiring 4719 logarithms to be stored with an expected number of iterations of 628,284. This is compared to $b=12$, 746 logarithms and 25,431 iterations for the new algorithm. The equivalent information in Table 2 for $GF(2^{127})$ shows an optimal value of $b=23$, 766,149 logarithms and $5{,}549{,}006\times 10^3$ expected iterations for the Adleman algorithm and $b=20$, 111,012 logarithms and 56,563,479 expected iterations for the new one. Thus in terms of numbers of iterations alone the new algorithm is approximately 100 times as efficient as the Adleman algorithm for $n=127$ and requires less storage. Since factoring one high degree polynomial is less efficient than factoring two polynomials of much smaller degree, it appears that the factor of 100 is conservative although some caution is required in interpreting this statement.

A brief comment should be made on the factoring algorithm used. It is desired to find the largest irreducible factor of the given polynomial first, since if its degree exceeds that chosen for the database that iteration of the algorithm is terminated. For an arbitrary polynomial $g(x)$ define Sim $(g(x))$ as the product of all the irreducible factors of $g(x)$ which occur to the first power. The repeated part of $g(x)$, Rep $(g(x))$ is then $g(x)/\mathrm{Sim}\,(g(x))$. It is a relatively simple matter to isolate Sim $(g(x))$. The Berlekamp algorithm for polynomial factorization over $GF(2)$ [6] is used and requires Gaussian elimination on a $d\times d$ matrix where d is the degree of $g(x)$. By factoring Sim $(g(x))$ first, and using a modification of Berlekamp's algorithm to extract irreducible factors of largest degree first, an "early abort" strategy was developed.

3. Construction of the database. The algorithm described in the previous section depends critically in the ability to find the logarithms of all irreducible polynomials of degree less than or equal to some degree b. It appears unlikely at this point that an algorithm that does not require the construction of some data base will be found and that for useful field sizes this database will have to be substantial. There is no systematic construction of the database of interest here, but several techniques which have proven useful and successful for the fields $GF(2^{31})$ and $GF(2^{61})$ are described.

The underlying idea is to attempt to establish relationships between an element with a known logarithm and its factors if all of its factors have degree at most b. Such a relationship gives a linear equation over the integers modulo 2^n-1 and the problem is to establish a sufficient number of such equations to permit solution.

For an arbitrary nonzero element $\beta \in GF(2^n)$ define the square orbit of β as

$$SO(\beta) = \{\beta^{2^i}, 0 \leqq i \leqq n-1\}.$$

If the logarithm of any element in $SO(\beta)$ is known, the logarithms of all elements in the set are easily found. In particular, if α is a primitive element and $SO(\beta)$ contains polynomials of low degree we say this field representation exhibits "orbital weakness" which can be exploited to obtain many equations.

As an example consider α a primitive element of $GF(2^{127})$ which satisfies the primitive trinomial $f(x) = x^{127} + x + 1$. For $i \geqq 7$

$$\alpha^{2^i} = (\alpha^{2^7})^{2^{i-7}} = (\alpha + \alpha^2)^{2^{i-7}} = \alpha^{2^{i-6}} + \alpha^{2^{i-7}}$$

and by repeated application it is concluded that every element of the form α^{2^i}, $0 \leqq i \leqq 127$ can be expressed as a linear sum of elements in $SO(\alpha)$. Furthermore, since

$$(1+\alpha^{2^i}) = (1+\alpha)^{2^i} = \alpha^{127.2^i}$$

the logarithm of every element in the span of the set $\{1, \alpha, \alpha^2, \alpha^{2^2}, \cdots \alpha^{2^6}\}$ is known and this fact will yield many equations.

The following theorem has also proven useful in assisting with obtaining equations for the construction of the database.

THEOREM. *Let $f(x)$ be an irreducible polynomial of degree n over $F = GF(q)$ and let $g(x)$ be an arbitrary polynomial over $GF(q)$. If $m(x)$ is any divisor of $f(g(x))$ then the degree of $m(x)$ is a multiple of n.*

Proof. It is sufficient to prove the result for irreducible factors $m(x)$. Let $m(x)$ be a divisor of $h(x) = f(g(x))$, K the splitting field of $h(x)$, α a root of $m(x)$ in K and L the splitting field of $m(x)$. Since $m(x) | h(x)$ we have $h(\alpha) = f(g(\alpha))$ so that $g(\alpha)$ lies in the splitting field of $f(x)$, which is $R = GF(q^n)$. Also $g(\alpha)$ does not lie in any proper subfield of $GF(q^n)$ since all zeros of $f(x)$ generate R. Since $g(\alpha)$ lies in L then $R \subset L$ and therefore $n | \dim [L\colon GF(q)]$. Since $m(x)$ is irreducible and L is its splitting field we have that

$$\deg m(x) = \dim [L\colon GF(q)]$$

and the result follows.

To illustrate the use of this theorem and the notion of orbital weakness in finding equations consider $GF(2^{127})$ generated by the primitive trinomial $f(x) = 1 + x + x^{127}$. The polynomial $g(x) = 1 + x^2 + x^5$ is irreducible and, since $x^{2^7} = x + x^2$,

$$g(x^{2^7}) = g(x+x^2) = 1 + x^2 + x^4 + x^5 + x^6 + x^9 + x^{10}$$

$$= (1 + x^2 + x^3 + x^4 + x^5)(1 + x^3 + x^5) = p_1(x)p_2(x).$$

Now $g(x^{2^7}) = g(x)^{2^7}$ and so

$$2^7 \log g(x) = \log p_1(x) + \log p_2(x),$$

which gives a linear equation between the logarithms of three quintics.

Equations generated as in this example are referred to as systematic. The system of linear equations required to construct the database for $GF(2^{127})$ cannot be made up entirely of systematic equations. It seems likely however that a large portion of them can be systematic and the remaining ones must be found using other methods. A start was made, for example, in constructing the data base of $GF(2^{127})$. There are 226 irreducible polynomials of degree at most 10 and 142 linearly independent systematic equations were developed for them. There are 126 irreducible polynomials of degree at most 9 and 90 linearly independent systematic equations were found for them. As mentioned previously, to increase this initial database to include all polynomials of degree up to 17, a variety of techniques were used. With the probabilities involved, it was not feasible to use the algorithm itself to build up the database. Techniques similar to those introduced earlier in the section were applied very effectively here.

In generating new equations it is noted the equations are over $GF(2^n - 1)$ with m variables where m is the size of the database. If t independent equations are available, the probability that a new random equation will be linearly independent is

$$\frac{(2^n-1)^m - (2^n-1)^t}{(2^n-1)^m},$$

which has a minimum nonzero value of

$$\frac{2^n-2}{2^n-1}$$

when $t = m - 1$. Thus as each new equation is found it has a high probability of being independent of previous equations.

A general conclusion of the experience gained in constructing the databases is that the problem does not appear to be hard although a systematic and deterministic approach is not yet available.

4. Other issues. Several questions arose during the course of the work which have not been well investigated. They are briefly mentioned here.

Much of the available literature on the discrete logarithm implementation of a public key distribution system assumes the order of the multiplicative group of the field, $q-1$, is a Mersenne prime. It has been shown [3] that when q is prime and $q-1$ has a prime factorization with all prime factors small, then it is a fairly simple matter to determine logarithms. A similar technique is also applicable to fields of interest here. The question remains however if there is any technique to weaken the security of the system if $2^n - 1$ has at least one relatively small factor and a very large factor. Intuitively one might argue the security of the system is determined by the size of the largest factor, but there has been no serious investigation to determine whether the existence of small factors compromises it or not. The problem is of some interest because of the sparseness of Mersenne primes.

Of the finite fields considered in our study $GF(2^{31})$ and $GF(2^{127})$ were generated by trinomials. This turned out to be quite convenient for our purposes since much of the experimentation was done by hand. For implementation this would seem to be relatively unimportant in terms of both speed and complexity. The question arises as to whether the choice of a trinomial field generator polynomial has any implications for the security of the system. Said another way, is it significantly easier to determine logarithms in a field generated by a trinomial than it is in one generated by any other polynomial? Our feeling is that the effect of choosing a trinomial will have a negligible effect on the security of the system, but the question has not been seriously investigated.

5. Comments. An algorithm for the determination of logarithms in a finite field has been discussed. The interesting aspects of the work pertain to the analysis and implementation of it. For $GF(2^{127})$ it was shown that the construction of the database, while nontrivial, is a feasible task. It is also observed that the probability that a randomly chosen polynomial in the field has all of its factors in the database is sufficiently (and surprisingly) high to imply a relatively modest expected number of iterations to complete the algorithm. Taken together they demonstrate the feasibility of the algorithm for this field, which was the major aim of the work.

The RSA algorithm for a public key distribution system [2] depends for its security on the difficulty of factoring large integers. Because of the inherently binary character of the logarithm problem for fields of characteristic two it is felt there may be implementation advantages for the discrete logarithm problem over the RSA, in terms of speed and complexity of hardware required for the same level of security. This aspect has not been sufficiently well investigated to state with any confidence. It is noted however that the 127 bits used in $GF(2^{127})$ corresponds approximately to 38 digits, for which integer factorization algorithms can be quite easily implemented.

Ultimately, confidence in a security system can only be established by its continued analysis over a long period of time by highly qualified people working with sufficient resources. This has certainly been true for the RSA system where a successful attack would imply the existence of an integer factorization algorithm far more efficient than is presently felt possible. The discrete logarithm problem has not yet withstood such a test of time.

Acknowledgments. The authors would like to thank Dr. D. Coppersmith of the IBM Research Center, Yorktown Heights and Dr. A. Odlyzko of Bell Laboratories, Murray Hill, for many interesting conversations on this problem. They would also like to thank Dr. Evi Nemeth of the University of Colorado for her considerable assistance with the construction of the database for $GF(2^{127})$.

REFERENCES

[1] W. DIFFIE AND M. E. HELLMAN, *New directions in cryptography*, IEEE Trans. Inform Theory, IT-22 (1976), pp. 644–654.

[2] R. L. RIVEST, A. SHAMIR AND L. ADLEMAN, *A method for obtaining digital signatures and public key cryptosystems*, Comm. ACM, 21 (1978), pp. 120–126.

[3] S. C. POHLIG AND M. E. HELLMAN, *An improved algorithm for computing logarithms over GF(p) and its cryptographic significance*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 106–110.

[4] L. ADLEMAN, *A subexponential algorithm for the discrete logarithm problem with applications to cryptography*, Proc. IEEE 20th Annual Symposium on Foundations of Computer Science (1979), pp. 55–60.

[5] R. J. MCELIECE, *The Theory of Information and Coding: A Mathematical Framework for Communications*, Addison-Wesley Publishing, Reading, MA 1977.

[6] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

EXPLICIT CONCENTRATORS FROM GENERALIZED N-GONS*

R. MICHAEL TANNER†

Abstract. Concentrators are graphs used in the construction of switching networks that exhibit high connectivity. A technique for establishing the concentration properties of a graph by analysis of its eigenvalues is given. If the ratio of the subdominant eigenvalue to the dominant eigenvalue is small, the graph is a good concentrator. Generalized N-gons are very sparse, locally tree-like graphs for which the eigenvalues can be calculated with relative ease. The eigenvalue ratios for the known N-gons are calculated and show that the N-gons are excellent concentrators.

1. Introduction. A variety of problems concerned with information transmission and complexity have shown the importance of constructing graphs that are highly connected yet sparse. A survey article by Pippenger [12] discusses the complexity of constructing switching networks. The existence of nonblocking networks requiring only $O(N \log N)$ switches was first proven by Bassalygo and Pinsker [1] using a recursive construction based on sparse graphs called concentrators. Concentrators are also a key building block for the construction of a class of graphs called superconcentrators [15][13] useful in studies of algorithmic complexity [11][8], and they are the basis for another class of graphs called generalized connectors [13]. In a seemingly unrelated area, Tanner [14] has shown the importance of concentrator-like bipartite graphs for the construction of low complexity error-correcting codes.

For the Bassalygo–Pinsker network construction, a sequence of graphs is required that can be concatenated, connecting the outputs of one graph to the inputs of the next, to form the entire switching network. Using a counting argument, Bassalygo and Pinsker proved the existence of the necessary concentrators, but did not give an explicit construction. Margulis [10] gave the first explicit construction for concentrators, but his result was deficient in that he could not give an exact value for the expansion factor. In the language of Gabber and Galil [6], an (n, k, d) *expander* is a bipartite graph with n inputs, n outputs and at most kn edges, such that for subset X of inputs the subset Γ_x of outputs satisfies $|\Gamma_X| \geqq [1+d(1-|X|/n)]|X|$, where Γ_X is the set of outputs connected to X and $|\cdot|$ is the cardinality of the set. Margulis exhibited $(n, 5, d)$ expanders for $n=m^2$, but could only show that $d>0$. Gabber and Galil gave an explicit construction for a family of $(n, 5, d_0)$ expanders with $d_0=(2-\sqrt{3})/4$, as well as a family of $(n, 7, 2d_0)$ expanders.

In this paper, we provide first a simple lower bound on the concentration level of an arbitrary bipartite graph. In Gabber and Galil's terminology, an $(n, \theta, k, \alpha, c)$ *bounded strong concentrator* (*bsc*) is a bipartite graph with n inputs, θn outputs, and at most kn edges, such that if X is a set of inputs with $|X| \leqq \alpha n$, then $|\Gamma_X| \geqq c|X|$. We bound c as a function of α in terms of the eigenvalues of the graph. In brief, if the ratio of the subdominant eigenvalue to the dominant eigenvalue is small, $c(\alpha)$ is large. Second, we show that a class of graphs derived from finite geometries, called generalized N-gons, have an eigenvalue ratio that gives a relatively large c. These graphs are distance regular graphs that form metric association schemes, which permits the eigenvalues of the adjacency matrix to be determined quite easily. They have the remarkable property that the girth is exactly twice the diameter, and thus the graph is locally tree-like. Using the N-gons as expanders, we show that for those values of

* Received by the editors November 18, 1980, and in revised form April 15, 1983.

† Department of Computer and Information Sciences, University of California, Santa Cruz, California 95064.

n and k for which N-gons exist, they are (n, k, d) expanders with much larger values of d than those of the Gabber and Galil construction.

Unfortunately, the generalized N-gons do not provide a complete solution to the problem of constructing concentrators for two reasons: First, there do not exist generalized N-gons for arbitrary parameters n and k. We discuss techniques for pruning a generalized N-gon to obtain a bsc with a smaller number of input and output nodes. However, the pruned graphs are no longer distance regular, and the concentration factor c must be bounded by modification of the bound for the original N-gon. The modified bound weakens as the pruning becomes more severe. Second, as shown by Feit and Higman [5], generalized N-gons do not exist for arbitrarily large N. While the N-gons can be used to build very large bsc's, they do not provide arbitrarily large bounded-degree bsc's.

2. An eigenvalue argument. Consider a bipartite graph G with a set of input nodes A of size n, a set of output nodes B of size m, edges connecting input nodes to output nodes. In keeping with the notation of finite geometries, let the degree of each input node be $s+1$ and that of each output node be $r+1=n(s+1)/m$. Let M be the real valued incidence matrix of the bipartite graph: $M=[m_{ij}]$, $m_{ij}=1$ if the ith input node a_i is connected to the jth output node b_j and 0 otherwise. Since MM^T is a real symmetric nonnegative definite matrix, it is diagonalizable and has real nonnegative eigenvalues and orthogonal eigenvectors. Let $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n$ be the ordered eigenvalues, and $e_1, e_2, \cdots, e_n$ the corresponding orthonormal eigenvectors.

THEOREM 2.1. *If $\lambda_1 > \lambda_2$, then G is an $(n, m/n, s+1, \alpha, c(\alpha))$ bsc with*

$$c(\alpha) \geqq \frac{(s+1)^2}{[\alpha((s+1)(r+1)-\lambda_2)+\lambda_2]}.$$

Proof. Let row vector A be the characteristic vector of some set X_A of size $|X_A| = \alpha n$. That is, $A=[A_i]$ with $A_i = 1$ if the ith input is in X_A, 0 otherwise. Thus $AA^T = \|A\|^2 = \alpha n$. Let Y_i be the set of output nodes connected to the ith input node, X_i. Let $Y_B = \cup\{Y_i: x_i \in X_A\}$, and let B be its characteristic vector. Note that if $Y_i \cap Y_j = \varnothing$ for all $x_i, x_j \in X_A$, $i \neq j$, then

$$AMM^TA^T = \alpha n(s+1) = AM[1, 1, \cdots, 1]^T,$$

$AM = B$, and $|Y_B| = \|B\|^2 = \alpha n(s+1)$. Any overlap in the Y_i will cause AMM^TA^T to increase, and our strategy is to lower bound $|Y_B|$ by upper bounding AMM^TA^T. Let $C = AM$. By convexity of $f(x) = x^2$,

$$(2.1) \qquad \|C\|^2 = \sum_{j=1}^{m} C_j^2 \geqq \left(\sum_{j=1}^{m} C_j \Big/ |Y_B| \right)^2 |Y_B| = (\alpha n(s+1))^2/|Y_B|.$$

Thus $|Y_B| \geqq (\alpha n(s+1))^2/\|C\|^2$. Expanding A in terms of the eigenvectors, let $A = \sum_{i=1}^{n} \gamma_i e_i$. Then $AMM^T = \sum_{i=1}^{n} \lambda_i \gamma_i e_i$ and $\|C\|^2 = \sum_{i=1}^{n} \lambda_i \gamma_i^2$ by orthonormality of the eigenvectors. Since all input nodes have degree $s+1$ and all output nodes have degree $r+1$, it is easy to prove that $\lambda_1 = (s+1)(r+1)$ and, without loss of generality, we may take $e_1 = [1, 1, \cdots, 1]/\sqrt{n}$. (Clearly this e_1 is an eigenvector with eigenvalue $(s+1)(r+1)$. Any other $\lambda_i \leqq (s+1)(r+1)$, as can be seen by considering the equations satisfied by the eigenvector component with largest absolute value [2, p. 14].) This implies immediately that $\gamma_1 = Ae_1^T = \alpha\sqrt{n}$. Thus

$$\|C\|^2 = \lambda_1\gamma_1^2 + \sum_{i=2}^{n} \lambda_i\gamma_i^2 = \lambda_1\alpha^2 n + \sum_{i=2}^{n} \lambda_i\gamma_i^2$$

$$\leqq \lambda_1\alpha^2 n + \lambda_2 \sum_{i=2}^{n} \gamma_i^2 = \alpha^2 n(\lambda_1 - \lambda_2) + \lambda_2\|A\|^2 \tag{2.2}$$

$$\leqq \alpha^2 n((s+1)(r+1) - \lambda_2) + \lambda_2\alpha n.$$

Therefore, the concentration $c(\alpha)$ satisfies

$$c(\alpha) = |Y_B|/|X_A| \geqq \alpha^2 n^2 \frac{(s+1)^2}{\alpha^2 n^2[\alpha((s+1)(r+1) - \lambda_2) + \lambda_2]} \tag{2.3}$$

$$\geqq \frac{(s+1)^2}{[\alpha((s+1)(r+1) - \lambda_2) + \lambda_2]}.$$

3. Generalized polygons. The generalized polygons [4] are incidence structures consisting of points and lines; for our purposes we can restrict our attention to those in which every point is incident on $s+1$ lines and every line is incident on $r+1$ points, for some positive integers s and r. By identifying points with input nodes and lines with output nodes, a generalized N-gon defines a bipartite graph G that satisfies the following conditions:

(1) For all nodes $u, v \in G$, $d(u, v) \leqq N$, where $d(u, v)$ is the length of the minimum path connecting u and v.

(2) If $d(u, v) = h < N$, then there is a unique path of length h joining u and v.

(3) Given a node $u \in G$ there exists a node $v \in G$ such that $d(u, v) = N$.

When $s = r = 1$, the graph is the standard N-gon. Note that (2) implies that every cycle of even length has length at least $2N$; since every cycle in a bipartite graph must be even, the girth of the N-gon is at least $2N$.

These conditions ensure that the graph forms a metric association scheme [9]. Moreover, the set of input nodes alone form an association scheme with nodes u and v in the ith relation whenever $d(u, v) = 2i$. Consequently, the eigenvalues of MM^T can be determined for a generalized N-gon by finding the eigenvalues of an associated $(\lfloor N/2 \rfloor + 1) \times (\lfloor N/2 \rfloor + 1)$ matrix. Alternatively, and equivalently, $MM^T = (s+1)I + D$, where $(D)_{uv} = 1$ if $d(u, v) = 2$ and 0 otherwise. The matrix D can be used as the adjacency of a distance regular graph with the input nodes as its node set. As shown in [2, pp.140–142], the distinct eigenvalues of D are the same as those of an intersection matrix P whose (i, j)th entry, p_{ij}, $0 \leqq i, j \leqq \lfloor N/2 \rfloor$ is the number of nodes w that satisfy $d(u, w) = 2$ and $d(u, v) = 2i$, for any pair of u, v, with $d(u, v) = 2j$.

The meaning of the intersection matrix is clarified by considering the form of a possible eigenvector for the graph defined by D. Starting from an arbitrary node in the graph, let E be a vector for which all components corresponding to nodes at distance i in the graph have the same value, x_i, $i = 0, \cdots, \lfloor N/2 \rfloor$. Let X be the $(\lfloor N/2 \rfloor + 1)$-dimensional vector with components x_i. The matrix P defines the set of equations that must be satisfied by X in order to make E an eigenvector of D; that is, X is an eigenvector of P implies E is an eigenvector of D. The matrix entry p_{ij} is the number of edges connecting a node with value x_i to any other node with value x_j.

The P matrices of the generalized N-gons for $N = 3$, 4, 6, and 8 are given in Fig. 1, along with the corresponding eigenvalues of P and the N-gon. Of course, one can readily verify a given eigenvalue λ by checking that the matrix $P - \lambda I$ is singular. The

3-gons

$$\begin{bmatrix} 0 & 1 \\ s(s+1) & s(s+1)-1 \end{bmatrix}$$

P eigenvalues: $s(s+1)$, -1
3-gon eigenvalues: $(s+1)(s+1)$, s

4-gons

$$\begin{bmatrix} 0 & 1 & 0 \\ (s+1)r & r-1 & s+1 \\ 0 & sr & (s+1)(r-1) \end{bmatrix}$$

P eigenvalues: $r(s+1)$, $r-1$, $-(s+1)$
4-gon eigenvalues: $(s+1)(r+1)$, $s+r$, 0

6-gons

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ (s+1)r & r-1 & 1 & 0 \\ 0 & sr & r-1 & s+1 \\ 0 & 0 & sr & (s+1)(r-1) \end{bmatrix}$$

P eigenvalues: $(s+1)r$, $r-1+\sqrt{rs}$, $r-1-\sqrt{rs}$, 0
6-gon eigenvalues: $(s+1)(r+1)$, $r+s+\sqrt{rs}$, $r+s-\sqrt{rs}$, 0

8-gons

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ (s+1)r & r-1 & 1 & 0 & 0 \\ 0 & sr & r-1 & 1 & 0 \\ 0 & 0 & sr & r-1 & s+1 \\ 0 & 0 & 0 & sr & (s+1)(r-1) \end{bmatrix}$$

P eigenvalues: $(s+1)r$, $r-1+\sqrt{2rs}$, $r-1$, $r-1-\sqrt{2rs}$, 0
8-gon eigenvalues: $(s+1)(r+1)$, $r+s+\sqrt{2rs}$, $r+s$, $r+s-\sqrt{2rs}$, 0

FIG. 1. *P matrices and eigenvalues of the N-gons.*

eigenvalues can be deduced from the roots of the characteristic equations provided by [5] (e.g., [5, p.122], the value given for θ).

The concentration $c(\alpha)$ derived in § 2 can be rewritten

$$c(\alpha) \geqq \frac{(s+1)}{(r+1)(\alpha+\rho(1-\alpha))} = \frac{m}{n(\alpha+\rho(1-\alpha))}$$

with $\rho = \lambda_2/(s+1)(r+1)$. Clearly, for all the generalized N-gons, ρ can be made as small as desired by choosing s and r sufficiently large, thus permitting $c(\alpha)$ to be as large as $m/(n\alpha)$.

In Table 1 we list the parameters of all currently known generalized N-gons [7] with $s>1$ and $t>1$.

As an example, take $s=r=7$ in a 6-gon. The graph has $(7^6-1)/(7-1)$ input nodes and the same number of output nodes. The eigenvalues ratio is $\rho = 21/64$, and the concentration satisfies $c(\alpha) \geqq 64/(43\alpha+21)$ for $0 \leqq \alpha \leqq 1$.

TABLE 1
Known generalized polygons.

N	s	r	n
3	q	q	s^2+s+1
4	q q q^2 $q-1$ $(q\neq 2)$	q q^2 q^3 $q+1$	$sr^2+sr+r+1$
6	q q	q q^3	$(r+1)(s^2r^2+sr+1)$
8	2^m	2^{2m}	$s^{15}+(s^{12}-1)\left(1+\frac{s+1}{s^4-1}\right)$
	(q a prime power)		

4. Comparison with the Gabber–Galil construction. For those parameters n, m, and $k=s+1$ for which N-gons exist, they are excellent concentrators, as we will now show by comparing the expanders derived from an N-gon with those of the more general construction of Gabber and Galil.

An $(n, 1, k, \alpha, c=1+d(1-\alpha))$ bsc is an (n, k, d) expander. An N-gon with $n=m$ is therefore an expander with parameter d satisfying

$$d=[1/(1-\alpha)]\left[\frac{1}{(\alpha+\rho(1-\alpha))}-1\right]=\frac{(1-\rho)}{\alpha+\rho(1-\alpha)}.$$

Here d is a function of α. Since $\rho<1$, a lower bound for d over the entire range of α is obtained by letting $\alpha=1$, which gives $d\geqq(1-\rho)/(1+\rho)$.

Theorems 2 and 2′ of Gabber and Galil furnish $(n, 5, (2-\sqrt{3})/4)$ and $(n, 7, (2-\sqrt{3})/2)$ expanders, respectively. The example $s=r=7$ hexagon above is a (19,608, 8, 43/85 = 0.506) expander. The most closely comparable expander of Gabber and Galil is a (19,600, 7, 0.134) expander. Taking $s=r=4$ in a hexagon produces a (1,365, 5, 13/37 = 0.351) expander, compared with a (1,369, 5, 0.067) expander of the Gabber and Galil construction.

5. Pruning. Although generalized N-gons do not exist for all possible parameters n, s, and r, it is possible to obtain good bsc's by pruning, eliminating nodes and edges from an N-gon. The first observation is that given an $(n, m/n, k, \alpha, c(\alpha))$ bsc, an $(n', m/n', k, \alpha', c((n'/n)\alpha'))$ bsc can be formed merely by eliminating some arbitrary set of $n-n'$ input nodes. Second, if n' is much smaller than n it will be possible to eliminate some of the output nodes also, those incident on none of the n' remaining input nodes, without decreasing the concentration factor c. Third, suppose it is possible to prune a bipartite graph G of Theorem 2.1 to create a graph G' in which all input nodes have degree at least $t+1$ and at most $s'+1$. Then we have the following modification of Theorem 2.1:

THEOREM 5.1. *G' is an $(n', m'/n', s'+1, a', c(\alpha'))$ bsc with*

$$c(\alpha')\geqq\frac{(t+1)^2}{[\alpha'(n'/n)((s+1)(r+1)-\lambda_2)+\lambda_2]},$$

where λ_1 and λ_2 are the eigenvalues of the original graph G.

Proof. Equation (2.1) can be replaced by

$$\|C\|^2=\sum_{j=1}^{m} C_j^2 \geqq \left(\sum_{j=1}^{m} C_j \Big/ |Y_B|\right)^2 |Y_B| \geqq (\alpha' n'(t+1))^2/|Y_B|. \tag{4.1}$$

Equation (2.2) becomes

$$\|C\|^2=\lambda_1\gamma_1^2+\sum_{i=2}^{n}\lambda_i\gamma_i^2=\lambda_1\alpha'^2(n'^2/n)+\sum_{i=2}^{n}\lambda_i\gamma_i^2$$

$$\leqq \lambda_1\alpha'^2(n'^2/n)+\lambda_2\sum_{i=2}^{n}\gamma_i^2=\alpha'^2(n'^2/n)(\lambda_1-\lambda_2)+\lambda_2\|A\|^2 \tag{4.2}$$

$$\leqq \alpha'^2(n'^2/n)((s+1)(r+1)-\lambda_2)+\lambda_2\alpha' n'.$$

Combining these two as in (2.3) yields the statement of the theorem.

The main source of weakness in this modified bound is in (4.2), where the use of the eigenvalues for the original graph gives a relatively loose upper bound on $\|C\|^2$. For example, by taking an input node and an output node that are connected in a generalized N-gon with parameters r and s and eliminating all nodes that are at distance $N-2$ or less from either, one obtains a graph in which all input nodes have degree s and all output nodes degree r. The dominant eigenvalue for this graph is rs, and the subdominant is strictly less than the dominant because the graph is connected. When $r=s$, the bound of Theorem 2.1 shows that $c(\alpha)\geqq 1$ for all α, but the bound obtained from Theorem 5.1 allows $c(\alpha')<1$ at $\alpha'=1$.

6. Tensor product extension. The eigenvalue argument makes it easy to prove the properties of bsc's that are constructed by forming the tensor product $((\times))$ of two known bsc's [3, pp. 67–70].

THEOREM 6.1. *Let M_i be an $m_i \times n_i$ bsc incidence matrix such that $M_iM_i^T$ has eigenvalue ratio $\rho_i=\lambda_{i2}/\lambda_{i1}$, $i=1, 2$. Then the matrix $M'=M_1(\times)M_2$ is an $(m_1m_2)\times(n_1n_2)$ bsc incidence matrix with $\rho'=\max(\rho_1,\rho_2)$.*

Proof.

$$M'M'^T=(M_1(\times)M_2)(M_1(\times)M_2)^T=(M_1M_1^T)(\times)(M_2M_2^T).$$

Consequently the eigenvalues of $M'M'^T$ are $\lambda_{1i}\lambda_{2i}$ for $i=1,\cdots,n_1$, $j=1,\cdots,n_2$. Then

$$\rho'=\frac{\max(\lambda_{11}\lambda_{22},\lambda_{12}\lambda_{21})}{\lambda_{11}\lambda_{21}}=\max(\rho_1,\rho_2).$$

Applying this theorem when both constituents are the 6-gon with $s=r=7$ gives a square bsc with node degree $8^2=64$, $[(7^6-1)/(7-1)]^2$ input nodes, a like number of output nodes, and the same $c(\alpha)$ as the constituents.

7. Conclusion. We have exhibited a straightforward technique for lower bounding the concentration level of a bipartite graph. This technique may simplify substantially the analysis of new concentrator constructions and suggests new construction algebras such as the tensor product. Although the bound is not tight, particularly for small α, it is nonetheless strong when compared to previous bounding techniques. It can be strengthened by more detailed analysis of the relative size of the components of the input set vector in the eigenspaces of MM^T.

The bounding method is applied to the generalized N-gons and shows that these graphs are excellent concentrators. We surmise that they are indeed the best possible uniform concentrators of a given size and graph degree. While very large N-gons can be constructed, it is known that N-gons do not exist with arbitrarily large N, nor for arbitrary node degree, and thus our results fall short of giving an asymptotic construction.

Acknowledgments. The author wishes to thank Bruce Cooperstein for first drawing the generalized N-gons to his attention, Geoffrey Mason and Roger Wellington-Oguri for calculation and verification of their eigenvalues, and all three for many helpful discussions.

REFERENCES

[1] L. A. BASSALYGO AND M. S. PINSKER, *Complexity of an optimum non-blocking switching network without reconnections*, Problemy Peredachi Informatsii, 9 (1973), pp. 84–87, English translation in Problems of Information Transmission, Plenum, New York, 1975.

[2] N. BIGGS, *Algebraic Graph Theory*, Cambridge Tracts in Mathematics No. 67, Cambridge Univ. Press, Cambridge, 1974.

[3] D. M. CVETKOVIC, M. DOOB AND H. SACHS, *Spectra of Graphs*, Academic Press, New York, 1980.

[4] P. DEMBOWSKI, *Finite Geometries*, Springer, New York, 1968.

[5] W. FEIT AND G. HIGMAN, *The non-existence of certain generalized polygons*, J. Algebra, 1 (1964), pp. 114–131.

[6] O. GABBER AND Z. GALIL, *Explicit construction of linear size superconcentrators*, Proc. 20th Annual IEEE Symposium on Foundations of Computer Science, 29–31 Oct. 1979, San Juan, PR, pp. 364–370.

[7] W. HAEMERS, *Eigenvalue Techniques in Design and Graph Theory*, Mathematical Centre Tracts, 121, North-Holland, Amsterdam, 1979, pp. 50–51.

[8] T. LENGAUER, *Upper and lower bounds on time-space tradeoffs in a pebble game*, Dept. Computer Science Report 79-745, Stanford Univ., Stanford, CA, July 1979.

[9] R. J. MACWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977, pp. 651–672.

[10] G. A. MARGULIS, *Explicit construction of concentrators*, Problemy Peredachi Informatsii, 9 (1973), pp. 71–80; English translation in Problems of Information Transmission, Plenum, New York, 1975.

[11] W. J. PAUL, R. E. TARJAN AND J. R. CELONI, *Space bounds for a game on graphs*, Proc. 8th Annual ACM Symposium on Theory of Computing, Hershey, PA, May 1976, pp. 149–190.

[12] N. PIPPENGER, *Complexity theory*, Scientific American, June 1978, pp. 114–125.

[13] ———, *Superconcentrators*, SIAM J. Comput. 6 (1977), pp. 298–304.

[13] ———, *Generalized connectors*, SIAM J. Comput. 7 (1978), pp. 510–514.

[14] R. M. TANNER, *A recursive approach to low complexity codes*, IEEE Trans. Inform. Theory, IT-17 (1981), pp. 533–547.

[15] L. G. VALIANT, *On non-linear lower bounds in computational complexity*, Proc. 7th Annual ACM Symposium on Theory of Computing, Albuquerque, NM, May 1975, pp. 45–53.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

002

ON THE VORONOI REGIONS OF CERTAIN LATTICES*

J. H. CONWAY† AND N. J. A. SLOANE‡

Abstract. The Voronoi region of a lattice $L_n \subseteq \mathbb{R}^n$ is the convex polytope consisting of all points of $\mathbb{R}^n$ that are closer to the origin than to any other point of L_n. In this paper we calculate the second moments of the Voronoi regions of the lattices E_6^*, E_7^*, K_{12}, Λ_{16} and Λ_{24}. The results show that these lattices are the best quantizers presently known in dimensions 6, 7, 12, 16 and 24. The calculations are performed by Monte Carlo integration, and make use of fast algorithms for finding the closest lattice point to an arbitrary point of the space. We also establish two general theorems concerning the number of faces of the Voronoi region of a lattice.

AMS(MOS) subject classifications. Primary, 10E05, 52A45

1. Introduction. The Voronoi region of an n-dimensional lattice $L_n \subseteq \mathbb{R}^n$ is the convex polytope

$$\mathcal{P} := \{x \in \mathbb{R}^n : N(x) \leqq N(x-l) \text{ for all } l \in L_n\},$$

where $N(x) = x \cdot x$ denotes the norm of a vector (cf. [10] and the references given there). Figure 1 for example shows the Voronoi region of the body-centered cubic lattice A_3^*. If L_n is used as a quantizer (or analog-to-digital convertor), its average mean squared error per symbol is given by

$$(1) \qquad G(L_n) := \frac{1}{n} \frac{\int_{\mathcal{P}} x \cdot x \, dx}{\det(L_n)^{(n+2)/(2n)}},$$

where $\det(L_n)$, the determinant of L_n, is the square of the volume of $\mathcal{P}$ ([10], [22], [39]). $G(L_n)$ is a normalized second moment of $\mathcal{P}$ about the origin. (The formula (1) assumes that the input to the quantizer is uniformly distributed over a large region of $\mathbb{R}^n$, and the number of output levels is very large.)

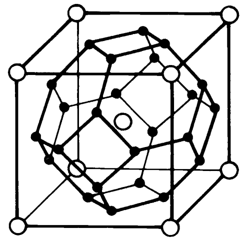


FIG. 1. *The Voronoi region of the body-centered cubic lattice $A_3^*(\cong D_3^*)$ is a truncated octahedron. The open circles represent lattice points, the solid circles the vertices of the Voronoi region.*

$G(L_n)$ measures the average error introduced when points of $\mathbb{R}^n$ are replaced by the closest lattice points. If we use the 1-dimensional integer lattice $\mathbb{Z}$ as a quantizer, the average error is $G(\mathbb{Z}) = 1/12 = 0.08333 \cdots$.But by using higher-dimensional lattices this can be reduced (see Table 1 and Fig. 2). The expression for $G(L_n)$ in (1) has been scaled so as to provide a proper comparison between quantizers of different dimensions.

* Received by the editors February 7, 1983, and in revised form May 27, 1983.

† Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB2 1SB, England.

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

TABLE 1
Smallest mean squared error $G(L_n)$ of any known n-dimensional lattice L_n

dimension n	lattice L_n	mean squared error $G(L_n)$
1	$\mathbb{Z}$	$0.0833333\cdots$
2	A_2	$0.0801875\cdots$
3	A_3^*	$0.0785433\cdots$
4	D_4	$0.0766032\cdots$
5	D_5^*	$0.0756254\cdots$
6	E_6^*	0.074239 ± 0.000018
7	E_7^*	0.073124 ± 0.000013
8	E_8	$0.0716821\cdots$
12	K_{12}	0.070100 ± 0.000024
16	Λ_{16}	0.068299 ± 0.000027
24	Λ_{24}	0.065771 ± 0.000074

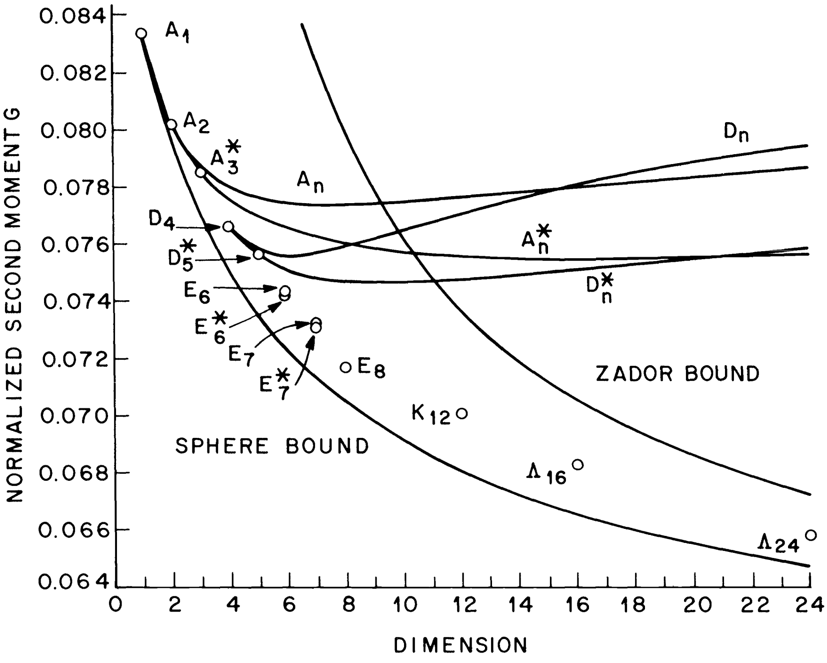


FIG. 2. *Normalized second moment G for various lattices, and the Zador and sphere bounds. It is known that the best quantizers must lie between the two bounds.*

In 1953 Fejes Tóth ([20], see also [30]) proved that the hexagonal lattice A_2 is the optimal lattice quantizer in two dimensions, i.e., has the smallest value of $G(L_2)$, namely $5/36\sqrt{3} = 0.0801875\cdots$. He also showed that no nonlattice quantizer can do better. Gersho [22] computed $G(A_3)$, $G(A_3^*)$ and $G(D_4)$, and conjectured that A_3^* ($\cong D_3^*$) is the optimal lattice in three dimensions. This conjecture was established in [2]. In an earlier paper [10] we determined the Voronoi regions and evaluated G for all the root lattices $A_n (n \geqq 1)$, $D_n (n \geqq 3)$, E_6, E_7, E_8 ($=E_8^*$) and the dual lattices A_n^* ($n \geqq 1$) and D_n^* ($n \geqq 3$). We observed that the optimal lattice quantizer was often the dual of the densest lattice packing, and conjectured that this may be true in general. It is true in dimensions 1, 2 and 3, and is supported by the available data in dimensions 4, 5 and 8. In the present paper we evaluate G for the duals of the densest known packings in 6, 7, 12, 16 and 24 dimensions, namely for the lattices E_6^*, E_7^*, the Coxeter–Todd lattice K_{12} ($\cong K_{12}^*$), the Barnes–Wall lattice Λ_{16} ($\cong \Lambda_{16}^*$), and the Leech lattice Λ_{24} ($= \Lambda_{24}^*$). The results are summarized in Table 1 and Fig. 2 and support our conjecture.

It is worth pointing out that this conjecture would imply the somewhat surprising result that the best quantizer is in general different from the most efficient lattice covering of space by spheres. Indeed, the two problems already have different answers in dimensions 4 and 5. There the best lattice coverings are known to be A_4^* and A_5^* [33], yet D_4^* and D_5^* are better quantizers (see Fig. 2). In higher dimensions, the best coverings known are A_n^*, if $n \leqq 23$, and then various lattices constructed from Λ_{24}[1]. So it seems likely that, in all dimensions between 4 and 23, the best lattice quantizers and coverings are distinct.

The values of G for various lattices are compared in Fig. 2, the values for A_n, A_n^*, D_n, D_n^*, E_6, E_7 and E_8 being taken from [10]. In 1964 Zador proved by a nonconstructive argument that good quantizers exist in sufficiently high dimensions, and observed that the second moment of a sphere gives a bound in the other direction (see [39], and also [22], [10, Eq. (3)]). These two bounds are also plotted in the figure.

We use Monte Carlo integration to compute G. The technique is briefly described in § 2. It requires that we have a fast algorithm for performing the quantizing, that is, given an arbitrary point of $\mathbb{R}^n$, for finding the closest lattice point. (If the lattice is used as a code for a band-limited channel, this algorithm performs the decoding [9], [11].) The best quantizing algorithms we have found for these lattices, and the values of G that were obtained, are given in §§ 3–7.

If a more complete description of the polytopes $\mathcal{P}$ were available, we could determine G exactly by decomposing each polytope into simplices and using the methods of [10]. Unfortunately little is known about the polytopes of these lattices. For E_6^* and E_7^*, for example, even the covering radius (the distance of the furthest vertex of $\mathcal{P}$ from the origin) is unknown, although it is known for all the other lattices mentioned [8], [10], [13], [15], [17]. In § 8 we establish two theorems which help determine the number of $(n-1)$-dimensional faces of $\mathcal{P}$, and in the last section we use them to study the Voronoi region of K_{12}.

Notation. If L_n is a lattice in $\mathbb{R}^n$ (the subscript indicates the dimension), the dual lattice $L_n^* = \{x \in \mathbb{R}^n : x \cdot y \in \mathbb{Z} \text{ for all } y \in L_n\}$. Two lattices L_n and M_n are equivalent, written $L_n \cong M_n$, if they differ only by a rotation and possibly a change of scale. The direct sum of k copies of L_n is written L_n^k. For further information about these lattices see [4], [10], [13]–[17], [27], [31], [36], [37].

2. Monte Carlo integration over a Voronoi region. The Monte Carlo technique that we use to calculate G is slightly unusual. We wish to find $I := \int_{\mathcal{P}} x \cdot x \, dx$. Conventional Monte Carlo methods ([24], [25], [28], [34]) would begin by replacing I by $\int_{\mathcal{Q}} \chi_{\mathcal{P}}(x) x \cdot x \, dx$, where $\mathcal{Q}$ is a region enclosing $\mathcal{P}$, usually a sphere or a cube, and $\chi_{\mathcal{P}}(x)$ is 1 if $x \in \mathcal{P}$, 0 otherwise. This is wasteful, since points in $\mathcal{Q} \setminus \mathcal{P}$ do not contribute to the estimate. The following approach avoids this difficulty, by exploiting our fast quantizing algorithms.

Let $v^{(1)}, \cdots, v^{(n)}$ be linearly independent vectors spanning the lattice, and let $u_1, \cdots, u_n$ be independent random numbers, uniformly distributed between 0 and 1. Then $y = \Sigma u_i v^{(i)}$ is uniformly distributed over the fundamental parallelepiped generated by the $v^{(i)}$. Let l be the closest lattice point to y (found by the quantizing algorithm for this lattice); then $w(y) := y - l$ is uniformly distributed over the Voronoi region $\mathcal{P}$. Figure 3 illustrates this for the hexagonal lattice A_2.

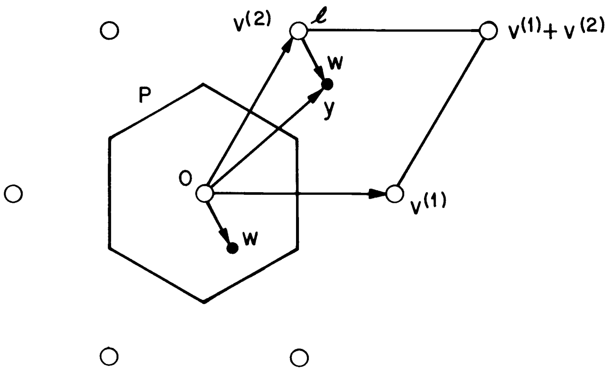


FIG. 3. *Hexagonal lattice A_2 with spanning vectors $v^{(1)}$, $v^{(2)}$; y is a random point in the parallelogram* 0, $v^{(1)}$, $v^{(2)}$, $v^{(1)}+v^{(2)}$; *l is the closest lattice point to y; and $w = y - l$ is a random point in the Voronoi region* $\mathcal{P}$.

Then if $m = gh$ random points $y^{(0)}, \cdots, y^{(m-1)}$ are selected in the manner just described,

$$\hat{I} = \frac{1}{m} \sum_{i=0}^{m-1} N(w(y^{(i)})) \tag{2}$$

is an estimate of I, and when suitably scaled (see (1)) produces our estimate $\hat{G}$ of G. To estimate the variance of $\hat{I}$, we group the measurements into g sets of h, and use the jackknife estimator (see [29], [32], [38]):

$$\widehat{\operatorname{var}(\hat{I})} = \frac{1}{g(g-1)} \sum_{j=0}^{g-1} \left(\frac{1}{h} A_j - \hat{I} \right)^2, \tag{3}$$

where

$$A_j = \sum_{i=jh}^{(j+1)h-1} N(w(y^{(i)})), \qquad j = 0, 1, \cdots, h-1. \tag{4}$$

By taking the square root of (3) and scaling, we obtain an estimate $\hat{\sigma}$ for the standard deviation of $\hat{G}$, and then $\hat{G} \pm 2\hat{\sigma}$ is our final estimate for G. The portable random number generator on the PORT library [21], which combines a congruential generator and a Tausworthe generator, was used to produce the u_i's.

To test this procedure we applied it first to the lattices E_6, E_7 and E_8, for which the exact value of G is known [10]. The estimates agreed closely with the exact values. For example, for E_8 we found using $m = 10^7$ points (g was always taken to be 100) that

$$\hat{G}(E_8) = 0.071689 \pm 0.000008,$$

while the true value is

$$G(E_8) = \frac{929}{12960} = 0.0716821 \cdots.$$

3. The lattices E_6 and E_6^*. E_6 is most easily obtained from the complex 3-dimensional $\mathbb{Z}[\omega]$-lattice with generator matrix

$$\begin{bmatrix} \theta & 0 & 0 \\ 0 & \theta & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad \omega = e^{2\pi i/3}, \quad \theta = \omega - \bar{\omega} = i\sqrt{3}, \tag{5}$$

([16, p. 421], [35, Example 5], [36, § 5.8.2], [37, § 17]), and the dual lattice E_6^* from

$$\begin{bmatrix} \theta & 0 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}. \tag{6}$$

Thus the real 6-dimensional lattices E_6 and E_6^* have generator matrices

$$\begin{bmatrix} 0 & \sqrt{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{3} & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ \frac{-3}{2} & \frac{-\sqrt{3}}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{-3}{2} & \frac{-\sqrt{3}}{2} & 0 & 0 \\ \frac{-1}{2} & \frac{\sqrt{3}}{2} & \frac{-1}{2} & \frac{\sqrt{3}}{2} & \frac{-1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}, \tag{7a}$$

$$\begin{bmatrix} 0 & \sqrt{3} & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ \frac{-3}{2} & \frac{-\sqrt{3}}{2} & 0 & 0 & 0 & 0 \\ \frac{-1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} & \frac{-\sqrt{3}}{2} & 0 & 0 \\ \frac{-1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 & \frac{1}{2} & \frac{-\sqrt{3}}{2} \end{bmatrix} \tag{7b}$$

respectively. (For example (7a) is obtained from the real and imaginary parts of (5) and ω times (5).)

To find a quantizing algorithm for any of these lattices, we proceed as follows, following [11]. Inside our lattice L_n we look for a sublattice S_n, of small index t (say), for which quantizing is easy. Suppose

$$L_n = (a^{(0)} + S_n) \cup \cdots \cup (a^{(t-1)} + S_n), \tag{8}$$

where $a^{(0)}, \cdots, a^{(t-1)}$ are coset representatives for S_n in L_n. Let $\phi : \mathbb{R}^n \to S_n$ be a quantizer for S_n, so that $\phi(x)$ is the closest point of S_n to a given vector x. Then a quantizer for L_n is obtained by taking the given vector w, forming the t candidates

$$c^{(i)} = \phi(w - a^{(i)}) + a^{(i)}, \qquad i = 0, \cdots, t-1,$$

and choosing the closest candidate to w.

Both E_6 and E_6^* contain a sublattice S_6 isomorphic to A_2^3, namely the real version of the lattice $\theta\mathbb{Z}[\omega]^3$ with generator matrix

$$\begin{bmatrix} \theta & 0 & 0 \\ 0 & \theta & 0 \\ 0 & 0 & \theta \end{bmatrix}.$$

S_6 has index 3 in E_6, with coset representatives

$$a^{(0)}=(0,0,0,0,0,0), \quad a^{(1)}=(1,0,1,0,1,0), \quad a^{(2)}=-a^{(1)},$$

and has index 9 in E_6^*, with coset representatives given in Table 2.

TABLE 2
Coset representatives $a^{(0)}, \cdots, a^{(8)}$ for S_6 in E_6^*

(0,	0,	0,	0,	0,	0)
±(1,	0,	−1,	0,	0,	0)
±(0,	0,	1,	0,	−1,	0)
±(−1,	0,	0,	0,	1,	0)
±(1,	0,	1,	0,	1,	0)

It is easy to design a quantizer for the hexagonal lattice A_2 (and hence A_2^3), either using the fact that A_2 is the union of a rectangular lattice and a translate, as suggested by Gersho [23, p. 165], or via 3-dimensional coordinates as we suggest in [10, § VII]. Gersho's method seems slightly simpler and is the one we adopted. From the previous discussion we have the following quantizing algorithms for E_6 and E_6^*.

QUANTIZING ALGORITHMS FOR E_6 $(t=3)$ AND E_6^* $(t=9)$

Given $w=(w_1, \cdots, w_6)$, to find the closest lattice point of E_6 (or E_6^*).

Subtract one of t coset representatives $a^{(i)}$, obtaining

$$z=w-a^{(i)}=(z_1, \cdots, z_6).$$

Divide the corresponding complex vector by θ, i.e., form

$$z'=\left(\frac{z_2}{\sqrt{3}}, \frac{-z_1}{\sqrt{3}}, \frac{z_4}{\sqrt{3}}, \frac{-z_3}{\sqrt{3}}, \frac{z_6}{\sqrt{3}}, \frac{-z_5}{\sqrt{3}}\right).$$

Apply the quantizer for A_2 to the three pairs

$$\left(\frac{z_2}{\sqrt{3}}, \frac{-z_1}{\sqrt{3}}\right), \left(\frac{z_4}{\sqrt{3}}, \frac{-z_3}{\sqrt{3}}\right), \left(\frac{z_6}{\sqrt{3}}, \frac{-z_5}{\sqrt{3}}\right)$$

obtaining say

$$(m_1, m_2), (m_3, m_4), (m_5, m_6).$$

Multiply by "θ", to get

$$m'=(-\sqrt{3}m_2, \sqrt{3}m_1, -\sqrt{3}m_4, \sqrt{3}m_3, -\sqrt{3}m_6, \sqrt{3}m_5).$$

Then $c^{(i)}=m'+a^{(i)}$ is the ith candidate.

The final answer is the candidate which minimizes $N(w-c^{(i)})$.

We used this algorithm in the Monte Carlo procedure described in § 2, in order to estimate $G(E_6^*)$. With 5×10^6 random points we obtained

$$\hat{G}(E_6^*)=0.074239\pm0.000018. \tag{9}$$

(As a check, 10^7 points for E_6 gave $\hat{G}(E_6)=0.074342\pm0.000013$, while the exact value from [10] is $G(E_6)=3^{-1/6}5/56=0.07434671\cdots$, well within the range of the estimate.)

4. The lattices E_7 and E_7^*. E_7 may be obtained by applying Construction A of [27] to the little [7, 3, 4] Hamming code $\mathcal{H}_1$, and a generator matrix for E_7 may be found on [37, p. 335] (see also [4], [16], [13], [27]). Similarly E_7^* may be obtained from the [7, 4, 3] Hamming code $\mathcal{H}_2$. Both E_7 and E_7^* contain a sublattice $S_7=2\mathbb{Z}^7$, of index 8 and 16 respectively; the coset representatives are the codewords of either $\mathcal{H}_1$ or $\mathcal{H}_2$. Since there is a trivial quantizer for $\mathbb{Z}^7$ (see [11, § III]), this leads to fast quantizing algorithms for E_7 and E_7^*; we omit the details. (Alternative algorithms, based on the sublattice A_7, were proposed in [11, § VII].) Our Monte Carlo estimate for $G(E_7^*)$, based on 10^7 points, is

$$\hat{G}(E_7^*)=0.073124\pm0.000013. \tag{10}$$

This is slightly better than the value for E_7, which is $0.07323063\cdots$.

5. The Coxeter–Todd lattice K_{12}. The real 12-dimensional lattice K_{12} was first described in [18]. It is the subject of our earlier paper [15], and further properties will be found in § 9 below and in [13], [14], [27]. Since it is the densest sphere packing known in 12 dimensions [13], [27], and is also equivalent to its dual, according to the conjecture mentioned in § 1 it is a good candidate for a quantizer. The same remark applies to the lattices Λ_{16} and Λ_{24} studied in the following sections.

Regarded as a complex 6-dimensional $\mathbb{Z}[\omega]$-lattice, K_{12} has the generator matrix

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & \omega & \bar{\omega} \\ 1 & 0 & 1 & 0 & \omega & \bar{\omega} \\ \omega & \bar{\omega} & 0 & 1 & 0 & 1 \end{bmatrix}, \tag{11}$$

and so as a real 12-dimensional lattice it has the generator matrix shown in Fig. 4. K_{12} has a sublattice isomorphic to A_2^6, of index 64, namely the real version of the lattice $2\mathbb{Z}[\omega]^6$. The coset representatives correspond to the codewords of the hexacode (the code over $GF(4)=\{0,1,\omega,\bar{\omega}\}$ spanned by the last three rows of (11)). This leads to a quantizing algorithm similar to those for E_6 and E_6^* described in § 3.

There are alternative definitions of K_{12} (see [15]), which make four other sublattices visible, namely the real lattices corresponding to $\mathbb{Z}[\omega]\otimes A_6$, $\mathbb{Z}[\omega]\otimes D_6$, $\mathbb{Z}[\omega]\otimes E_6$, and the lattice $\{(x_1,\cdots,x_6)\colon \text{all } x_i\in\mathbb{Z}[\omega],\ \Sigma x_i\equiv 0 \pmod{\theta}\}$. Each of these leads to a decoding algorithm. However the one described above seems to be the simplest.

Our Monte Carlo estimate, based on 10^6 points, is

$$\hat{G}(K_{12})=0.070100\pm0.000024. \tag{12}$$

6. The Barnes–Wall lattice Λ_{16}. The lattice Λ_{16} was first described in [3]. Other references are [13] and [27], and a generator matrix is given on page 336 of [37]. Λ_{16} has a sublattice $2D_{16}$ of index 32, with coset representatives which are the codewords of the [16, 5, 8] first-order Reed–Muller code. Then the quantizing algorithm for D_n

2	0	0	0	0	0	0	0	0	0	0	0
0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0	$-1/2$	$\sqrt{3}/2$	$-1/2$	$-\sqrt{3}/2$
1	0	0	0	1	0	0	0	$-1/2$	$\sqrt{3}/2$	$-1/2$	$-\sqrt{3}/2$
$-1/2$	$\sqrt{3}/2$	$-1/2$	$-\sqrt{3}/2$	0	0	1	0	0	0	1	0
-1	$\sqrt{3}$	0	0	0	0	0	0	0	0	0	0
0	0	-1	$\sqrt{3}$	0	0	0	0	0	0	0	0
0	0	0	0	-1	$\sqrt{3}$	0	0	0	0	0	0
0	0	$-1/2$	$\sqrt{3}/2$	0	0	$-1/2$	$\sqrt{3}/2$	$-1/2$	$-\sqrt{3}/2$	1	0
$-1/2$	$\sqrt{3}/2$	0	0	$-1/2$	$\sqrt{3}/2$	0	0	$-1/2$	$-\sqrt{3}/2$	1	0
$-1/2$	$-\sqrt{3}/2$	1	0	0	0	$-1/2$	$\sqrt{3}/2$	0	0	$-1/2$	$\sqrt{3}/2$

FIG. 4. *Generator matrix for Coxeter–Todd lattice K_{12}.*

described in [11, § IV] leads to an efficient quantizer for Λ_{16}. Using $2 \cdot 10^6$ points we found

$$\hat{G}(\Lambda_{16}) = 0.068299 \pm 0.000027. \tag{13}$$

7. The Leech lattice Λ_{24}. The Leech lattice Λ_{24} ([5]–[8], [12]–[14], [26], [27], [36], [37]) may be constructed in many ways. The standard MOG (or miracle octad generator [7], [8], [19]) basis is shown in Fig. 5. Λ_{24} has a sublattice $4D_{24}$ of index 8,192, with coset representatives

$$2c \text{ and } 2c+u,$$

where $u = (-3, 1, 1, \cdots, 1)$, and c runs through the vectors of the [24, 12, 8] Golay code with generator matrix shown in Fig. 6. Because of the large index of the sublattice, this is by far the slowest of our quantizing algorithms. Using 25,000 points we found

$$\hat{G}(\Lambda_{24}) = 0.065771 \pm 0.000074. \tag{14}$$

8. The number of faces of the Voronoi region. The Voronoi region $\mathcal{P}$ of a lattice $L_n \subseteq \mathbb{R}^n$ may be expressed as

$$\mathcal{P} = \bigcap_{v \in L_n, v \neq 0} \mathcal{S}(v), \tag{15}$$

where $\mathcal{S}(v)$ is the half-space $\{x \in \mathbb{R}^n : x \cdot v \leqq \frac{1}{2} v \cdot v\}$, bounded by the hyperplane

$$\Pi(v) = \{x \in \mathbb{R}^n : x \cdot v = \tfrac{1}{2} v \cdot v\}.$$

Of course only finitely many of the $\mathcal{S}(v)$ are really needed to define $\mathcal{P}$. Let

$$\mathcal{P} = \bigcap_{v \in \mathcal{R}} \mathcal{S}(v),$$

where $\mathcal{R}$ is a minimal subset of $L_n \backslash \{0\}$ that will define $\mathcal{P}$. We call the lattice vectors in $\mathcal{R}$ *relevant*, and the remaining vectors of L_n *irrelevant*. The number of relevant vectors is the number of $(n-1)$-dimensional faces of $\mathcal{P}$. For example, in the body-centred cubic lattice (Fig. 1) there are 14 relevant vectors, namely the six minimal vectors and the eight vectors of the next smallest norm.

8 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 4 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 4 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 4	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	4 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	0 4 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	0 0 4 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
2 2 2 2	2 2 2 2	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	0 0 0 0	4 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	0 0 0 0	0 4 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	0 0 0 0	0 0 4 0	0 0 0 0	0 0 0 0	0 0 0 0
2 2 2 2	0 0 0 0	2 2 2 2	0 0 0 0	0 0 0 0	0 0 0 0
4 0 0 0	0 0 0 0	0 0 0 0	4 0 0 0	0 0 0 0	0 0 0 0
2 2 0 0	2 2 0 0	2 2 0 0	2 2 0 0	0 0 0 0	0 0 0 0
2 0 2 0	2 0 2 0	2 0 2 0	2 0 2 0	0 0 0 0	0 0 0 0
2 0 0 2	2 0 0 2	2 0 0 2	2 0 0 2	0 0 0 0	0 0 0 0
4 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	4 0 0 0	0 0 0 0
2 0 2 0	2 0 0 2	2 2 0 0	0 0 0 0	2 2 0 0	0 0 0 0
2 0 0 2	2 2 0 0	2 0 2 0	0 0 0 0	2 0 2 0	0 0 0 0
2 2 0 0	2 0 2 0	2 0 0 2	0 0 0 0	2 0 0 2	0 0 0 0
0 2 2 2	2 0 0 0	2 0 0 0	2 0 0 0	2 0 0 0	2 0 0 0
0 0 0 0	0 0 0 0	2 2 0 0	2 2 0 0	2 2 0 0	2 2 0 0
0 0 0 0	0 0 0 0	2 0 2 0	2 0 2 0	2 0 2 0	2 0 2 0
−3 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1

FIG. 5. *Generator matrix for Leech lattice* Λ_{24} *in standard* MOG *form.*

1 1 1 1	1 1 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
1 1 1 1	0 0 0 0	1 1 1 1	0 0 0 0	0 0 0 0	0 0 0 0
1 1 0 0	1 1 0 0	1 1 0 0	1 1 0 0	0 0 0 0	0 0 0 0
1 0 1 0	1 0 1 0	1 0 1 0	1 0 1 0	0 0 0 0	0 0 0 0
1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 1	0 0 0 0	0 0 0 0
1 0 1 0	1 0 0 1	1 1 0 0	0 0 0 0	1 1 0 0	0 0 0 0
1 0 0 1	1 1 0 0	1 0 1 0	0 0 0 0	1 0 1 0	0 0 0 0
1 1 0 0	1 0 1 0	1 0 0 1	0 0 0 0	1 0 0 1	0 0 0 0
0 1 1 1	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0
0 0 0 0	0 0 0 0	1 1 0 0	1 1 0 0	1 1 0 0	1 1 0 0
0 0 0 0	0 0 0 0	1 0 1 0	1 0 1 0	1 0 1 0	1 0 1 0
1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1

FIG. 6. *Generator matrix for Golay code in standard* MOG *form.*

[10, § III] contains two theorems which give sufficient conditions for a lattice to have the property that the minimal vectors are the only relevant vectors. The lattices A_n, D_n, E_6, E_7 and E_8 have this property [10, Corollary to Theorem 5]. The following theorems can also be used to show that certain vectors are irrelevant.

THEOREM 1. *If* L_n *has covering radius* R_c (*cf.* [8]), *then any vector* $v \in L_n$ *of norm* $\geqq 4R_c^2$ *is irrelevant.*

Proof. Suppose $v \in L_n$ is relevant and $N(v) \geqq 4R_c^2$. Then $\Pi(v)$ meets $\mathcal{P}$ in a set of positive measure. On the other hand the closest point of $\Pi(v)$ to the origin, $\frac{1}{2}v$, has norm $\geqq R_c^2$, so $\Pi(v)$ can contain at most a single vertex of $\mathcal{P}$, which is a contradiction. Q.E.D.

For example, Theorem 1 and the main result of [8] imply that for the Leech lattice only the minimal vectors and those of the next smallest norm are relevant. Thus the Voronoi region has 16,969,680 faces.

When attempting to show that a vector $v \in L_n$ is irrelevant, it is sometimes possible to prove that the point $\frac{1}{2}v$ is not in the Voronoi region. In general this is not enough to prove that v is irrelevant, since it is certainly possible for the hyperplane $\Pi(v)$ to intersect $\mathcal{P}$ in an asymmetric region not containing $\frac{1}{2}v$. The next theorem establishes a condition under which this does not happen.

THEOREM 2. *Let $\mathcal{G}_v$ be the subgroup of the automorphism group* Aut (L_n) *fixing a lattice vector v. Suppose $\mathcal{G}_v$ fixes exactly one* 1*-dimensional subspace, namely the* 1*-dimensional subspace containing v. Then if the hyperplane $\Pi(v)$ contains a point x of the Voronoi region $\mathcal{P}$ different from $\frac{1}{2}v$, it follows that $\frac{1}{2}v \in \mathcal{P}$.*

Proof. Consider the point

$$y = \frac{1}{|\mathcal{G}_v|} \sum_{g \in \mathcal{G}_v} x^g.$$

Since $x \in \mathcal{P} \cap \Pi(v)$, $x^g \in \mathcal{P} \cap \Pi(v)$. Because $\mathcal{P}$ is convex, $y \in \mathcal{P} \cap \Pi(v)$. Clearly y is fixed by $\mathcal{G}_v$. But the only point of $\Pi(v)$ fixed by $\mathcal{G}_v$ is $\frac{1}{2}v$. Therefore $y = \frac{1}{2}v$ and $\frac{1}{2}v \in \mathcal{P}$. Q.E.D.

An application of these theorems will be found in the following section.

9. The Voronoi region of K_{12}. We shall determine the number of faces of the Voronoi region $\mathcal{P}$ of K_{12}. We assume K_{12} is scaled so that the minimal norm is 2. It was shown in [15] that the covering radius of K_{12} is $\sqrt{4/3}$, and that there are 20,412 vertices of $\mathcal{P}$ at this distance from the origin, all equivalent under Aut (K_{12}). Unfortunately nothing is known about other vertices of $\mathcal{P}$.

K_{12} is best studied via the corresponding 6-dimensional lattice Λ_6^ω (see [15]). The latter has several equivalent definitions, one of which may be seen in (11). In this section, however, another construction is more convenient, the so-called 3-base. In this form Λ_6^ω is defined as the set of vectors

$$\{(x_1, \cdots, x_6)_3\text{: all } x_i \in \mathbb{Z}[\omega],\ x_1 \equiv \cdots \equiv x_6 \pmod{\theta} \text{ and } \Sigma x_i \equiv 0 \pmod{3}\},$$

where $(x_1, \cdots, x_6)_3$ is an abbreviation for $\theta^{-1}(x_1, \cdots, x_6)$.

THEOREM 3. *The Voronoi region $\mathcal{P}$ of K_{12} has* 4,788 11*-dimensional faces, and is bounded by the hyperplanes determined by the* 756 *vectors of norm* 2 *and the* 4,032 *vectors of norm* 3.

Proof. We shall show that the vectors of norm $\geqq 4$ in K_{12} are irrelevant. Theorem 1 already implies that the vectors of norm $\geqq 6$ are irrelevant. Also it is easy to verify that Aut (K_{12}) is transitive on the vectors of norms 2, 3, 4 and 5. The vectors of norms 2, 3 and 4, expressed in the 3-base, are listed in Table II of [15].

(i) *The vectors of norm* 5 *are irrelevant.* In view of the transitivity just mentioned, it is enough to show that a single norm 5 vector, say $v_5 = (2\theta, \theta, 0, 0, 0, 0)_3$, is irrelevant. We first show that $\frac{1}{2}v_5 \notin \mathcal{P}$. In fact it is easily checked that $\frac{1}{2}v_5$ is outside the hyperplane $\Pi(v_2)$ determined by $v_2 = (\omega, \omega, \omega, \omega, \bar{\omega}, 1)_3$, and therefore by (15) is not in $\mathcal{P}$.

Next, let $\mathscr{F}$ be the stabilizer of v_5 in Aut (Λ_6^ω). We show that $\mathscr{F}$ only fixes a (complex) 1-dimensional subspace. Since $\mathscr{F}$ contains the diagonal matrices

$$\operatorname{diag}\{1, 1, 1, 1, \omega, \bar{\omega}\},$$

where the ω and $\bar{\omega}$ may be in any of the last four positions, any vector fixed by $\mathscr{F}$ must have the form $f = (\alpha, \beta, 0, 0, 0, 0)_3$. Also $\mathscr{F}$ contains the reflection

$$R_w : x \mapsto x - x \cdot \bar{w}\, w,$$

where $w = (1, -2, 1, 1, 1, 1)_3$, and so $f = (2\beta, \beta, 0, 0, 0, 0)_3$; in other words f is a multiple of v_5.

Let $\mathscr{G}_{v_5}$ be the stabilizer in Aut (K_{12}) of the real vector corresponding to v_5. Since Aut (K_{12}) contains a transformation which corresponds to complex conjugation in Λ_6^ω, only real multiples of v_5 are fixed by $\mathscr{G}_{v_5}$. Thus $\mathscr{G}_{v_5}$ only fixes a single 1-dimensional space. We can conclude from Theorem 2 that v_5 is irrelevant.

(ii) *The vectors of norm* 4 *are irrelevant.* It is enough to consider one vector of norm 4, say $v_4 = (1 + 3\omega, 1, 1, 1, 1, 1)_3$. One can now show that the equation to the hyperplane $\Pi(v_4)$ is already implied by the equations to the ten hyperplanes $\Pi(v_2)$ passing through $\frac{1}{2}v_4$, and defined by the following vectors v_2:

$$\begin{gathered}(\omega, \bar{\omega}, 1, 1, 1, 1)_3, \\ \cdots \\ (\omega, 1, 1, 1, 1, \bar{\omega})_3, \\ (\theta, -\omega\theta, 0, 0, 0, 0)_3, \\ \cdots \\ (\theta, 0, 0, 0, 0, -\omega\theta)_3.\end{gathered}$$

We leave this verification to the reader, as well as the easy justification that the vectors of norms 2 and 3 are relevant. Q.E.D.

Acknowledgments. We should like to thank W. M. Kantor for some helpful discussions about K_{12}, and C. L. Mallows for valuable statistical advice concerning the design of the Monte Carlo experiments. These experiments were carried out on the Bell Labs Cray-1 computer, and the total computing time used (excluding test runs) was about $8\frac{1}{2}$ hours.

REFERENCES

[1] R. P. Bambah and N. J. A. Sloane, *On a problem of Ryskov concerning lattice coverings*, Acta Arithmetica, 42 (1982), pp. 107–109.

[2] E. S. Barnes and N. J. A. Sloane, *The optimal lattice quantizer in three dimensions*, this Journal, 4 (1983), pp. 30–41.

[3] E. S. Barnes and G. E. Wall, *Some extreme forms defined in terms of Abelian groups*, J. Austral. Math. Soc., 1 (1959), pp. 47–63.

[4] N. Bourbaki, *Groupes et algèbres de Lie*, Hermann, Paris, 1968, Chap. 4–6.

[5] J. H. Conway, *A characterisation of Leech's lattice*, Inventiones Math., 7 (1969), pp. 137–142.

[6] ———, *Three lectures on exceptional groups*, in Finite Simple Groups, M. B. Powell and G. Higman, eds., Academic Press, New York, 1971, pp. 215–247.

[7] ———, *The miracle octad generator*, in Topics in Group Theory and Computation, M. P. J. Curran, ed., Academic Press, New York, 1977, pp. 62–68.

[8] J. H. Conway, R. A. Parker and N. J. A. Sloane, *The covering radius of the Leech lattice*, Proc. Royal Soc. London, A 380 (1982), pp. 261–290.

[9] J. H. Conway and N. J. A. Sloane, *Fast* 4- *and* 8-*dimensional quantizers and decoders*, National Telecommunications Record 1981, IEEE Press, New York, 1981, Vol. 3, pp. F4.2.1 to F4.2.4.

[10] J. H. CONWAY AND N. J. A. SLOANE, *Voronoi regions of lattices, second moments of polytopes, and quantization*, IEEE Trans. Information Theory, IT-28 (1982), pp. 211–226.

[11] ———, *Fast quantizing and decoding algorithms for lattice quantizers and codes*, IEEE Trans. Information Theory, IT-28 (1982), pp. 227–232.

[12] ———, *Twenty-three constructions for the Leech lattice*, Proc. Royal Soc. London, A 381 (1982), pp. 275–283.

[13] ———, *Laminated lattices*, Ann. Math. 116 (1982), pp. 593–620.

[14] ———, *Complex and integral laminated lattices*, Trans. Amer. Math. Soc., in press.

[15] ———, *The Coxeter–Todd lattice, the Mitchell group, and related sphere packings*, Math. Proc. Cambridge Phil. Soc., 93 (1983), pp. 421–440.

[16] H. S. M. COXETER, *Extreme forms*, Canad. J. Math., 3 (1951), pp. 391–441.

[17] ———, *Regular Polytopes*, Dover, New York, third edition, 1973.

[18] H. S. M. COXETER AND J. A. TODD, *An extreme duodenary form*, Canad. J. Math., 5 (1953), pp. 384–392.

[19] R. T. CURTIS, *On subgroups of* $\cdot 0$. 1: *Lattice stabilizers*, J. Algebra, 27 (1973), pp. 549–573.

[20] L. FEJES TÓTH, *Lagerungen in der Ebene, auf der Kugel und im Raum*, Springer-Verlag, Berlin, 1953. Second edition, corrected and expanded, Springer-Verlag, 1972.

[21] *The* PORT *Mathematical Subroutine Library*, P. A. Fox, ed., Bell Labs, Murray Hill, NJ, second edition, 1977.

[22] A. GERSHO, *Asymptotically optimal block quantization*, IEEE Trans. Information Theory, IT-25 (1979), pp. 373–380.

[23] ———, *On the structure of vector quantizers*, IEEE Trans. Information Theory, IT-28 (1982), pp. 157–166.

[24] J. H. HALTON, *A retrospective and prospective survey of the Monte Carlo method*, SIAM Rev., 12 (1970), pp. 1–63.

[25] J. M. HAMMERSLEY AND D. C. HANDSCOMB, *Monte Carlo Methods*, John Wiley, NY, 1964.

[26] J. LEECH, *Notes on sphere packings*, Canad. J. Math., 19 (1967), pp. 251–267.

[27] J. LEECH AND N. J. A. SLOANE, *Sphere packing and error-correcting codes*, Canad. J. Math., 23 (1971), pp. 718–745.

[28] *Symposium on Monte Carlo Methods*, H. A. Meyer, ed., John Wiley, New York, 1956.

[29] R. G. MILLER, *The jackknife—a review*, Biometrika, 61 (1974), pp. 1–15.

[30] D. J. NEWMAN, *The hexagonal theorem*, IEEE Trans. Information Theory, IT-28 (1982), pp. 137–139.

[31] H.-V. NIEMEIER, *Definite quadratische Formen der Dimension* 24 *und Diskriminante* 1, J. Number Theory, 5 (1973), pp. 142-178.

[32] M. H. QUENOUILLE, *Notes on bias in estimation*, Biometrika, 43 (1956), pp. 353–360.

[33] S. S. RYSKOV AND E. P. BARANOVSKII, *C-types of n-dimensional lattices and* 5*-dimensional primitive parallelohedra* (*with application to the theory of coverings*) (in Russian), Trudy Mat. Inst. Steklov., 137 (1976); English translation in Proc. Steklov. Inst. Math., Issue 4, 1978.

[34] YU. A. SHREIDER, *Method of Statistical Testing*: *Monte Carlo Method.* Elsevier, New York, 1964.

[35] N. J. A. SLOANE, *Codes over GF*(4) *and complex lattices*, J. Algebra, 52 (1978), pp. 168–181.

[36] ———, *Self-dual codes and lattices*, in Relations Between Combinatorics and Other Parts of Mathematics, Proc. Symposia in Pure Mathematics, 24, American Mathematical Society, Providence, RI, 1979, pp. 273–308.

[37] ———, *Tables of sphere packings and spherical codes*, IEEE Trans. Information Theory, IT-27 (1981), pp. 327-338.

[38] J. W. TUKEY, *Bias and confidence in not-quite large samples* (Abstract), Ann. Math. Stat., 29 (1958), pp. 614.

[39] P. ZADOR, *Asymptotic quantization error of continuous signals and the quantization dimension*, IEEE Trans. Information Theory, IT-28 (1982), pp. 139–149.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

003

A SEPARATOR THEOREM FOR CHORDAL GRAPHS*

JOHN R. GILBERT†, DONALD J. ROSE‡ AND ANDERS EDENBRANDT§

Abstract. Chordal graphs are undirected graphs in which every cycle of length at least four has a chord. They are sometimes called triangulated graphs, monotone transitive graphs, rigid circuit graphs, or perfect elimination graphs; the last name reflects their utility in modelling Gaussian elimination on sparse matrices. The main result of this paper is that a chordal graph with n vertices and m edges can be cut in half by removing $O(\sqrt{m})$ vertices. A similar result holds if the vertices have nonnegative weights and we want to bisect the graph by weight, or even if we want to bisect the graph simultaneously by several unrelated sets of weights. We present an $O(m)$ time algorithm to find the separating set.

AMS(MOS) subject classifications. 05C40, 68E10

1. Introduction.

Many divide-and-conquer algorithms on graphs are based on finding a small set of vertices or edges whose removal divides the graph roughly in half. Examples include layout of circuits in a model of VLSI [11], efficient sparse Gaussian elimination [5, 13], and the solution of various geometric problems [14].

Most graphs do not have small separators that divide them evenly in half, but some useful ones do. Lipton and Tarjan's planar separator theorem gives an example.

PROPOSITION [15]. *A planar graph with n vertices has a set of at most $2\sqrt{2n}$ vertices whose removal leaves no component with more than $2n/3$ vertices.* □

This theorem is the best possible within a constant factor. Djidjev [3] improved the constant $2\sqrt{2}$ to $\sqrt{6}$; the tightest possible constant is not known. Other kinds of graphs that can be separated evenly by deleting $o(n)$ vertices are trees ($O(1)$ vertices [8, 12]), outerplanar graphs ($O(1)$ vertices [11]), hypercubes ($O(n/\sqrt{\log n})$ vertices [5]), graphs of genus at most g ($O(\sqrt{gn})$ vertices [6]), and several interconnection graphs for parallel computation [9, 10, 11].

An undirected graph is said to be *chordal* if every cycle of length at least four has a *chord*, which is an edge joining two vertices that are not adjacent on the cycle. Chordal graphs are perfect; that is, every induced subgraph of a chordal graph has a clique covering and an independent set of the same size [7]. Much of the basic theory of chordal graphs was developed by Dirac [2] and Fulkerson and Gross [4]. Chordal graphs have also been called triangulated graphs, monotone transitive graphs, rigid circuit graphs, and perfect elimination graphs.

Rose [18] discovered a connection between chordal graphs and systems of linear equations whose coefficient matrices are sparse, symmetric, and positive definite. Such a system can be solved using Gaussian elimination with pivots chosen from the diagonal. The coefficient matrix is the adjacency matrix of an undirected graph; the graph is chordal if and only if the elimination can be done in some order without fill-in, that is, without changing any zero entries to nonzeros.

Since a complete graph is chordal and has only trivial separators, chordal graphs in general cannot be separated by removing $o(n)$ vertices. The main result of this paper is that chordal graphs do satisfy a separator theorem in which the size of the separator depends on the density of the graph. We prove that a chordal graph with

* Received by the editors November 18, 1982, and in revised form May 27, 1983.

† Computer Science Department, Cornell University, Ithaca, New York 14853. The work of this author was supported in part by the National Science Foundation under grant MCS-82-02948.

‡ Bell Laboratories, Murray Hill, New Jersey 07974.

§ Computer Science Department, Cornell University, Ithaca, New York 14853. The work of this author was supported in part by the National Science Foundation under grant MCS-81-01220.

n vertices and m edges has a set of $O(\sqrt{m})$ vertices whose removal leaves no component with more than $n/2$ vertices. (This is immediate at the extremes of density, for complete graphs and for trees). We show that the separator can in fact be chosen to be a complete subgraph. We also show that the result holds if the vertices have nonnegative weights and we want to bisect the graph by weight, or even if we want to bisect the graph simultaneously by several unrelated sets of vertex weights.

The next section contains some definitions and results from the literature that we will need later. § 3 proves the main result. § 4 presents a linear algorithm to find the separator. § 5 extends the main result to graphs whose vertices have multiple weights. The final § describes possible applications and open problems.

2. Results from the literature. The first results we require concern the graph model of Gaussian elimination. Let $G=(V,E)$ be a (not necessarily chordal) graph. Let v be a vertex of G. The *deficiency* of v is the set of nonedges between neighbors of v,

$$D(v)=\{\{x,y\}:\{v,x\}\in E, \{v,y\}\in E, \{x,y\}\notin E\}.$$

The deficiency of v corresponds to the zeros of the coefficient matrix that become nonzero when the equation in v's row is used to eliminate the variable in v's column. The graph G_v produced by *eliminating* v from G is obtained by adding v's deficiency and deleting v and its incident edges, so

$$G_v=(V-\{v\}, E(V-\{v\})\cup D(v)).$$

When a sequence of vertices is eliminated from a graph, the edges in the deficiencies that are added are called *fill-in* edges. A *simplicial vertex* of a graph is a vertex that has a null deficiency, so it can be eliminated without fill-in; thus, it is a vertex whose neighbors form a clique. A graph G is a *perfect elimination graph* if its vertices can all be eliminated in some order without any fill-in. Such an order is called a *perfect elimination ordering* of the vertices of G.

The lemmas that follow are due to Dirac [2], Fulkerson and Gross [4], and Rose, Tarjan, and Lueker [16], [18], [19].

LEMMA 1. *A graph G is chordal if and only if it is a perfect elimination graph.* □

A perfect elimination ordering must start with a simplicial vertex. Any simplicial vertex will do, and a choice of simplicial vertices is always available.

LEMMA 2. *If G is chordal and C is any complete proper subgraph, then there is a simplicial vertex in $G-C$. Any simplicial vertex can be eliminated first in some perfect elimination ordering.* □

Now we give a condition that determines the fill-in for any elimination ordering on any graph.

LEMMA 3. *Fix an elimination ordering for a graph G. Let v and w be nonadjacent vertices of G. Then $\{v,w\}$ is a fill-in edge if and only if there is a path from v to w consisting of vertices that are eliminated earlier than both v and w.* □

In an ordered graph, a path is a *monotone path* if the indices of its vertices are strictly increasing.

LEMMA 4. *Fix a perfect elimination ordering $v_1, \cdots, v_n$ for a chordal graph G. If $k<h$ and there is a path from v_k to v_h through vertices numbered at most h, then there is a monotone path from v_k to v_h.* □

A *separation clique* is a complete subgraph whose removal leaves a disconnected graph.

LEMMA 5. *If G is chordal and not complete, then G has at least one separation clique.* □

We mention the following result to contrast it with the first theorem of the next section; the proofs below do not use it. A v, w *separator* is a set of vertices that cuts every path from v to w.

LEMMA 6. *A graph G is chordal if and only if for all vertices v and w, every minimal v, w separator in G is a clique.* □

3. A $\sqrt{m}$-vertex separator theorem. Let G be a chordal graph with n vertices and m edges. Suppose that each vertex of G has a nonnegative *weight*, and that the sum of the weights is n. The main result of this section is that there is a clique that divides the weight roughly in half.

THEOREM 1. *Let G be a weighted chordal graph as above, with p vertices in its largest clique. Then G contains a clique whose removal leaves no connected component of weight more than $n/2$. Unless $n = 1$, the clique can be chosen to have at most $p-1$ vertices.*

Remark. This theorem resembles Lemma 6 above, but seems not to follow from it. Let us call the separator in the statement of Theorem 1 an *$n/2$ separator*. Then a minimal $n/2$ separator need not be a clique; for example, if G is a path with 5 vertices, then one minimal $n/2$ separator is the second and fourth vertices. Also, there need not be a minimal v, w separator that is an $n/2$ separator; for example, if G is an $n/2$-vertex clique with an additional vertex of degree one adjacent to each clique vertex, then the only minimal v, w separators are single clique vertices.

Proof. The idea of the proof is to start with an arbitrary clique and make it ooze around the graph like an amoeba until it is an $n/2$ separator. It oozes by disgorging vertices that can join or become components of weight less than $n/2$, and by engulfing vertices that are in a component of weight more than $n/2$.

Here are the details. We will not distinguish between a set of vertices of G and the subgraph of G it induces. Unless G is empty, it has at least one clique. Let C be the clique that minimizes the maximum weight of a connected component of $G-C$. In case of ties, minimize the number of vertices in a maximum-weight component of $G-C$. If ties remain, minimize the number of vertices in C. If ties still remain, choose arbitrarily.

Assume for the sake of contradiction that $G-C$ has a component A of weight greater than $n/2$. Then the total weight of $G-A$ is less than $n/2$. We shall state and prove three facts about A and C.

FACT 1. *Every vertex of C is adjacent to some vertex of A.*

Proof. If $v \in C$ were not adjacent to any vertex of A, then $C-\{v\}$ would have been chosen in preference to C.

FACT 2. *If B is a nonempty subset of A, then B contains a vertex that is simplicial in $B \cup C$.*

Proof. Immediate from Lemma 2.

FACT 3. *Component A contains a vertex v adjacent to every vertex of C.*

Proof. We will take v to be the last vertex of A in a perfect elimination ordering of $A \cup C$ with C ordered last. Thus $v = a_k$ where $\{a_1, \cdots, a_k, c_1, \cdots, c_h\}$ is a perfect elimination ordering of $A \cup C$. Such an ordering exists because by Fact 2 we can repeatedly choose simplicial vertices that are not in the clique C.

Let x be a vertex of C. Since A is connected and (by Fact 1) x is adjacent to a vertex of A, there is a path from x to a_k in $A \cup C$ that uses only vertices of $A-\{a_k\}=\{a_1, \cdots, a_{k-1}\}$. Lemma 3 says that if $\{x, a_k\}$ is not an edge of $A \cup C$, then it is a fill-in edge. But a perfect elimination ordering has no fill-in, so x is adjacent to a_k in $A \cup C$ and in G. Thus a_k is adjacent to every vertex of C, so we can take $v = a_k$.

Fact 3 leads to a contradiction: $C \cup \{v\}$ is a clique, and it should have been chosen in preference to C. Thus C is the desired $n/2$ separator.

The argument above shows that each component of $G-C$ contains a vertex adjacent to all of C's vertices. If G is not complete then C is not the largest clique in G, and C has at most $p-1$ vertices. If G is complete we can take C to be all of G except the lightest vertex. □

COROLLARY 1. *Let G be a chordal graph with n vertices and m edges. Suppose G's vertices have nonnegative weights that add up to n. Then G has a set of $O(\sqrt{m})$ vertices whose removal leaves no connected component of weight more than $n/2$.*

Proof. Theorem 1 says that G has a clique that separates the graph as required. This clique has at most m edges and hence only $O(\sqrt{m})$ vertices. □

COROLLARY 2. *Let G be a chordal graph with n vertices and m edges. Then G has a set of $O(\sqrt{m})$ vertices whose removal leaves no connected component with more than $n/2$ vertices.* □

A *k-tree* [17] is a graph constructed by starting with a k-vertex clique and adding vertices one at a time, making each new vertex adjacent to k mutually adjacent old vertices. Thus a 1-tree is a tree. A k-tree is chordal, and its largest clique has $k+1$ vertices unless the k-tree is a k-clique. Therefore k-trees have separators whose size is independent of the size of the tree.

COROLLARY 3. *Let T be a k-tree whose vertices have nonnegative weights that add up to n. Then T has a set of k vertices whose removal leaves no connected component of weight more than $n/2$.* □

4. An $O(m)$ algorithm. Throughout this section G will be a connected chordal graph with n vertices having nonnegative weights that add up to n, and with m edges. We shall present an algorithm to find the separator of Theorem 1 in $O(m)$ time.

The proof of Theorem 1 leads directly to the following algorithm.

```
procedure SLOW CHORDAL SEPARATOR (graph $G$);
  begin
  $C \leftarrow \{\}$;
  while some component $A$ of $G-C$ has weight more than $n/2$ do
    while some vertex $x$ of $C$ is adjacent to no vertex of $A$ do
      $C \leftarrow C - \{x\}$
      od;
    $v \leftarrow$ some vertex of $A$ adjacent to every vertex of $C$;
    $C \leftarrow C \cup \{v\}$
    od;
  return $C$
  end
```

Since a vertex is added to C at most once, the main loop is executed at most n times. The whole algorithm is easily implemented to run in $O(mn)$ time. To speed it up, we need to make four observations.

First, a close examination of the proof of Theorem 1 reveals that vertices are added to C in the reverse of a perfect elimination order. We can save searching by precomputing this order.

Second, we do not really need to delete vertices from C as we go along. We can just keep adding vertices to C until no component of $G-C$ weighs more than $n/2$; it then turns out that the desired separator is the last vertex added to C, plus all vertices of C adjacent to that vertex. This observation is formalized and proved in Lemma 8 below.

Third, we can find "the last vertex added to C" by looking at the vertices in the opposite order, that is, in perfect elimination order. We simply start with an empty graph H and add vertices of G to it in perfect elimination order until some connected component of H weighs more than $n/2$. Then the last (highest-numbered) vertex added to H is the same as the last (lowest-numbered) vertex added to C by the slow algorithm.

Finally, we do not have to represent the connected components of H explicitly. The fast algorithm maintains an array $w(1..n)$ indexed by vertex number. When the ith vertex is added to H, the array entry $w(i)$ is the weight of the connected component containing that vertex. Because G is chordal, it turns out to be possible to maintain w by doing only one operation per vertex. This observation is formalized and proved in Lemma 7 below.

```
procedure FAST CHORDAL SEPARATOR (graph $G$);
  begin
  real array $w(1\cdot\cdot n)$;
  find a perfect elimination ordering $\{v_1, \cdots, v_n\}$ of $G$;
  for $i \leftarrow 1$ to $n$ do $w(i) \leftarrow$ weight of $v_i$ od;
  $i \leftarrow 1$;
  while $w(i) \leqq n/2$ do
    comment $w(i)$ is the weight of the connected component
      of $\{v_1, \cdots, v_i\}$ that contains $v_i$;
    $v_k \leftarrow$ lowest-numbered neighbor of $v_i$ with $k > i$;
    $w(k) \leftarrow w(k) + w(i)$;
    $i \leftarrow i + 1$
    od;
  comment $i$ is minimum such that some component
    of $\{v_1, \cdots, v_i\}$ weighs more than $n/2$;
  $C \leftarrow v_i$ plus all of $v_{i+1}, \cdots, v_n$ adjacent to $v_i$;
  return $C$
  end
```

We will now prove that this algorithm correctly finds the separator C mentioned in Theorem 1. This requires two lemmas.

LEMMA 7. *Consider the subgraph of G induced by $v_1, \cdots, v_i$. At the beginning of the ith iteration of the* **while** *loop, $w(i)$ is the weight of the connected component of this subgraph that includes v_i. (That is, the comment in the* **while** *loop is correct.)*

Proof. The lemma is immediate for $i = 1$. Suppose it is true for $1, 2, \cdots, i-1$, and consider the situation at the beginning of the ith iteration of the loop.

Originally $w(i)$ was the weight of v_i, and something was added to $w(i)$ during the kth iteration if and only if $k < i$, v_k is adjacent to v_i, and v_k is adjacent to no v_h with $k < h < i$. We will show that each connected component of $\{v_1, \cdots, v_{i-1}\}$ containing a vertex adjacent to v_i had its weight added to $w(i)$ exactly once.

Let A be a connected component of $\{v_1, \cdots, v_{i-1}\}$. If A has no vertex adjacent to v_i, nothing was added to $w(i)$ during an iteration for any vertex in A.

Suppose A has a vertex v_k adjacent to v_i, and suppose v_h is the highest-numbered vertex of A. Thus $k \leqq h < i$. There is a path from v_h to v_i through vertices numbered no higher than i, so by Lemma 4 there is a monotone path from v_h to v_i. Since v_h is the highest-numbered vertex in A, the monotone path must be a single edge. Therefore v_h is adjacent to v_i and, by the inductive hypothesis, the weight of A was added to $w(i)$ during the hth iteration. If $k < h$ then, again by Lemma 4, there is a monotone path from v_k to v_h. This implies that v_k is adjacent to a vertex numbered higher than k but lower than i, so nothing was added to $w(i)$ during the kth iteration.

This shows that each connected component of $\{v_1, \cdots, v_{i-1}\}$ adjacent to v_i has its weight added to $w(i)$, and nothing else is ever added to $w(i)$. Thus $w(i)$ is the weight of the connected component of $\{v_1, \cdots, v_i\}$ that contains v_i. □

LEMMA 8. *Consider the smallest i such that some component of* $\{v_1, \cdots, v_i\}$ *weighs more than* $n/2$. *Then* $C=\{v_i\}\cup\{v_k\colon k>i$ *and* v_k *is adjacent to* $v_i\}$ *is a clique whose removal from G leaves no component that weighs more than* $n/2$.

Proof. Define the *boundary* of a set A of vertices, written ∂A, to be the set of vertices of $G-A$ that are adjacent to vertices of A. Let i be as in the statement of the lemma, and let A_i be the heaviest component of $\{v_1, \cdots, v_i\}$.

First, C is a clique, because v_i is simplicial in $\{v_{i+1}, \cdots, v_n\}$ by the definition of a perfect elimination ordering.

Second, v_i must be in A_i, because A_i weighs more than $n/2$ but no component of $\{v_1, \cdots, v_{i-1}\}$ weighs more than $n/2$. Then v_i is the last vertex of A_i in the perfect elimination ordering, so the same argument as the proof of Fact 3 in Theorem 1 shows that v_i is adjacent to every vertex in ∂A_i. Therefore C is $\partial A_i \cup \{v_i\}$. Since A_i is the only component of $G-\partial A_i$ that weighs more than $n/2$, $\partial A_i \cup \{v_i\}$ is a set whose deletion from G leaves no component that weighs more than $n/2$. □

The correctness of the algorithm follows immediately from these two lemmas: Lemma 7 implies that the **while** loop finds the smallest i such that some component of $v_1, \cdots, v_i$ weighs more than $n/2$, and then Lemma 8 says that the algorithm finds the desired separator.

It remains only to analyze the algorithm's running time. Finding a perfect elimination ordering takes $O(m)$ time by an algorithm of Rose, Tarjan, and Lueker [19]. The **while** loop is executed at most n times. Finding the lowest-numbered neighbor of v_i means looking at all its adjacent vertices, for a total of $O(m)$ over all vertices. Finally, the computation of C takes $O(m)$ time. Thus the total running time is $O(m)$.

5. More separator theorems.

We can generalize the separator theorem in § 3 to find separators that chop a chordal graph into fragments no larger than a specified weight, or to separate a chordal graph according to two or more unrelated sets of weights simultaneously.

These generalizations are suggested by similar results on planar graphs. Lipton, Tarjan, and Gilbert [14, Thm. 2], [5, Thm. 1.3.2] used the $\sqrt{n}$-vertex separator theorem for planar graphs quoted in § 1 to derive analogues of the following two theorems for planar graphs. (In the planar theorems the sizes depend only on n, the number of vertices, and not on m, the number of edges.) The chordal versions follow from our Theorem 1 by the same methods, so we omit the proofs.

THEOREM 2. *Let G be a chordal graph with n vertices and m edges, with nonnegative vertex weights that add up to* 1. *Let* $\varepsilon>0$ *be given. Then there is a set of* $\sqrt{m/\varepsilon}$ *vertices of G whose removal leaves no component with weight more than* ε. *The separator can be found in* $O(m \log n)$ *time.* □

THEOREM 3. *Fix* $\varepsilon>0$. *Let G be a chordal graph with n vertices and m edges, with two sets of vertex weights that both add up to n. Then the vertices of G can be partitioned into sets A, B, and C such that C separates A from B, C has* $O(\sqrt{m})$ *vertices, neither A nor B has weight of the first kind more than* $n/2$, *and neither A nor B has weight of the second kind more than* $(\frac{1}{2}+\varepsilon)n$. *The separator can be found in* $O(m)$ *time.* □

For example, we can take one kind of weight to be 1 for every vertex, and the other kind to be proportional to the vertex degree. Then Theorem 3 allows us to divide a chordal graph into two pieces each with at most $n/2$ vertices and at most $(\frac{1}{2}+\varepsilon)m$ edges.

Theorem 3 can be applied recursively to obtain a $\sqrt{m}$-vertex separator theorem for any fixed number of sets of vertex weights. In general the separator divides the graph into two pieces, each with at most half the first kind of weight and at most $\frac{1}{2}+\varepsilon$ of each of the other kinds of weight. This quickly ceases to be practical, because the constant factor grows double-exponentially with the number of kinds of weight.

6. Remarks. The algorithm in § 4 finds a separator in $O(m)$ time. No algorithm faster than $O(m)$ is possible if the input graph is represented by listing its edges or by listing the vertices adjacent to each vertex. However, the graph might be more compactly represented. For example, the chordal graph corresponding to an acyclic hypergraph (as described below) is given as a union of cliques. The sum of the sizes of the cliques can be much less than m. It might be possible to find a separator in time linear in this sum.

Chordal graphs have applications in a number of areas, and we expect the separator theorems above to be useful in some of them. One such area is solving sparse linear systems by Gaussian elimination, where one wishes to find an elimination ordering for a sparse matrix that causes relatively few zeros to become nonzero. If the matrix is symmetric and only symmetric permutations are allowed, this corresponds to finding a small set of edges whose addition makes a graph chordal. Finding the smallest such set of edges is an NP-complete problem [20], but there are heuristics that perform well in many cases. One heuristic, called nested dissection, uses separators in planar graphs to give good orderings for systems that come from differential equations on two-dimensional regions [13]. We are investigating the use of the chordal separator theorem to show that any graph has some nested dissection ordering that is close to optimum.

Chordal graphs and their separators also appear in database theory, and we believe that our results might be applicable there. A database represents a relation (called a *universal relation*) on a set of attributes. The universal relation can be represented implicitly by storing its projections on some subsets of the attributes. Consider the hypergraph whose vertices are attributes and whose hyperedges are the subsets whose projections are explicitly stored. A separator in this hypergraph implies an association (technically, a *multivalued dependency*) between each component and the separator. This association often corresponds to some relationship in the real world among the attributes involved.

The hypergraphs of database schemes from the real world are nearly all *acyclic* [1]. Acyclic database schemes have a number of desirable properties; roughly, in an acyclic scheme pairwise consistency between the projections implies that the universal relation is consistent. A hypergraph is acyclic if and only if the graph formed by replacing each hyperedge with a clique is chordal.

Acknowledgment. We thank Dale Skeen for several stimulating discussions, and particularly for helping us understand acyclic database schemes.

REFERENCES

[1] CATRIEL BEERI, RONALD FAGIN, DAVID MAIER, ALBERTO MENDELZON, JEFFREY ULLMAN AND MIHALIS YANNAKAKIS, *Properties of acyclic database schemes*, Proc. 13th Annual ACM Symposium on Theory of Computing, pp. 355–362, 1981.

[2] G. A. DIRAC, *On rigid circuit graphs*, Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg, 25 (1961), pp. 71–76.

[3] HRISTO NICOLOV DJIDJEV, *On the problem of partitioning planar graphs*, this Journal, 3 (1982), pp. 229–240.

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

[5] JOHN RUSSELL GILBERT, *Graph separator theorems and sparse Gaussian elimination*, Ph.D. thesis, Stanford University, Stanford, CA, 1980.

[6] JOHN R. GILBERT, JOAN P. HUTCHINSON AND ROBERT ENDRE TARJAN, *A separator theorem for graphs of bounded genus*, Cornell Univ. Dept. of Computer Science technical report 82-506, 1982, J. Algorithms, to appear.

[7] ANDRÁS HAJNAL AND JÁNOS SURÁNYI, *Über die Auflösung von Graphen in vollständige Teilgraphen*, Annales Universitatis Scientarium Budapest, Sectio Mathematica 1 (1958), pp. 113–121.

[8] CAMILLE JORDAN, *Sur les assemblages de lignes*, Journal Reine Angew. Math., 70 (1869), pp. 185–190.

[9] DAN HOEY AND CHARLES E. LEISERSON, *A layout for the shuffle-exchange network*, Carnegie–Mellon Univ. Department of Computer Science technical report CMU-CS-80-139, Pittsburgh, PA, 1980.

[10] FRANK THOMSON LEIGHTON, *New lower bound techniques for* VLSI, Proc. 22nd Annual IEEE Symposium on Foundations of Computer Science, (1981), pp. 1–12.

[11] CHARLES E. LEISERSON, *Area-efficient graph layouts* (*for* VLSI), Proc. 21st Annual IEEE Symposium on Foundations of Computer Science, (1980), pp. 270–281.

[12] P. M. LEWIS II, R. E. STEARNS AND J. HARTMANIS, *Memory bounds for the recognition of context-free and context-sensitive languages*, IEEE Conference Record on Switching Theory and Logical Design, (1965), pp. 191–202.

[13] RICHARD J. LIPTON, DONALD J. ROSE AND ROBERT ENDRE TARJAN, *Generalized nested dissection*, SIAM J. Numer. Anal., 16 (1979) 346–358.

[14] RICHARD J. LIPTON AND ROBERT ENDRE TARJAN, *Applications of a planar separator theorem*, SIAM J. Comput. 9 (1980), pp. 615–627.

[15] ———, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–189.

[16] DONALD J. ROSE, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*. In Graph Theory and Computing, R. C. Read, ed. Academic Press, 1972, pp. 183–217.

[17] ———, *On simple characterizations of k-trees*, Discrete Math., 7 (1974), pp. 317–322.

[18] ———, *Triangulated graphs and the elimination process*, J. Math. Anal. Appl., 32 (1970), pp. 597–609.

[19] DONALD J. ROSE, R. ENDRE TARJAN AND GEORGE S. LUEKER, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

[20] MIHALIS YANNAKAKIS, *Computing the minimum fill-in is* NP-*complete*, this Journal, 2 (1981), pp. 77–79.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

004

STABILITY OF BLOCK LU-DECOMPOSITIONS OF MATRICES ARISING FROM BVP*

R. M. M. MATTHEIJ†

Abstract. An analysis is made of the stability of block LU-decompositions of matrices arising from boundary value problems of ODE. It is based on an investigation of the growth properties of the related recursion (or ODE) solution spaces. It is shown how blocks in the upper right corner or the lower left corner of the matrix may generate blocks in the decomposition that exhibit a growth like some of these solutions, unstable ones not excluded. In particular, for partially separated boundary conditions the desire to reduce memory space may thus conflict with that for actual stability of this decomposition.

AMS(MOS) subject classifications. 65F05, 65L10

1. Introduction. An important part in the solution of boundary value problems (BVP) is played by the question of how the linear systems that arise from discretization and/or linearization have to be solved. Since these systems usually involve sparse matrices, much of the attention has been focused on efficient and memory space saving techniques. A particularly interesting class of methods is based on an appropriate rearrangement of the rows of this matrix after which a block LU-decomposition is performed in which pivoting is restricted somehow. These methods do quite well in the case of so-called separated boundary conditions, where the initial conditions make up for the first rows of the matrix and the terminal conditions for the last rows. For some examples; see e.g. [1], [2], [3], [4], [12], [13]. In particular, if we are dealing with a method that can be thought of as a one step recursion, the system matrix can be repartitioned as a block tridiagonal system. In a previous paper [10], we showed that in the last case a block LU-decomposition, where the L and the U are block bidiagonal matrices, is stable, if it exists. It is a natural question to ask if a similar nice result can be established for more general boundary conditions (BC), where the matrix has a more complicated structure. An obvious generalization seems to be given by so called partially separated boundary conditions, cf. [6, pp. 2ff], where an appropriate rearrangement and a block partitioning give rise to a system which is block tridiagonal and moreover has a nonzero block in the upper right or lower left corner. The advantage of such an approach is obvious: by this we save memory space since either the upper triangular matrix or the lower triangular matrix then is again block bidiagonal. Unfortunately, however, it will turn out that these savings may have to be paid for by lack of stability. Indeed we shall see that the proper rearrangement of the BC in the system is not a matter of placement of more or less accidental zero rows in them, but rather is dictated by inherent growth properties of the solutions of the difference and usually also of the differential equations. In the separated case the nonzero rows in either the initial or the terminal condition, in fact take care of the decreasing or increasing modes, respectively; this is why utilization of zero rows to minimize the storage problem is tightly connected with the proper rearrangement and partitioning in order to have stability there.

* Received by the editors September 21, 1982, and in revised form May 20, 1983. This research was supported in part by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.) and the Niels Stensen Stochting.

† Mathematisch Instituut, Katholieke Universiteit, Toernooiveld, 6525 ED Nijmegen, The Netherlands. This work was written while the author was visiting the Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York. This paper was presented at the SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 26–29, 1982.

This paper is built up as follows. First, in § 2 we give a number of notational conventions that will be used in this paper. Then in § 3 we derive the BVP matrices and indicate how we may find an LU-decomposition. In § 4 we show how important it may be to have a proper splitting of the BC. Section 5 describes a special LU-decomposition in which the incremental matrices of the one step recursion are transformed onto upper triangular form; the L and U of this representation can be found fairly easily and will be used for estimation purposes. In § 6 we derive some useful expressions for blocks of the inverse of the system matrix; this is used in § 7, where we give the actual stability estimates and the conclusions based on them. Finally, in § 8, we derive a strategy for determining a proper splitting of the BC in practice.

2. Notational conventions. In the sequel we use some notation, which is summarized here for convenience.

2.1. Partitioning of matrices. Open letters like $\mathbb{A}$ denote $Nn \times Nn$ matrices. For $n \times n$ matrices A we use superscripts to denote a partitioning like

$$A = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}, \tag{2.1}$$

where A^{11} is a $k \times k$ matrix (k a generic constant). A single right superscript then denotes the corresponding column partitioning

$$A = (A^1 \mid A^2) \tag{2.2}$$

(i.e., A^1 is an $n \times k$ matrix).

A single left superscript denotes the corresponding row partitioning

$$A = \begin{bmatrix} {}^1A \\ \hdashline {}^2A \end{bmatrix} \tag{2.3}$$

(i.e., ${}^1A$ is a $k \times n$ matrix).

Wherever necessary we shall provide an identity matrix with an index in order to indicate its order, e.g., I_k. For all other matrices such an index will never refer to order.

2.2. Norms. The norm $\|\cdot\|$ is the 2-Hölder norm. For the associated matrix norm

$$\|A^{ij}\| \leqq \|A\|, \quad \|A^i\| \leqq \|A\|, \quad \|{}^iA\| \leqq \|A\|. \tag{2.4}$$

2.3. Products and sums of matrices. We define

$$\prod_{j=l}^{m} A_j = \begin{cases} A_m \cdots A_l & \text{if } m \geqq l, \\ I & \text{if } m < l, \end{cases} \tag{2.5}$$

$$\sum_{j=l}^{m} A_j = \begin{cases} A_m + \cdots + A_l & \text{if } m \geqq l, \\ 0 & \text{if } m < l. \end{cases} \tag{2.6}$$

3. Block matrices arising from one step recursions. Quite a few discretization methods for boundary value problems lead to a one step recursion, which, together with the BC, give rise to a sparse linear system for the solution values at a certain grid. Such methods are multiple shooting, where the grid consists of the shooting points, one step difference methods, where the grid equals the grid of discretization points, or collocation, where the grid is formed by the endpoints of the collocation intervals (cf. [1], [5], [6], [11], [12]). Suppose the desired solution values are denoted by $x_1, \cdots, x_N$, then in its most simple form this recursion reads

$$x_{i+1} = G_i x_i - c_i, \qquad 1 \leqq i \leqq N-1. \tag{3.1}$$

Here the G_i are $n\times n$ matrices and the x_i and c_i are n-vectors. For such problems one could just as well write down a similar recursive relation for decreasing index. Therefore, we shall assume that the G_i are nonsingular. Although we often have a recursion like $F_ix_{i+1}=G_ix_i+d_i$, where F_i is nonsingular, we restrict ourselves to the form (3.1) to avoid complications in the notations later on. The more general case can be deduced fairly straightforwardly from the results of the present discussion.

A general BC can be written as

$$M_1x_1+M_Nx_N=b, \tag{3.2}$$

where M_1 and M_N are $n\times n$ matrices and b is an n-vector.

Assumption 3.3. Let M_1 and M_N be normalized such that $\max(\|M_1\|, \|M_N\|)=1$.

The relations (3.1) and (3.2) lead to a linear system

$$\mathbb{A}\mathbf{x}=\mathbf{b}, \tag{3.4}$$

where

$$\mathbb{A}=\begin{bmatrix} G_1 & -I & 0 & & \\ & G_2 & -I & & \\ 0 & & \ddots & \ddots & \\ & & & G_{N-1} & -I \\ M_1 & & & & M_N \end{bmatrix}, \quad \mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \quad \mathbf{b}=\begin{bmatrix} c_1 \\ \vdots \\ c_{N-1} \\ b \end{bmatrix}. \tag{3.5}$$

Of course we may associate many other linear systems with (3.1) and (3.2). As we shall see later it might be advisable to split the BC into parts and write them as first and last equations in the system in order to assure the stability of block LU-decompositions. Such a splitting can be described by premultiplying $\mathbb{A}$ by an appropriate permutation matrix $\mathbb{P}$. For notational reasons we adopt the convention that the last k rows are written first. (In § 5 this will appear to lead a more natural notation of the blocks.)

We can then write $\mathbb{P}$ as

$$\mathbb{P}=\left[\begin{array}{cccc|c} & & & & I_{n-k} \\ \hline I_n & & 0 & & \\ & \ddots & & & \\ 0 & & I_n & & \\ & & & I_k & \end{array}\right]\cdot\begin{bmatrix} I_n & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & I_n & \\ & & & P \end{bmatrix}, \tag{3.6}$$

where P is a permutation matrix of order n. The resulting matrix $\mathbb{P}\mathbb{A}$ then reads

$$\bar{\mathbb{A}}:=\mathbb{P}\mathbb{A}=\begin{bmatrix} {}^2\bar{M}_1 & & & & {}^2\bar{M}_N \\ G_1 & -I & & 0 & \\ & G_2 & -I & & \\ & & \ddots & \ddots & \\ & 0 & & G_{N-1} & -I \\ {}^1\bar{M}_1 & & & & {}^1\bar{M}_N \end{bmatrix}, \text{ where } \bar{M}_j=PM_j, j=1,2. \tag{3.7}$$

By repartitioning $\bar{\mathbb{A}}$ into a matrix of $n\times n$ blocks we obtain an almost tridiagonal block matrix. We shall use the notation

$$\bar{\mathbb{A}} = \begin{bmatrix} B_1 & C_1 & & & H_N \\ A_2 & B_2 & C_2 & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & A_{N-1} & B_{N-1} & C_{N-1} \\ H_1 & & & A_N & B_N \end{bmatrix}. \tag{3.8}$$

Apparently the C_i systematically have zeros in the first $n-k$ rows and, similarly, the A_i have zeros in the last k rows.

The aim of regrouping the BC and repartitioning it into the form (3.8), is that we would like to obtain a representation that allows for a stable block LU-decomposition, which preserves as much of the sparsity structure as possible. It is not restrictive to require then that the lower triangular matrix has identity diagonal blocks. Hence we look for the decomposition

$$\bar{\mathbb{A}} = \mathbb{L}\mathbb{U}, \tag{3.9}$$

where

$$\mathbb{L} = \begin{bmatrix} I & & & & \\ L_2 & I & & 0 & \\ & \ddots & \ddots & & \\ & 0 & L_{N-1} & I & \\ S_1 & \cdots & S_{N-2} & S_{N-1} & I \end{bmatrix} \tag{3.10a}$$

and

$$\mathbb{U} = \begin{bmatrix} U_1 & C_1 & & & T_1 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & C_{N-2} & T_{N-2} \\ & 0 & & U_{N-1} & T_{N-1} \\ & & & & U_N \end{bmatrix}. \tag{3.10b}$$

PROPERTY 3.11. *The L_i, S_i and T_i have the same systematical zero rows as G_1, H_1 and H_N, respectively.*

If we premultiply $\mathbb{A}$ by a block diagonal matrix $\hat{\mathbb{D}}$,

$$\hat{\mathbb{D}} = \operatorname{diag}(\hat{D}_1, \cdots, \hat{D}_{N-1}, I). \tag{3.12}$$

where the $\hat{D}_j$ are nth order nonsingular matrices, and premultiply this by a matrix $\hat{\mathbb{P}}$ as in (3.6) with $\hat{P}$ instead of P, and moreover if we postmultiply $\mathbb{A}$ by $\hat{\mathbb{E}}$

$$\hat{\mathbb{E}} = \operatorname{diag}(\hat{E}_1, \cdots, \hat{E}_N), \tag{3.13}$$

where the $\hat{E}_j$ are nth order nonsingular matrices, then it can easily be seen that the matrix $\hat{\mathbb{A}}$, defined by

$$\hat{\mathbb{A}} = \hat{\mathbb{P}}\hat{\mathbb{D}}\mathbb{A}\hat{\mathbb{E}} \tag{3.14}$$

systematically has zero rows in the same blocks as $\bar{\mathbb{A}}$ has. In this case a pivoting strategy for a block LU-decomposition, as in [5], can be described by such $\mathbb{P}$, $\mathbb{D}$ and $\mathbb{E}$, where these matrices are suitable permutation matrices (depending on the strategy and the problem). Partitioning $\hat{\mathbb{A}}$ as in (3.8) and using a similar notation for the blocks by providing them with a cap, we obtain, from the decomposition,

$$\hat{\mathbb{A}} = \hat{\mathbb{L}}\hat{\mathbb{U}}, \tag{3.15}$$

the following relations for the blocks of $\hat{\mathbb{L}}$ and $\hat{\mathbb{U}}$:

$$\hat{L}_i = \hat{A}_i \hat{U}_{i-1}^{-1}, \qquad i = 2, \cdots, N-1; \tag{3.16}$$

$$\text{(3.17)}\quad \begin{array}{ll} \text{(a)} & \hat{U}_1 = \hat{B}_1, \\ \text{(b)} & \hat{U}_i = \hat{B}_i - \hat{A}_i \hat{U}_{i-1}^{-1} \hat{C}_{i-1}, \qquad i = 2, \cdots, N-1, \\ \text{(c)} & \hat{U}_N = \hat{B}_N - \sum_{j=1}^{N-1} \hat{S}_j \hat{T}_j; \end{array}$$

$$\text{(3.18)}\quad \begin{array}{ll} \text{(a)} & \hat{S}_1 = \hat{H}_1 \hat{U}_1^{-1}, \\ \text{(b)} & \hat{S}_i = -\hat{S}_{i-1} \hat{C}_{i-1} \hat{U}_i^{-1}, \qquad i = 2, \cdots, N-2, \\ \text{(c)} & \hat{S}_{N-1} = [-\hat{S}_{N-2} \hat{C}_{N-2} + \hat{A}_N] \hat{U}_{N-1}^{-1}; \end{array}$$

$$\text{(3.19)}\quad \begin{array}{ll} \text{(a)} & \hat{T}_1 = \hat{H}_N, \\ \text{(b)} & \hat{T}_i = -\hat{L}_i \hat{T}_{i-1}, \qquad i = 2, \cdots, N-2, \\ \text{(c)} & \hat{T}_{N-1} = -\hat{L}_{N-1} \hat{T}_{N-1} + \hat{C}_{N-1}. \end{array}$$

We are interested in bounds for $\|\hat{\mathbb{L}}^{-1}\|$ and $\|\hat{\mathbb{U}}^{-1}\|$, since they appear as stability constants in error analyses of the solution of (3.4), cf. [14]. We shall use the following

PROPERTY 3.20. *Let* $\|\hat{\mathbb{A}}^{-1}\| \leqq \kappa$. *If* $\|\hat{\mathbb{L}}\| \leqq \mu_1$ *then* $\|\hat{\mathbb{U}}^{-1}\| \leqq \kappa\mu_1$. *If* $\|\hat{\mathbb{U}}\| \leqq \mu_2$ *then* $\|\hat{\mathbb{L}}^{-1}\| \leqq \kappa\mu_2$.

Hence, if κ is not large and in addition μ_1 and μ_2 are not large, we may call the LU-decomposition *stable.*

4. On proper splittings of the BC.

An important question in forming the system matrix $\bar{\mathbb{A}}$ is how many and which rows of the BC are used to make up for the first k rows in this matrix. Since B_1 (or $\hat{B}_1$) will act as the first pivotal block it is obvious that ${}^2\bar{M}_1$ has to have full rank. Since we can always premultiply the BC by a nonsingular matrix P say, such that, if rank $M_1 = l < n$, PM_1 has zeros in its first $(n-l)$ rows, it is not restrictive to assume the following.

Assumption 4.1 If rank $M_1 = l < n$, then the first $(n-l)$ rows of M_1 consist of zeros.

If the BC are *separated*, i.e., if the last l rows of M_N only contain zeros, then it is natural to try to utilize this fact by taking $k = n - l$, i.e., $\bar{\mathbb{A}}$ has a block tridiagonal form. The stability of this case has been investigated in [10], [12]. It is also most tempting to similarly utilize zero rows in so-called *partially separated* BC (cf. [6]), where nonzero rows in the one matrix do not necessarily correspond to zero rows in the other. If $l < n$ and we again take $k = n - l$ then it follows that the S_i in $\mathbb{L}$ are zero. (In a similar way we may utilize zeros in ${}^2M_N$ to obtain an LU-decomposition, where all T_i, $i \leqq N-2$ are zero.) However, the stability of such an LU-decomposition is no longer assured in general. To show this we consider the following example.

Example 4.2. Let $\{x_i\}$ satisfy the recursion

$$x_{i+1}=\begin{bmatrix}\frac{1}{2}&1&0\\0&3&0\\0&0&\frac{1}{4}\end{bmatrix}x_i+\begin{bmatrix}-\frac{1}{2}\\-2\\\frac{3}{4}\end{bmatrix},\qquad 1\leqq i\leqq N-1 \tag{4.3}$$

and the BC

$$\begin{bmatrix}1&0&0\\0&1&0\\0&0&-1\end{bmatrix}x_1+\begin{bmatrix}1&0&0\\0&1&0\\0&0&0\end{bmatrix}x_N=\begin{bmatrix}2\\2\\-1\end{bmatrix}. \tag{4.4}$$

It is simple to check that

$$\forall_i \quad x_i=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Apparently (4.4) is partially separated. Utilizing the zero row in M_N we therefore may choose for $\hat{\mathbb{A}}=\bar{\mathbb{A}}$

$$\hat{\mathbb{A}}=\left[\begin{array}{cccccc}
\begin{matrix}0&0&-1\\\frac{1}{2}&1&0\\0&3&0\end{matrix} & \begin{matrix}0&0&0\\-1&0&0\\0&-1&0\end{matrix} & & & 0 & \\
\begin{matrix}0&0&\frac{1}{4}\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}0&0&-1\\\frac{1}{2}&1&0\\0&3&0\end{matrix} & \begin{matrix}0&0&0\\-1&0&0\\0&-1&0\end{matrix} & & & \\
& \ddots & \ddots & \ddots & & \\
0 & & & \begin{matrix}0&0&\frac{1}{4}\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}0&0&-1\\\frac{1}{2}&1&0\\0&3&0\end{matrix} & \begin{matrix}0&0&0\\-1&0&0\\0&-1&0\end{matrix} \\
\begin{matrix}0&0&0\\1&0&0\\0&1&0\end{matrix} & & & & \begin{matrix}0&0&\frac{1}{4}\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}0&0&-1\\1&0&0\\0&1&0\end{matrix}
\end{array}\right]. \tag{4.5}$$

From the theory in [9] it already follows that (4.2) and (4.3) is a well-conditioned problem. This is also shown by Table 4.1.

TABLE 4.1

N	5	10	20	30
$\|\hat{A}^{-1}\|$	2.71	2.99	3.00	3.00

We obtain the following $\mathbb{L}$ and $\mathbb{U}$:

$$\mathbb{L}=\left[\begin{array}{ccccc}
\begin{array}{ccc}1&0&0\\0&1&0\\0&0&1\end{array} & & & & \\
\begin{array}{ccc}-\frac{1}{4}&0&0\\0&0&0\\0&0&0\end{array} & \begin{array}{ccc}1&0&0\\0&1&0\\0&0&1\end{array} & & & \\
 & \ddots & \ddots & & \\
 & & \begin{array}{ccc}-\frac{1}{4}&0&0\\0&0&0\\0&0&0\end{array} & \begin{array}{ccc}1&0&0\\0&1&0\\0&0&1\end{array} & \\
\begin{array}{cc}0&\begin{matrix}0&0\end{matrix}\\\begin{matrix}0\\0\end{matrix}&X\end{array} & & \begin{array}{cc}0&\begin{matrix}0&0\end{matrix}\\\begin{matrix}0\\0\end{matrix}&X^{N-2}\end{array} & \begin{array}{cc}-\frac{1}{4}&\begin{matrix}0&0\end{matrix}\\\begin{matrix}0\\0\end{matrix}&X^{N-1}\end{array} & \begin{array}{ccc}1&0&0\\0&1&0\\0&0&1\end{array}
\end{array}\right], \tag{4.6}$$

where

$$X=\begin{bmatrix}2&-\frac{2}{3}\\0&\frac{1}{3}\end{bmatrix}.$$

$$\mathbb{U}=\left[\begin{array}{cccc}
\begin{array}{ccc}0&0&-1\\\frac{1}{2}&1&0\\0&3&0\end{array} & \begin{array}{ccc}0&0&0\\-1&0&0\\0&-1&0\end{array} & & 0\\
 & \ddots & \ddots & \\
 & & \begin{array}{ccc}0&0&-1\\\frac{1}{2}&1&0\\0&3&0\end{array} & \begin{array}{ccc}0&0&0\\-1&0&0\\0&-1&0\end{array}\\
 & 0 & & \begin{array}{cc}\begin{matrix}0&0\end{matrix}&-1\\X^{N-1}&\begin{matrix}0\\0\end{matrix}\end{array}
\end{array}\right]. \tag{4.7}$$

We see that the last block row in $\mathbb{L}$ has elements of the order of 2^{N-1} and similarly U_N contains an element of this order. We therefore have to conclude that this splitting of the BC is not a desirable one. In Table 4.2 we have indicated the dramatic effects of this; "max error" denotes the maximal error (in norm) of the x_i, $i=1,\cdots,N$ (NB the machine constant is $\approx .2\ 10^{-16}$).

In all cases we obtained an accuracy of the order of the machine constant if we solved the system by a Crout routine.

We see from Table 4.2 that the error in the solution is almost proportional to 2^{N-1}. Obviously, this error growth factor is of the same order as $\|\mathbb{L}\|$ or $\|\mathbb{U}\|$ (both

TABLE 4.2

N	5	10	20	30
max error	.5 −15	.6 −13	.5 −10	.6 −7
2^{N-1}	.2 +2	.5 +3	.5 +6	.5 +9

about $\|X^{N-1}\|$). It seems plausible to relate this factor to the growth in the backward direction of the solution $(\frac{1}{2})^i(1,0,0)^T$ (which satisfies the homogeneous part of (4.3)). It is easily seen that backsolving via $\mathbb{U}$ allows errors to blow up by 2^{N-1} indeed. For reasons that will become more transparent in the next sections we better, e.g., choose $^2M_1 = (\begin{smallmatrix}0 & 0 & -1\\ 1 & 0 & 0\end{smallmatrix})$ and permute the first and second row in the A_i; what we can see already is that this would lead to pivotal blocks with no eigenvalues smaller than one. From this example we therefore conclude that we should look for splittings of the BC such that in the resulting block LU-decomposition the back substitution does not generate solutions that grow like the decaying solutions in the backward direction, and also such that the forward substitution via $\mathbb{L}$ does not generate solutions that grow like increasing solutions. One should realize that such a stability argument may be in conflict with memory space considerations.

5. A special decomposition. In order to investigate the stability of block LU-decompositions it is useful to compare such decompositions to a special one, resulting from a certain recursive transformation that brings the incremental matrices G_i onto upper triangular form (cf. [8]). The initialization of this transformation is based on the following:

PROPERTY 5.1. *Let M_1 be as in Assumption* 4.1. *Let P be a nonsingular matrix of which the lower right $l\times l$ block is also nonsingular. Then there exists an orthogonal matrix Q_1 such that $\tilde{M}_1 := PM_1Q_1$ is upper triangular. Moreover, if l $(=\operatorname{rank} M_1) < n$ then the first $(n-l)$ columns of $\tilde{M}_1$ are zero and the lower right $l\times l$ block of $\tilde{M}_1$ is nonsingular.*

Proof. Apparently the last l rows of PM_1 are linearly independent. By orthogonalization of these (cf. Gram–Schmidt) we find the last l columns of Q_1, after which Q_1 can be completed by using a basis for the orthoplement as first $(n-l)$ columns. □

The matrix P in Property 5.1 may, e.g., be the permutation matrix P in (3.6) or any other "rearranging" matrix. Now calculate orthogonal matrices $\{Q_i\}$ and upper triangular matrices $\{V_i\}$ which satisfy the relation

$$G_iQ_i = Q_{i+1}V_i, \qquad i=1,\cdots,N-1, \tag{5.2}$$

and finally a matrix $\tilde{M}_N$, defined by

$$\tilde{M}_N := PM_NQ_N. \tag{5.3}$$

Denote the global transformation matrices (cf. (3.12)) by

$$\tilde{\mathbb{D}} = \begin{bmatrix} Q_2^{-1} & & 0 & \\ & \ddots & & \\ 0 & & Q_N^{-1} & \\ & & & I \end{bmatrix}, \tag{5.4}$$

$$(5.5)\qquad \tilde{\mathbb{E}}=\begin{bmatrix} Q_1 & & 0 \\ & \ddots & \\ 0 & & Q_N \end{bmatrix}.$$

Finally define a permutation $\tilde{\mathbb{P}}$ as in (3.6). Then the matrix $\tilde{\mathbb{A}}$ defined by

$$(5.6)\qquad \tilde{\mathbb{A}}=\tilde{\mathbb{P}}\tilde{\mathbb{D}}\mathbb{A}\tilde{\mathbb{E}}$$

has the form (where the minus sign in the first row has been added for convenience)

$$(5.7)\quad \tilde{\mathbb{A}}:=\left[\begin{array}{cc|cc|c|cc|cc|cc}
0 & -\tilde{M}_1^{22} & 0 & 0 & & & & & & -\tilde{M}_N^{21} & -\tilde{M}_N^{22} \\
V_1^{11} & V_1^{12} & -I & 0 & & & & & & 0 & 0 \\
\hline
0 & V_1^{22} & 0 & -I & & & & & & & \\
0 & 0 & V_2^{11} & V_2^{12} & & & & & & & \\
\hline
 & & \ddots & & \ddots & & & & & & \\
\hline
 & & & & & 0 & V_{N-2}^{22} & 0 & -I & & \\
 & & & & & 0 & 0 & V_{N-1}^{11} & V_{N-1}^{12} & & \\
\hline
0 & 0 & & & & & & 0 & V_{N-1}^{22} & 0 & -I \\
\tilde{M}_1^{11} & \tilde{M}_1^{12} & & & & & & 0 & 0 & \tilde{M}_N^{11} & \tilde{M}_N^{12}
\end{array}\right].$$

We remark that if $k\leqq n-l$, then $\tilde{M}_1^{11}$ is zero (however note that $k\geqq n-l$ is necessary to make B_1 nonsingular). We now look for an LU-decomposition of $\tilde{A}$, viz.,

$$(5.8)\qquad \tilde{\mathbb{A}}=\tilde{\mathbb{L}}\tilde{\mathbb{U}}.$$

By our construction $\tilde{M}_1^{22}$ must be nonsingular; moreover nonsingularity of G_i implies nonsingularity of V_i and hence of V_i^{11}. It now turns out that all diagonal blocks in (5.7) are actually equal to the pivotal blocks, except for the last one. By straightforward calculations we obtain the following blocks of $\tilde{\mathbb{L}}$ and $\tilde{\mathbb{U}}$ (cf. [10]) (we use a similar notation similar to (3.16) ff, now with tildes)

$$(5.9)\qquad \begin{aligned} &\text{(a)}\quad \tilde{U}_1=\begin{bmatrix} 0 & -\tilde{M}_1^{22} \\ V_1^{11} & V_1^{12}\end{bmatrix}, \\ &\text{(b)}\quad \tilde{U}_i=\begin{bmatrix} 0 & -I \\ V_i^{11} & V_i^{12}\end{bmatrix}, \qquad i=2,\cdots,N-1; \end{aligned}$$

$$(5.10)\qquad \begin{aligned} &\text{(a)}\quad \tilde{L}_2=\begin{bmatrix} -V_1^{22}[\tilde{M}_1^{22}]^{-1} & 0 \\ 0 & 0\end{bmatrix}, \\ &\text{(b)}\quad \tilde{L}_i=\begin{bmatrix} -V_{i-1}^{22} & 0 \\ 0 & 0\end{bmatrix}, \qquad i=3,\cdots,N-1; \end{aligned}$$

$$\text{(a)}\qquad \tilde{S}_1=\left[\begin{array}{c|c} 0 & 0 \\ \hline [\tilde{M}_1^{11}[V_1^{11}]^{-1}V_1^{12}-\tilde{M}_1^{12}][\tilde{M}_1^{22}]^{-1} & \tilde{M}_1^{11}[V_1^{11}]^{-1} \end{array}\right],$$

$$(5.11)\quad\text{(b)}\qquad \tilde{S}_i=\left[\begin{array}{c|c} 0 & 0 \\ \hline \tilde{M}_1^{11}[\prod_{j=1}^{i} V_j^{11}]^{-1}V_i^{12} & \tilde{M}_1^{11}[\prod_{j=1}^{i} V_j^{11}]^{-1} \end{array}\right],$$

$$i=2,\cdots,N-2,$$

$$\text{(c)}\qquad \tilde{S}_{N-1}=\left[\begin{array}{c|c} -V_{N-1}^{22} & 0 \\ \hline \tilde{M}_1^{11}[\prod_{j=1}^{N-1} V_j^{11}]^{-1}V_{N-1}^{12} & \tilde{M}_1^{11}[\prod_{j=1}^{N-1} V_j^{11}]^{-1} \end{array}\right];$$

$$\text{(a)}\qquad \tilde{T}_1=\begin{bmatrix} -\tilde{M}_N^{21} & -\tilde{M}_N^{22} \\ 0 & 0 \end{bmatrix},$$

$$(5.12)\quad\text{(b)}\qquad \tilde{T}_i=\left[\begin{array}{c|c} \prod_{j=1}^{i-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21} & \prod_{j=1}^{i-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{22} \\ \hline 0 & 0 \end{array}\right],$$

$$i=2,\cdots,N-2,$$

$$\text{(c)}\qquad \tilde{T}_{N-1}=\left[\begin{array}{c|c} \prod_{j=1}^{N-2} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21} & \prod_{j=1}^{N-2} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{22} \\ \hline -I & 0 \end{array}\right].$$

Finally we obtain

(5.13)

$$\tilde{U}_N=\left[\begin{array}{c|c} \prod_{j=1}^{N-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21} & -I+\prod_{j=1}^{N-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21} \\ \hline \begin{array}{l}\tilde{M}_N^{11}+\tilde{M}_1^{12}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21} \\ \quad -\tilde{M}_1^{11}\{\Omega[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21}+[\prod_{j=1}^{N-1} V_j^{11}]^{-1}\}\end{array} & \begin{array}{l}\tilde{M}_N^{12}+\tilde{M}_1^{12}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{22} \\ \quad -\tilde{M}_1^{11}\Omega[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{22}\end{array} \end{array}\right]$$

where

$$(5.14)\qquad \Omega=\sum_{l=1}^{N-1}\left[\prod_{j=1}^{l} V_j^{11}\right]^{-1} V_l^{12}\prod_{j=1}^{l-1} V_j^{22}.$$

Although these expressions look fairly complicated, they show directly that all elements of $\tilde{\mathbb{L}}$ and $\tilde{\mathbb{U}}$ are indeed "reasonably" bounded if both $\|\prod_{j=1}^{l} V_j^{22}\|$ and $\|[\prod_{j=1}^{l} V_j^{11}]^{-1}\|$ are reasonably bounded. They also show that if either one of these grows exponentially with l, (as in Example 4.2) we can expect instability. Therefore our attention in the next sections will first focus on the problem of how to find estimates for them.

6. Green's functions associated with $\tilde{\mathbb{A}}$. As an intermezzo we give below explicit expressions for blocks appearing in $\tilde{\mathbb{A}}^{-1}$, in terms of a *fundamental solution* of the following recursion. Let the sequence of $n\times n$ matrices $\{\Phi_1,\cdots,\Phi_N\}$ satisfy the transformed homogeneous recursion (cf. (3.1) and (5.2))

$$(6.1)\qquad \Phi_{i+1}=V_i\Phi_i,\qquad i=1,\cdots,N-1,$$

and also the BC

$$(6.2)\qquad \tilde{M}_1\Phi_1+\tilde{M}_N\Phi_N=I.$$

Note that although the Φ_i are not necessarily upper triangular, we see that $\Phi_i\Phi_{j+1}^{-1}$ should be upper triangular for each i and j. For $j=1,\cdots,N-1$ define Green's functions $\{Z_{i,j}\}_{i=1}^{N}$ satisfying

$$(6.3a)\qquad Z_{i+1,j}=V_iZ_{i,j}+\Delta_{i,j},$$

where

$$(6.3b)\qquad \Delta_{i,j}=\begin{cases} I & \text{if } i=j,\\ 0 & \text{if } i\neq j\end{cases}$$

and

$$(6.4)\qquad \tilde{M}_1 Z_{1,j}+\tilde{M}_N Z_{N,j}=0.$$

It can easily be checked that

$$(6.5)\quad \begin{aligned} &\text{(a)}\quad Z_{i,j}=-\Phi_i\tilde{M}_N\Phi_N\Phi_{j+1}^{-1}=-\Phi_i\Phi_{j+1}^{-1}+\Phi_i\tilde{M}_1\Phi_1\Phi_{j+1}^{-1} &&\text{for } i\leqq j,\\ &\text{(b)}\quad Z_{i,j}=\Phi_i\tilde{M}_1\Phi_1\Phi_{j+1}^{-1}=\Phi_i\Phi_{j+1}^{-1}-\Phi_i\tilde{M}_N\Phi_N\Phi_{j+1}^{-1} &&\text{for } i\geqq j+1.\end{aligned}$$

If we formally define

$$(6.6)\qquad Z_{i,N}:=\Phi_i,$$

and denote by $\mathbb{Z}$ the matrix with $Z_{i,j}$ as the ij block; then

$$(6.7)\qquad \tilde{\mathbb{A}}\mathbb{Z}=\mathbb{P}^{-1},$$

where $\mathbb{P}^{-1}$ is defined as in (3.6) and where $P=I$. Hence we see that

$$(6.8)\qquad \|\mathbb{Z}\|=\|\tilde{\mathbb{A}}^{-1}\|.$$

7. Estimates for $\|\mathbb{L}^{-1}\|$ and $\|\mathbb{U}^{-1}\|$. In this section we shall show that there exists at least one suitable arrangement of the BC such that a block LU-decomposition of the resulting matrix $\mathbb{A}$ gives lower and upper triangular matrices the inverses of which are bounded by a constant of the order of $\|\mathbb{A}^{-1}\|$ (cf. the "exponential growth" in Example 4.2).

A preliminary step consists of investigation of growth factors ("Floquet numbers") in the matrix $M_1G_1^{-1}\cdots G_{N-1}^{-1}$. For this we use the singular value decomposition (cf. [7])

$$(7.1)\qquad M_1G_1^{-1}\cdots G_{N-1}^{-1}=P_0\Sigma P_N^{-1},$$

where P_0 and P_N are orthogonal matrices and Σ is a (semi) positive diagonal matrix, say

$$(7.2)\qquad \Sigma=\operatorname{diag}(\sigma_1,\cdots,\sigma_n),$$

of which we assume $\sigma_1\leqq\sigma_2\leqq\cdots\leqq\sigma_n$. If rank $M_1=l<n$ then $\sigma_1=\cdots=\sigma_{n-l}=0$.

The matrix P_0 is used to premultiply the BC matrices M_1 and M_N. According to Property 5.1 there then exists an orthogonal matrix Q_1 such that

$$(7.3)\qquad \tilde{M}_1:=P_0^{-1}M_1Q_1$$

is upper triangular. In order to be able to apply Property 5.1 we have to show that the lower right block of P_0 is nonsingular. For this we can use

PROPERTY 7.4. *If the first $n-l$ rows of M_1 are zero, then the matrix P_0 in the singular value decomposition* (7.1) *has a block diagonal form, i.e., the upper left block is an $(n-l)\times(n-l)$ matrix and the lower right block is an $l\times l$ matrix.*

Proof. Write

$$\text{(a)}\qquad M_1G_1^{-1}\cdots G_{N-1}^{-1}P_N=P_0\Sigma.$$

The matrix on the left in (a) has the first $(n-l)$ rows equal to zero, whereas the matrix on the right has the first $(n-l)$ columns equal to zero. Hence the upper right block of P_0 and therefore also the lower left block must be zero. □

We now use the matrix Q_1 to generate $\{Q_i\}$, $\{V_i\}$ as in (5.2) and $\tilde{M}_N$ as in (5.3). It is not restrictive to identify P_N and Q_N as may be deduced from:

PROPERTY 7.5. *$P_N Q_N^{-1}$ is block diagonal like P_0; moreover the lower right block is a diagonal matrix (therefore only containing* ± 1).

Proof. From (5.2), (7.1) and (7.3) we obtain

$$\tilde{M}_1 V_1^{-1} \cdots V_{N-1}^{-1} = \Sigma P_N^{-1} Q_N = \begin{pmatrix} 0 & 0 \\ 0 & K \end{pmatrix}, \qquad K \text{ an } l \times l \text{ matrix.}$$

Thus K is upper triangular and at the same time the product of a diagonal (nonsingular) matrix and an orthogonal matrix; hence the lower right block of $P_N^{-1} Q_N$ must be a diagonal matrix. □

In constructing the singular value decomposition (7.1) we have the freedom to choose the first $(n-l)$ rows of P_N^{-1}; so we might as well take them equal to those of Q_N^{-1}. Moreover, premultiplying P_N and P_0 by a diagonal matrix consisting of ± 1 does not affect Σ. Hence we may assume that $P_N = Q_N$.

The next lemma now shows how the choice of the partitioning and the growth of the S_i and T_i blocks can be related in terms of the singular values σ_i.

LEMMA 7.6. *Let $k \geqq n-l$ (for k see* (3.6) *and for l see Assumption* 4.1). *Let $\|\mathbb{A}^{-1}\| \leqq \kappa$. Then*

(i) $$\forall_i \quad \left\| \tilde{M}_1^{11} \left[\prod_{j=1}^{i} V_j^{11} \right]^{-1} \right\| \leqq \kappa(1+\sigma_k);$$

(ii) $$\forall_i \quad \left\| \prod_{j=1}^{i} V_j^{22} [\tilde{M}_1^{22}]^{-1} \right\| \leqq \kappa(1+\sigma_{k+1}^{-1}).$$

Proof. From the proof of Property 7.5 we see that $\tilde{M}_1[\prod_{j=1}^{N-1} V_j]^{-1} = \Sigma$. Since we find from (6.2) that $\Phi_N^{-1} = \tilde{M}_1 \Phi_1 \Phi_N^{-1} + \tilde{M}_N$ and from (6.5b) that $\Phi_N^{-1} Z_{N,i} = \tilde{M}_1 \Phi_1 \Phi_{i+1}^{-1}$, we thus have

$$\tilde{M}_1 \left[\prod_{j=1}^{i} V_j \right]^{-1} = \tilde{M}_1 \Phi_1 \Phi_{i+1}^{-1} = \tilde{M}_N Z_{N,i} + \tilde{M}_1 \Phi_1 \Phi_N^{-1} Z_{N,i} = \tilde{M}_N Z_{N,i} + \Sigma Z_{N,i}.$$

By considering upper left $(k \times k)$ blocks in this relation and noting that $\|Z_{N,i}\| \leqq \|\mathbb{A}^{-1}\| = \|\mathbb{A}^{-1}\|$, we immediately obtain (i).

In order to show that (ii) holds we first deduce from (6.2)

(a) $$\Phi_i \Phi_1^{-1} = \Phi_i \tilde{M}_1 + \Phi_i \tilde{M}_N \Phi_N \Phi_1^{-1}.$$

Utilizing that $\Phi_i \Phi_1^{-1}$, $\tilde{M}_1$ and $\Phi_N \Phi_1^{-1}$ are uppertriangular, we then find

(b) $$(\Phi_i \tilde{M}_N)^{21} = -\Phi_i^{21} \tilde{M}_1^{11} [(\Phi_N \Phi_1^{-1})^{11}]^{-1} = -\Phi_i^{21} \tilde{M}_1^{11} (\Phi_1 \Phi_N^{-1})^{11}.$$

Moreover, since $(\tilde{M}_1 \Phi_1 \Phi_N^{-1})(\Phi_N \Phi_N^{-1}) = \tilde{M}_1$ there holds

(c) $$\tilde{M}_1^{11} (\Phi_1 \Phi_N^{-1})^{11} (\Phi_N \Phi_1^{-1})^{12} = M^{12}$$

(note that $\tilde{M}_1 \Phi_1 \Phi_N^{-1}$ is a diagonal matrix).

We now substitute (b) and (c) in the following relation, which is obtained from (a) by taking the lower right blocks on both sides

$$(\Phi_i \Phi_1^{-1})^{22} = \Phi_i^{21} \tilde{M}_1^{12} + \Phi_i^{22} \tilde{M}_1^{22} + (\Phi_i \tilde{M}_N)^{21} (\Phi_N \Phi_1^{-1})^{12} + (\Phi_i \tilde{M}_N)^{22} (\Phi_N \Phi_1^{-1})^{22},$$

which gives us

(d) $$(\Phi_i \Phi_1^{-1})^{22} = \Phi_i^{22} \tilde{M}_1^{22} + (\Phi_i \tilde{M}_N)^{22} (\Phi_N \Phi_1^{-1})^{22}.$$

Finally it follows from (7.1) (cf. also proof of Property 7.5) that $\tilde{M}_1^{22}(\Phi_1^{-1}\Phi_N)^{22}=\Sigma^{22}$, i.e.,

$$\text{(e)} \qquad \|(\Phi_N\Phi_1^{-1})^{22}[\tilde{M}_1^{22}]^{-1}\| = \|[\Sigma^{22}]^{-1}\| \leqq \frac{1}{\sigma_{k+1}}.$$

Using $\|\Phi_i\| \leqq \|\tilde{\mathbb{A}}^{-1}\| = \|\mathbb{A}^{-1}\|$, $\|\tilde{M}_N^{22}\| \leqq \|M_N\| \leqq 1$ and (e) in (d) yields the required estimate. □

Before giving our final stability estimate, we would first like to note that the expression for $\tilde{U}_N$ in (5.13) can be much simplified. Indeed we have:

PROPERTY 7.7. *For $\tilde{U}_N$ there holds*

$$\tilde{U}_N = \left[\begin{array}{c|c} \prod_{j=1}^{N-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{21} & -I+\prod_{j=1}^{N-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}\tilde{M}_N^{22} \\ \hline \tilde{M}_N^{11}+\tilde{M}_1^{11}\left[\prod_{j=1}^{N-1} V_j^{11}\right]^{-1} & \tilde{M}_N^{12} \end{array}\right].$$

Proof. We only have to show that $\tilde{M}_1^{11}\Omega=\tilde{M}_1^{12}$. This can be done as follows: Define a fundamental solution $\{\Psi_i\}$ (i.e., $\Psi_{i+1}=V_i\Psi_i$), which satisfies the BC ${}^2\Psi_1=(0 \mid I)$ and ${}^1\Psi_N=(I \mid 0)$. We then have

$$\Psi_1=\begin{bmatrix} \left[\prod_{j=1}^{N-1} V_j^{11}\right]^{-1} & -\Omega \\ 0 & I \end{bmatrix}, \qquad \Psi_N=\begin{bmatrix} I & 0 \\ 0 & \prod_{j=1}^{N-1} V_j^{22} \end{bmatrix}.$$

Hence, we see

$$0=\Sigma^{12}=-(\tilde{M}_1\Phi_1\Phi_N^{-1})^{12}=-(\tilde{M}_1\Psi_1\Psi_N^{-1})^{12}=\tilde{M}_1^{11}\Omega[\Pi V_j^{22}]^{-1}-\tilde{M}_1^{12}[\Pi V_j^{22}]^{-1},$$

from which the assertion simply follows. □

We can now prove the following:

THEOREM 7.8. *Let $k \geqq n-l$. Let $\|\mathbb{A}^{-1}\| \leqq \kappa$, $\forall_i$ $\|G_i\| \leqq \gamma$. Then*

$$\|\tilde{\mathbb{L}}\| \leqq N\left[1+\gamma\left(1+\frac{1}{\sigma_{k+1}}\right)\right][1+\kappa(1+\sigma_k)],$$

$$\|\tilde{\mathbb{U}}\| \leqq (N-1)\left[1+\gamma+\kappa\left(1+\frac{1}{\sigma_{k+1}}\right)\right]+2+\kappa(3+\sigma_k).$$

The proof follows from estimating $\tilde{\mathbb{L}}$ by block columns and $\tilde{\mathbb{U}}$ by block rows. Estimates for these blocks, see (5.9), (5.10), (5.11), (5.12) and Property 7.7, can be found using Property 7.6. (Note that a bound for $\|[\tilde{M}_1^{22}]^{-1}\|$ also follows from Property 7.6.)

Remark 7.9. The factor N in the estimates in Theorem 7.8 might be a slight overestimate since we used a very simple estimation technique for $\|\tilde{\mathbb{L}}\|$ and $\|\tilde{\mathbb{U}}\|$. Of much greater importance, however, is the fact that this result shows that an improper splitting (i.e., a wrong choice of k) may result in a large $\|\tilde{\mathbb{L}}\|$ and $\|\tilde{\mathbb{U}}\|$. It should be noted that Σ^{11} $(=\tilde{M}_1^{11}[\prod_{j=1}^{N-1} V_j^{11}]^{-1})$ as such appears at several places in $\tilde{\mathbb{L}}$ and $\tilde{\mathbb{U}}$, whereas $[\Sigma^{22}]^{-1}$ $(=\prod_{j=1}^{N-1} V_j^{22}[\tilde{M}_1^{22}]^{-1})$ only appears postmultiplied by $\tilde{M}_N^{21}$ or $\tilde{M}_N^{22}$. Hence, if either $\tilde{M}_1^{11}$ or $\tilde{M}_N^{21}$ or $\tilde{M}_N^{22}$ is nonzero, then we can expect that at least the σ_k term or the σ_{k+1}^{-1} term should appear in a realistic estimate thus making this estimate qualitatively sharp.

A special case occurs when $\tilde{M}_1^{11}$, $\tilde{M}_1^{21}$ and $\tilde{M}_N^{22}$ are zero. We then have:

PROPERTY 7.10. *Let $k=n-l$, so $\tilde{M}_1^{11}$, M_1^{12} are zero, and let $\tilde{M}_N^{21}$, $\tilde{M}_N^{22}$ be zero too* (*i.e., the* BC *are separated*). *Then the estimates in Theorem* 7.8 *hold without the σ_k and σ_{k+1}^{-1} terms.*

Proof. Apparently we have $\sigma_1=\cdots=\sigma_k=0$. Substitution of $\tilde{M}_1^{11}=0$ in (5.11) and Property 7.7 therefore makes the blocks containing $[\prod_{j=1}^{N-1} V_j^{11}]^{-1}$ disappear. A similar result follows for blocks containing $\prod_{j=1}^{N-1} V_j^{22}$. □

As was indicated in Property 3.20 we can given bounds for $\|\tilde{\mathbb{L}}^{-1}\|$ and $\|\tilde{\mathbb{U}}^{-1}\|$ using Theorem 7.8. We are, however, interested in bounds for $\|\hat{\mathbb{L}}^{-1}\|$ and $\|\hat{\mathbb{U}}^{-1}\|$, i.e., for the stability constants that arise in an actual LU-decomposition of $\mathbb{A}$, where a restricted pivoting strategy is used (see (3.14)). Because of the special structure of M_1 (see Assumption 4.1) the permutation of rows in the BC, i.e., the premultiplication by $\hat{P}$, can be described by a block diagonal matrix of which the upper left $(n-l)\times(n-l)$ block is an identity matrix. From Property 7.4 we see that we may assume that the matrix P_0 has a similar block diagonal form as this matrix $\hat{P}$, so $\hat{\mathbb{P}}\tilde{\mathbb{P}}^{-1}=\operatorname{diag}(I,\cdots,I,\hat{P}P_0^{-1})$. Hence to any LU-decomposition (3.14) we can associate a special LU-decomposition (5.6) for which there holds

$$\hat{\mathbb{P}}\hat{\mathbb{D}}\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{P}}^{-1}=\hat{\hat{\mathbb{D}}}:=\operatorname{diag}(I_l,\hat{\hat{D}}_1,\cdots,\hat{\hat{D}}_{N-1},I_{n-l}). \tag{7.11}$$

Comparing (3.14) and (5.6) now gives

$$\mathbb{A}=\hat{\mathbb{D}}^{-1}\hat{\mathbb{P}}^{-1}\hat{\mathbb{A}}\hat{\mathbb{E}}^{-1}=\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{P}}^{-1}\tilde{\mathbb{A}}\tilde{\mathbb{E}}^{-1}, \tag{7.12}$$

whence, using $\hat{\mathbb{A}}=\hat{\mathbb{L}}\hat{\mathbb{U}}$ and $\tilde{\mathbb{A}}=\tilde{\mathbb{L}}\tilde{\mathbb{U}}$, we obtain

$$\hat{\mathbb{U}}\hat{\mathbb{E}}^{-1}\tilde{\mathbb{E}}\tilde{\mathbb{U}}^{-1}=\hat{\mathbb{L}}^{-1}\hat{\mathbb{P}}\hat{\mathbb{D}}\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{P}}^{-1}\tilde{\mathbb{L}}=\hat{\mathbb{L}}^{-1}\hat{\hat{\mathbb{D}}}\tilde{\mathbb{L}}. \tag{7.13}$$

In (7.13) the leftmost matrix is block upper triangular and the rightmost matrix is "block" lower triangular, however with a structure like $\hat{\hat{\mathbb{D}}}$. Simple computation shows that the latter matrix can be repartitioned as

$$\begin{bmatrix} K_1 & F_1 & & \\ & \ddots & \ddots & \\ & & K_2 & F_{N-1} \\ & & \ddots & \\ & & & K_N \end{bmatrix}, \tag{7.14}$$

where K_j has a structure like

$$\begin{bmatrix} K_j^{22} & \phi \\ K_j^{12} & K_j^{11} \end{bmatrix}$$

(K_j^{22} being an $l\times l$ matrix) and F_j like

$$\begin{bmatrix} \phi & \phi \\ F_j^{12} & \phi \end{bmatrix}.$$

As was shown in [10], if we think of (3.14) as arisen from restricted pivoting in rows $s(n-1)+l+1$ through $sn+l$ $(s=1,\cdots,N-1)$ we obtain the following estimate:

PROPERTY 7.15. *Let $k\geqq n-l$ and $\|G_i\|\leqq\gamma$ for all i. Then $\|K_j\|\leqq 1+\max(\gamma, g(l,n-l))$ and $\|F_j\|\leqq 1$. Here $g(l,n-l)$ is the growth factor arising in partial pivoting, which is bounded by $l\sqrt{n-l}2^l$.*

Estimates for blocks of $\hat{\mathbb{L}}$, $\hat{\mathbb{L}}^{-1}$, $\mathbb{U}$ and $\hat{\mathbb{U}}^{-1}$ can now be given easily using (7.13). In particular we have:

THEOREM 7.16. *Let $k \geqq n-l$, $\|A^{-1}\| \leqq x$ and $\|G_i\| \leqq \gamma$ for all i. Then*

$$\|\hat{T}_i\| \leqq \kappa(2+\max(\gamma, g(l, n-l))(1+\sigma_{k+1}^{-1}),$$

$$\|\hat{S}_i\| \leqq \kappa(2+\max(\gamma, g(l, n-l))(1+\sigma_k).$$

From Theorem 7.16 we deduce that it is preferable for k to be such that both $\sigma_k \leqq 1$ and $\sigma_{k+1}^{-1} \leqq 1$. Moreover we need to require $k \geqq n-l$ in order to have a nonsingular first pivotal block. This leads to:

COROLLARY 7.17. *In order to have stability one should choose k such that $k = \max(n-l, n-m)$, where m is the largest integer such that $\sigma_1, \cdots, \sigma_{n-m} < 1$.*

Remark 7.18. Although we only gave upper bounds in Theorem 7.16, it follows from what has been said in Remark 7.9 and Property 7.10 that these bounds are fairly realistic. As a consequence the particular choice $k = n-l$, which is recommended in so-called *partially separated* BC may be very bad from a stability point of view. Another consequence is a confirmation of a result already established in [10], viz., that the restricted pivoting strategy for *separated* BC is stable.

As we already mentioned in § 3, one might as well consider the BVP as a backward problem. In that case a rank deficiency of M_N might be utilized. Based on similar considerations now for the singular values of $M_N G_{N-1} \cdots G_1$, an analysis could be given to indicate how the stability is affected by the partitioning of the BC. Such an analysis, however, is straightforward and will not be carried out.

8. A stable LU-decomposition technique. In the previous section we showed that block LU-decomposition might be a hazardous undertaking if no information about the singular values of $M_1 G_1^{-1} \cdots G_{N-1}^{-1}$ is available. However, this information is not easy to obtain as some of these singular values may be very small and others may be extremely large. Hence, apart from the fact that the calculation of the matrix $M_1 G_1^{-1} \cdots G_{N-1}^{-1}$ would make a method less attractive regarding the computational cost, it may also be practically impossible to compute the σ_i. In order to find a more useful way to decide on the proper splitting of the BC, we first note that if we could compute the matrices V_i, appearing in § 5, in advance, we might be able to *predict* the increments $[\prod_{j=1}^{N-1} V_j^{11}]^{-1}$ and $\prod_{j=1}^{N-1} V_j^{22}$. This would give reasonable bounds for

$$\Sigma^{11} = \tilde{M}_1^{11}\left[\prod_{j=1}^{N-1} V_j^{11}\right]^{-1} \quad \text{and} \quad [\Sigma^{22}]^{-1} = \prod_{j=1}^{N-1} V_j^{22}[\tilde{M}_1^{22}]^{-1}. \tag{8.1}$$

Indeed, if $\|[\tilde{M}_1^{22}]^{-1}\|$ is large, then the problem as such is ill-conditioned (cf. [9]), so $\|\mathbb{A}^{-1}\|$ is large; this also appears—though indirectly—from Lemma 7.6(ii), where $(1+\sigma_{k+1}^{-1})\kappa$ should be a bound for $\|[\tilde{M}_1^{22}]^{-1}\|$ for any choice of k. Moreover, $\|M_1\|$ and so $\|\tilde{M}_1^{11}\|$ is bounded by 1. Now assume that κ is not large; then we may expect that a proper splitting will assure that both $\|[\prod_{j=1}^{i} V_j^{11}]^{-1}\|$ and $\|\prod_{j=1}^{i} V^{22}\|$ remain bounded by moderate constants (cf. Lemma 7.6). The actual role of Q_1 is less important in this. Indeed, one should realize that Q_1 is in fact the initial value of some fundamental solution (of which the ith iterate equals $\prod_{j=1}^{i-1} G_j Q_1$), so it is most likely that in an arbitrary choice of Q_1 the first few, say k, columns are initial values of *unstable modes.* Now it was shown in [8] that the recursive computation of the V_i via (5.2) leads to matrices of which the magnitudes of the diagonal elements reflect the increments of the various growth classes that build up a fundamental solution of (3.1). In particular, if these first k columns of Q_1 define unstable modes then the larger increments will approximately appear as the first k diagonal elements, whereas the smaller ones (corresponding to the stable modes) appear thereafter. Of course this also implies $\text{lub}(V_i^{22}) \lesssim 1$ and $\text{glb}(V_i^{11}) \gtrsim 1$, which implies that the forward substitution and the backward substitution recursion are expected to be stable. This then provides a simple way to detect a proper splitting, i.e., a value for k. Only in special circumstances therefore may we expect the choice of Q_1 as, e.g., induced by Property 5.1 to be inappropriate. In those cases we will see a disorder of the diagonal elements of the V_i, at least initially. Then we stop after a few steps and try another Q_1, $\hat{Q}_1$ found, say, from permuting the columns of Q_1 (since a disorder at the diagonal of V_1 apparently means that at least one of the first k columns of Q_1 defines a stable solution).

This matter is discussed in more detail in [11]; also there it is indicated how a solution $\mathbf{x}$ of (3.4) may be computed without block LU-decomposition. Nevertheless we like to show that this algorithm may also be viewed as a special block LU-solver. We briefly describe the successive steps.

Step I. Compute Q_1 as in Property 5.1 (with $P = I$), and compute the first few matrices Q_{i+1} and V_i (say $i = 1, 2, 3$) via (5.2), and check whether these V_i or their product are "ordered" (i.e., elements along the diagonals appear in decreasing modulus going from above to below).

Step II. If these V_i are not ordered, choose a different matrix $\hat{Q}_1$ by permuting appropriate columns (see Remark 8.3), compute the thus induced upper triangular V_i (via (5.2)) and check the ordering. Repeat till ordering is found satisfactory. Then proceed to Step III.

Step III. Complete the computation of the $\{Q_i\}$ and $\{V_i\}$. Sum up the logarithms of corresponding diagonal elements of the V_i, $i = 1, \cdots, N-1$. Then define k as the largest diagonal element index (counting diagonal elements from above to below) for which this sum of logarithms is positive, under the restriction that $k \geqq n - l$.

Step IV. Compute an orthogonal matrix P such that PM_1Q_1 (or $PM_1\hat{Q}_1$) is uppertriangular.

Remark 8.2. As is intuitively clear only the nonzero rows in M_1 (in case $l < n$) can control solutions of the recursion and can only control the stable solutions in a meaningful way (as these need initial conditions). Since the first $(n-l)$ columns of Q_1 are orthogonal to these nonzero row vectors, they are not controlled by initial conditions and hence cannot induce stable modes (more formally this is also a consequence of [10, Thm. 4.6]). This means that we should expect the first $(n-l)$ diagonal elements of the V_i to represent increments of unstable modes only.

Remark 8.3. From Remark 8.2 it follows that in case $\text{rank}(M_1) < n$, no such permutation of columns of Q_1 is needed, where any of the first $(n-l)$ columns is involved. Therefore any $\hat{Q}_1$, being the product of Q_1 and such a permutation, will also leave the zero column structure of $M_1\hat{Q}_1$ the same as in M_1Q_1. Once we have found an acceptable matrix $\hat{Q}_1$ (inducing the required splitting in the V_i), we can find an orthogonal block diagonal matrix P of which the upper left block is an identity matrix, such that $\tilde{M}_1 = PM_1\hat{Q}_1$ has the form as described in Property 5.1.

Remark 8.4. The computation of k as is indicated in Step III is much safer than a computation based on information in the first few V_i. For instance the incremental growth may be close to 1 for a few solutions (thinking of small discretization steps of the ODE which gave rise to the problem). This might make a decision where to split very hard.

To illustrate this strategy we shall apply it to Example 4.2. Since M_1 is uppertriangular, we could start with $Q_1 = I$. Note that G_1 is already uppertriangular, but has a disordered diagonal, viz., $(\frac{1}{2}, 3, \frac{1}{4})$. Hence, we have to permute columns of Q_1.

There is no restriction, as $l=n$, so we take the obvious choice

$$\hat{Q}_1=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}. \tag{8.5}$$

We then compute orthogonal $\{\hat{Q}_i\}$ and upper triangular $\{V_i\}$ satisfying

$$G_i\hat{Q}_i=\hat{Q}_{i+1}V_i. \tag{8.6}$$

We obtain, e.g.,

$$\hat{Q}_2=\frac{1}{\sqrt{10}}\begin{bmatrix}1&3&0\\3&-1&0\\0&0&\sqrt{10}\end{bmatrix},\quad V_1=\begin{bmatrix}\sqrt{10}&\sqrt{10}/10&0\\0&3\sqrt{10}/20&0\\0&0&\frac{1}{4}\end{bmatrix}\approx\begin{bmatrix}3.2&.16&0\\0&.48&0\\0&0&.25\end{bmatrix}. \tag{8.7}$$

For i large we obtain

$$\hat{Q}_{i+1}\approx\frac{2}{\sqrt{29}}\begin{bmatrix}1&\frac{5}{2}&0\\\frac{5}{2}&-1&0\\0&0&\sqrt{29}/2\end{bmatrix},\quad V_i\approx\begin{bmatrix}3&-1&0\\0&5&0\\0&0&.25\end{bmatrix}. \tag{8.8}$$

Obviously we should choose $k=1$. The matrix $\tilde{M}_1$ becomes

$$\tilde{M}_1=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\0&0&-1\end{bmatrix}\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}=\begin{bmatrix}1&0&0\\0&-1&0\\0&0&-1\end{bmatrix}. \tag{8.9}$$

Summarizing, we obtain the following system:

$$\hat{\mathbb{A}}\approx\left[\begin{array}{cccccccccccccc}
0&-1&0&0&0&0&&&&&&1&0&0\\
0&0&-1&0&0&0&&&&&&0&0&0\\
3.2&.16&0&-1&0&0&&&&&&0&0&0\\
0&.48&0&0&-1&0&&&&&&&&\\
0&0&.25&0&0&-1&&&&&&&&\\
0&0&0&3.1&-.03&0&\ddots&&&&&&0&\\
&&&&&&\ddots&&&&&&&\\
&&0&&&&\ddots&&&&&&&\\
&&&&&&&0&-1&0&0&0&0&\\
&&&&&&&0&0&-1&0&0&0&\\
&&&&&&&3&-1&0&-1&0&0&\\
0&0&0&&&&&0&.5&0&0&-1&0&\\
0&0&0&&&&&0&0&.25&0&0&-1&\\
0&1&0&&&&&0&0&0&0&1&0&
\end{array}\right].$$

Apparently we have $\sigma_1\approx(\frac{1}{3})^{N-1}$, $\sigma_2\approx2^{N-1}$ and $\sigma_3\approx4^{N-1}$. From Table 4.1 we deduce $\|\mathbb{A}^{-1}\|\approx3$; finally we see that $\|G_i\|\approx3$. Now if we apply Theorem 7.8 or Theorem 7.16, we find that the choice $k=2$ (cf. (4.5)) would give us estimates for $\|\mathbb{L}^{-1}\|$, $\|\mathbb{U}^{-1}\|\sim3N2^{N-1}$. The choice $k=1$ just gives estimates of the order $3N$.

REFERENCES

[1] U. ASCHER, U. S. PRUESS AND R. D. RUSSELL, *On spline basis selection for solving differential equations*, SIAM J. Numer. Anal., 20 (1983), pp. 121–142.

[2] C. DE BOOR AND R. WEISS, SOLVEBLOK: *A package for solving almost block diagonal linear systems*, ACM Trans. Math. Software, 6 (1980), pp. 80–87.

[3] R. FOURER, *Sparse Gaussian elimination of staircase linear systems*, Tech. Rept. SOL 79-17, Dept. Operations Research, Stanford Univ., Stanford, CA, 1979.

[4] P. KEAST AND G. FAIRWEATHER, *On the solution of almost block diagonal systems*, Rept. University of Kentucky, Lexington, 1981.

[5] H. B. KELLER, *Accurate difference methods for nonlinear two-point boundary value problems*, SIAM J. Numer. Anal., 11 (1974), pp. 305–320.

[6] ———, *Numerical Solution of Two Point Boundary Value Problems*, CBMS Regional Conference Series in Applied Mathematics 24, Society for Industrial and Applied Mathematics, Philadelphia, 1976.

[7] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[8] R. M. M. MATTHEIJ, *Stable computation of unstable linear initial value recursions*, BIT, 22 (1982), pp. 79–93.

[9] ———, *The conditioning of linear boundary value problems*, SIAM J. Numer. Anal., 19 (1982), pp. 963–978.

[10] ———, *The stability of LU-decompositions of block tridiagonal systems*, Bull. Austr. Math. Soc., 29 (1984), pp. 177–205.

[11] R. M. M. MATTHEIJ AND G. W. M. STAARINK, *An efficient algorithm to solve general linear two point* BVP, Rept., Mathematisch Instituut, Kath. Universiteit, Nijmegen, the Netherlands, 1982; SIAM J. Sci. Stat. Comput., 5 (1984), to appear.

[12] J. M. VARAH, *On the solution of block-tridiagonal systems arising from certain finite-difference equations*, Math. Comp., 26 (1972), pp. 859–868.

[13] ———, *Alternate row and column elimination for solving certain linear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 71–75.

[14] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

005

THE k-DOMINATION AND k-STABILITY PROBLEMS ON SUN-FREE CHORDAL GRAPHS*

GERARD J. CHANG†‡ AND GEORGE L. NEMHAUSER†

Abstract. The k-domination problem is to find a minimum cardinality vertex set D of a graph such that every vertex of the graph is within distance k from some vertex of D, where k is a positive integer. The k-stability problem is to find a maximum cardinality vertex set S such that the distance between any two distinct vertices of S is greater than k. For sun-free chordal graphs, $2k$-stability and k-domination are dual problems. In particular, a minimum cardinality set of vertices D such that every vertex is within distance k of D has the same cardinality as a maximum cardinality set of vertices S such that the distance between every pair of vertices in S is greater than $2k$. To obtain this result we establish some theorems about the powers and radius of chordal graphs. Efficient algorithms for both problems on sun-free chordal graphs are obtained by transforming them to solvable cases of the clique covering and vertex packing problems. We also prove the NP-completeness of both problems on bipartite and chordal graphs.

Key words. combinatorial optimization, graph theory, domination, stability, chordal graphs

AMS(MOS) subject classifications. 05C70, 90C10

1. Introduction. All graphs in this paper are simple, i.e. finite, undirected, loopless and without multiple edges. The *length* of a path from a vertex x to a vertex y is the number of edges in the path. The *distance* $d_G(x, y)$ from vertex x to vertex y in a graph $G=(V, E)$ is the length of a minimum length path from x to y; $d_G(x, y)=\infty$ if there is no path from x to y. If $d(x, y)\leqq k$ for all $y\in S$, x is said to *dominate S within distance k*. A *k-dominating set* (*k-covering*) is a vertex set $D\subseteq V$ such that every vertex in V is dominated within distance k by some vertex in D. A *k-stable set* (*k-packing, k-independent set*) is a vertex set $S\subseteq V$ such that $d(x, y)>k$ for every distinct pair of vertices x and y in S. The *k-domination problem* is to find the *k-domination number* $\delta_k(G)$, which is the minimum cardinality of a k-dominating set in G. The *k-stability problem* is to find the *k-stability number* $\alpha_k(G)$, which is the maximum cardinality of a k-stable set in G.

S is a $2k$-stable set if and only if for every pair x and y in S there is no $z\in V$ such that $d(z, x)\leqq k$ and $d(z, y)\leqq k$. Thus $\delta_k(G)\geqq\alpha_{2k}(G)$, which establishes a weak duality between the $2k$-stability problem and the k-domination problem.

Suppose we assign each vertex v of the graph a real weight w_v. The *weighted k-domination problem* is to find a k-dominating set D such that $\sum_{v\in D} w_v$ is as small as possible. Similarly, the *weighted k-stability problem* is to find a k-stable set S such that $\sum_{v\in S} w_v$ is as large as possible.

Applications of domination and bounds on the 1-domination number have been presented in several papers; see Liu [1968], Berge [1973], Cockayne [1978], and Abbott and Liu [1979]. For trees, Meir and Moon [1975] have proved strong duality, i.e. $\delta_k(G)=\alpha_{2k}(G)$ for all $k\geqq 1$. Recently, Farber [1981] has proved $\delta_1(G)=\alpha_2(G)$ for a class of graphs that he called strongly chordal. In this paper, we use the more descriptive term—*sun-free chordal.* These graphs will be defined in § 2.

The k-domination and k-stability problems are NP-complete for general graphs. The 1-domination problem is NP-complete for planar graphs with maximum vertex

* Received by the editors August 11, 1982, and in revised form May 27, 1983. This research was supported by the National Science Foundation under grant ECS-8005350 to Cornell University. The results given here also appear in Gerard Chang's Ph.D. dissertation (Chang [1982]).

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853.

‡ Present address: Department of Mathematics, National Central University, Ching-li, Taiwan 320, Republic of China.

degree 3, planar graphs that are regular of degree 4 (Garey and Johnson [1979]), chordal graphs (Booth [1980]) and undirected path graphs (Booth and Johnson [1982]). For any fixed k, the k-domination problem is NP-complete for bipartite graphs and chordal graphs of diameter $2k+1$ (§ 5). Similar results for the k-stability problem are also given in § 5.

Although the 1-domination problem is NP-complete on chordal graphs, efficient algorithms are known for certain subclasses of chordal graphs. In particular, linear algorithms have been found for the 1-domination problem on trees (Cockayne, Goodman and Hedetniemi [1975]), powers of forests (Slater [1976]), powers of block graphs (Chang and Nemhauser [1982]), directed path graphs, which include interval graphs (Booth and Johnson [1982]), and for the weighted 1-domination problem on trees (Natarajam and White [1978], Kariv and Hakimi [1979]). Farber [1981] recently gave a polynomial algorithm for the weighted 1-domination problem on sun-free chordal graphs, which is a class that includes all of the above graphs for which polynomial algorithms are known.

The main purpose of this paper is to establish strong duality and to give good algorithms for the (unweighted) k-domination and k-stability problems on sun-free chordal graphs. To obtain these results, we need some properties of the powers and radii of sun-free chordal graphs. A main result here is that powers of sun-free chordal graphs are chordal. When the original version of this paper was written in April of 1982, we knew that if powers of sun-free chordal graphs also were sun-free chordal, then the duality and algorithmic results would generalize to the weighted k-domination and k-stability problems. Using their very nice characterization of totally balanced matrices, Edmonds and Lubiw (personal communication, April 1982) proved this conjecture, see Lubiw [1982].

To motivate our approach, we now sketch a new proof of Meir and Moon's duality result for trees. The *vertex packing* (*stability*) *problem* is just the 1-stability problem and we use $\alpha(G)$ for $\alpha_1(G)$. The *clique covering problem* is to find a minimum cardinality collection of cliques of a graph $G=(V,E)$ whose union is V. The *clique covering number* $\theta(G)$ is the minimum cardinality of a clique covering of G. For every graph G, we have the weak duality inequality $\theta(G)\geqq\alpha(G)$.

For a vertex set S of a graph $G=(V,E)$, the *subgraph induced* by S is defined by $G_S=(S,E_S)$, where $E_S=\{(x,y)|x,y\in S \text{ and } (x,y)\in E\}$. A graph G is *perfect* if $\theta(G_S)=\alpha(G_S)$ for all vertex induced subgraphs G_S of G. Examples of perfect graphs are chordal graphs, comparability graphs and unimodular graphs; see Berge [1973] and Golumbic [1980]. The *kth power* of a graph $G=(V,E)$ is the graph $G^k=(V,E^k)$ with $(x,y)\in E^k$ if and only if $1\leqq d_G(x,y)\leqq k$. G^2 is called the *square* of G.

Suppose G is a tree; then the following three properties hold.

(P1) S is k-stable in G if and only if S is stable in G^k.

(P2) G^k is chordal for all positive integers k.

(P3) A vertex set S is dominated by some x within distance k if and only if S is a clique in G^{2k}.

Note that (P1) transforms the $2k$-stability problem on G to the vertex packing problem on G^{2k}, i.e. $\alpha_{2k}(G)=\alpha(G^{2k})$. (P3) transforms the k-domination problem on G to the clique covering problem on G^{2k}, i.e. $\delta_k(G)=\theta(G^{2k})$. (P2) guarantees the perfection of G^{2k}, so that $\alpha(G^{2k})=\theta(G^{2k})$. Thus $\alpha_{2k}(G)=\delta_k(G)$ for any tree G.

This proof of Meir and Moon's result motivates the study of graphs that satisfy the above properties. (P1) and the "only if" statement of (P3) obviously are true for all graphs.

In § 2, we study graphs satisfying (P2). Graphs that satisfy (P3) are hard to characterize, so we will use a property that implies (P3). Let $d(G)$ and $r(G)$ be the diameter and radius of G, respectively. Since $r(G) \geqq d(G)/2$ for all graphs, we say that G has the *minimum radius* property if

(P4) $r(H) = \lceil d(H)/2 \rceil$ for any connected induced subgraph H of G.

(P4) is true for trees (Jordan [1869] and König [1950]). In § 3, we characterize graphs that satisfy (P4) and prove that (P4) implies (P3). By putting together the results of § 2 and § 3, we obtain $\delta_k(H) = \alpha_{2k}(H)$ for any subgraph H of G for sun-free chordal graphs.

2. *P*-chordal graphs—Graphs whose powers are chordal. In a graph $G = (V, E)$, a *hole* is a simple cycle without a *chord*; i.e. no pair of nonconsecutive vertices of the cycle is joined by an edge. A graph is *chordal* (*triangulated*) if it does not have a hole of length greater than 3.

A vertex x of $G = (V, E)$ is called *simplicial* if its *neighborhood* $\text{Nbd}(x) = \{z | (z, x) \in E\}$ is a clique. In general, the *n-neighborhood* of a vertex x is defined by $\text{Nbd}(x, n) = \{z | d(z, x) = n\}$. If $d(x, y) = k$ is finite and $0 \leqq n \leqq k$, then the set of all vertices of distance n from x and distance $k - n$ from y will be denoted by $\text{Bet}(x, n, y)$.

In a graph $G = (V, E)$, an ordering $[v_1, v_2, \cdots, v_n]$ of the vertex set V is called a *perfect vertex elimination scheme* (or simply a *perfect scheme*) if each v_i is a simplicial vertex of the subgraph induced by $\{v_i, \cdots, v_n\}$. In other words, each set $X_i = \{v_j \in \text{Nbd}(v_i) | j > i\}$ is a clique.

A subset S of V is a *vertex-separator of vertices x and y* (or simply an *x–y separator*) if $x, y \notin S$, and x and y are joined by a path in G, but not joined by a path in G_{V-S}. If no proper subset of S is an x–y separator, then S is a *minimal x–y* separator.

We now give three characterizations of chordal graphs.

THEOREM 2.1. *Each of the following conditions is necessary and sufficient for a graph G to be chordal.*

(1) (Fulkerson and Gross [1965]). *G has a perfect vertex elimination scheme.*

(2) (Dirac [1961]). *Every minimal vertex separator is a clique in G.*

(3) (Walter [1972], Gavril [1974], Buneman [1974]). *G is the intersection graph of a family of subtrees of a tree.*

The main purpose of this section is to study graphs whose powers are chordal; we call these graphs P-chordal (P stands for power). Balakrishnan and Paluraja [1982] and Duchet [1982] (see Theorem 2.4) proved that odd powers of chordal graphs are chordal. However, even powers of chordal graphs are not necessarily chordal.

An n-sun is a chordal graph $G = (V, E)$ whose vertex set V can be partitioned into $Y = \{y_1, y_2, \cdots, y_n\}$ and $Z = \{z_1, z_2, \cdots, z_n\}$ satisfying the following three conditions.

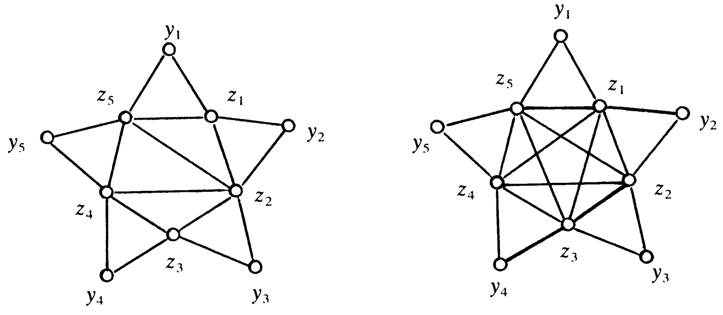
(S1) Y is a stable set in G.

(S2) $(z_1, \cdots, z_n, z_1)$ is a cycle in G.

(S3) $(y_i, z_j) \in E$ if and only if $i = j$ or $i = j + 1 \pmod{n}$.[1]

In the above definition, the z's are called *inner vertices* of the n-sun and the y's *outer vertices*. We may call G an *n-sun*, an *even sun*, an *odd sun*, or only a *sun* depending how much we specify about n. If Z is a clique, we call G a *complete n-sun*. Laskar

[1] It is to be understood in the sequel that whenever x_i is an arbitrary vertex on the cycle $(x_1, \cdots, x_n, x_1)$, addition of indices is assumed to be modulo n.



5-sun

Complete 5-sun

FIG. 2.1. *Examples of suns.*

and Shier [1980], [1982] used the term sunflower for sun; Farber [1981] used the terms incomplete trampoline for sun and trampoline for complete sun.

A chordal graph is called *sun-free chordal* if it does not have a sun as an induced subgraph.

The main result of this section is the following theorem.

THEOREM 2.2. *The following statements are equivalent for any graph G.*

(1) *G^k is chordal for every positive integer k.*

(2) *G and G^2 are both chordal.*

(3) *G is chordal and if G has an n-sun with $n \geqq 4$ as an induced subgraph, where the n-sun is defined on $\{y_1, \cdots, y_n, z_1, \cdots, z_n\}$, then $d_G(y_i, y_j) = 2$ for some i and j such that $j \notin \{i-1, i, i+1\}$.*

COROLLARY 2.3. *The following statements are equivalent for any graph G.*

(1) *H^k is chordal for every induced subgraph H of G and every integer $k \geqq 1$.*

(2) *H and H^2 are chordal for every induced subgraph H of G.*

(3) *G is n-sun-free chordal for $n \geqq 4$.*

After the original version of this paper was completed, we learned that Laskar and Shier [1982] had proved (2)⇔(3) of Theorem 2.2 and that Duchet [1982] had proved:

THEOREM 2.4. *If G^k is chordal, then G^{k+2} is chordal.*

Observe that (1)⇔(2) because of Theorem 2.4. So Duchet and Laskar and Shier independently proved parts of Theorem 2.2. In the rest of this section we will prove Theorem 2.2 from (1)⇒(2)⇒(3)⇒(1) by using Lemma 2.6 given below. Lemma 2.6 is also an important tool in a sequel to this paper (Chang and Nemhauser [1982a]).

LEMMA 2.5. *Suppose C is a cycle of a chordal graph G. Then for every edge (u, v) of the cycle there is a vertex w of the cycle that is adjacent to both u and v.*

Proof. We will prove the lemma by induction on the length of C. The case of length $(C) = 3$ is clear. Suppose the lemma holds for all cycles C' of length $<$ length $(C) \geqq 4$. Since G is chordal, C has a chord that decomposes C into two cycles C_1 and C_2 whose lengths are less than length (C). Suppose C_1 contains the edge (u, v). By the induction hypothesis, there is a vertex w of C_1 that is adjacent to both u and v. Since w is also a vertex of C, the lemma holds. □

LEMMA 2.6. *Suppose $G = (V, E)$ is a chordal graph and $k \geqq 2$ is a positive integer. If G^k has a hole $H = (x_1, \cdots, x_n, x_1)$ of length $n \geqq 4$ and p_i is a shortest path from x_i to x_{i+1} in G for $1 \leqq i \leqq n$, then the following three properties hold.*

(1) *k is even and $d_G(x_i, x_{i+1}) = k$ for $1 \leqq i \leqq n$.*

(2) *Let z_i be the vertex of p_i equidistant from x_i and x_{i+1} for $1 \leqq i \leqq n$; then $(z_1, \cdots, z_n, z_1)$ is a cycle in G.*

(3) *There exist $y_1, \cdots, y_n$ such that $\{y_1, \cdots, y_n, z_1, \cdots, z_n\}$ induces an n-sun such that $d_G(x_i, y_i) = k/2-1$ for $1 \leqq i \leqq n$.*

Proof. We will use the terminology *j is near to i* if $j \in \{i-1, i, i+1\}$.

CLAIM 1. *If j is not near to i and there are $u \in p_i$ and $v \in p_j$ such that $d_G(u, v) \leqq 1$, then $u \neq v$, $d_G(x_i, u) = d_G(x_{j+i}, v)$, $d_G(x_{i+1}, u) = d_G(x_j, v)$ and $d_G(x_i, x_{i+1}) = d_G(x_j, x_{j+1}) = k$.*

Proof of Claim 1. Since j is not near to i, $(x_i, x_j) \notin E^k$ and hence $d_G(x_i, x_j) > k$. Thus

$$d_G(x_j, v) + d_G(x_{j+1}, v) = d_G(x_j, x_{j+1}) \leqq k < d_G(x_i, x_j)$$
$$\leqq d_G(x_i, u) + d_G(u, v) + d_G(x_j, v),$$

which implies $d_G(x_{j+1}, v) < d_G(x_i, u) + d_G(u, v)$. Similarly, $d_G(x_i, u) < d_G(x_{j+1}, v) + d_G(u, v)$ since $j+1$ is not near to $i+1$. Hence $d_G(x_i, u) = d_G(x_{j+1}, v)$ and $d_G(u, v) = 1$. Similarly $d_G(x_{i+1}, u) = d_G(x_j, v)$. These equalities imply that

$$k < d_G(x_i, x_j) \leqq 1 + d_G(x_i, x_{i+1}) = 1 + d_G(x_j, x_{j+1}) \leqq 1 + k,$$

which implies $d_G(x_i, x_{i+1}) = d_G(x_{j,j+1}) = k$. So Claim 1 holds.

As a consequence of Claim 1, p_i intersects p_j if and only if j is near to i. Let C be the closed path in G given by $(p_1, p_2, \cdots, p_n)$. Choose a common vertex w_i of p_{i-1} and p_i so that $d_G(x_i, w_i)$ is as large as possible.

CLAIM 2.

$$(2.1) \qquad d_G(x_i, w_i) + d_G(x_{i+1}, w_{i+1}) + 1 \leqq d_G(x_i, x_{i+1})$$

and $d_G(x_i, w_i) < k/2$ for $1 \leqq i \leqq n$. (Thus C can be drawn as in Fig. 2.2.)

Proof of Claim 2. Note that

$$d_G(x_{i-1}, w_i) + d_G(x_i, w_i) = d_G(x_{i+1}, x_i)$$
$$= k < d_G(x_{i-1}, x_{i+1}) \leqq d_G(x_{i-1}, w_i) + d_G(x_{i+1}, w_i).$$

Thus

$$(2.2) \qquad 2d_G(x_i, w_i) + 1 \leqq d_G(x_i, w_i) + d_G(x_{i+1}, w_i) = d_G(x_i, x_{i+1}) \leqq k.$$

Similarly,

$$(2.3) \qquad 2d_G(x_{i+1}, w_{i+1}) + 1 \leqq d_G(x_i, x_{i+1}).$$

Together (2.2) and (2.3) imply (2.1), and (2.2) implies $d_G(x_i, w_i) < k/2$. So Claim 2 holds.

Now delete the subpaths from x_i to w_i and x_{i+1} to w_{i+1} of p_i for all i and consider the cycle $C' = (w_1, \cdots, w_2, \cdots, w_n, \cdots, w_1)$ of length $N \geqq n$ in G.

CLAIM 3. *Suppose (u, v) is a chord of C' such that $u \in p_i$ and $v \in p_j$, where j is not near to i. Then k is even and $d_G(x_i, u) = d_G(x_{i+1}, u) = d_G(x_j, v) = d_G(x_{j+1}, v) = k/2$.*

Proof of Claim 3. The chord (u, v) decomposes C' into two cycles $C_1 = (u, \cdots, w_{i+1}, \cdots, w_j, \cdots, v, u)$ and $C_2 = (v, \cdots, w_{j+1}, \cdots, w_i, \cdots, u, v)$. By Lemma 2.5, there is a vertex w in C_1 that is adjacent to both u and v. Assume $w \in p_m$. Since j is not near to i and $w \in C_1$, m is not near to both i and $j+1$. Thus either

$$k+1 \leqq d_G(x_m, x_i) \leqq d_G(x_m, w) + d_G(w, u) + d_G(x_i, u)$$

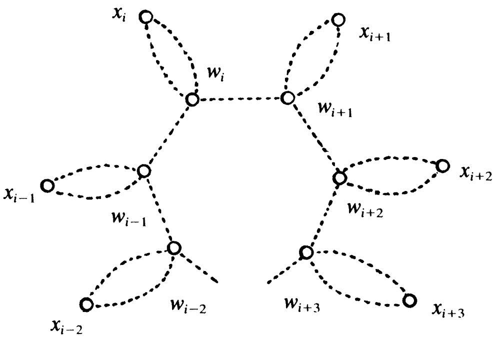


FIG. 2.2

or

$$k+1 \leqq d_G(x_m, x_{j+1}) \leqq d_G(x_m, w) + d_G(w, v) + d_G(x_{j+1}, v).$$

By Claim 1, $d_G(x_i, u) = d_G(x_{j+1}, v)$, so we have

$$k - d_G(x_m, w) \leqq d_G(x_i, u) = d_G(x_{j+1}, v) \tag{2.4}$$

in either case. Replace x_m by x_{m+1} in the above arguments, then

$$k - d_G(x_{m+1}, w) \leqq d_G(x_i, u) = d_G(x_{j+1}, v). \tag{2.5}$$

Summing (2.4) and (2.5) and using $d_G(x_m, w) + d_G(x_{m+1}, w) = d_G(x_m, x_{m+1}) \leqq k$ yields

$$k/2 \leqq d_G(x_i, u) = d_G(x_{j+1}, v). \tag{2.6}$$

Now choose w' in C_2 adjacent to both u and v, and repeat the above process to obtain

$$k/2 \leqq d_G(x_{i+1}, u) = d_G(x_j, v). \tag{2.7}$$

Claim 3 follows from (2.6) and (2.7).

Suppose $u = p_i \cap \text{Bet}(w_i, 1, w_{i+1})$. By Lemma 2.5, there exists a vertex v in C' that is adjacent to both w_i and u. Assume $v \in p_j$ and note that $j \neq i$. Since (w_i, v) is a chord of C' such that $v \in p_j$ and $w_i \in p_{i-1} \cap p_i$, and by Claim 2 $d_G(w_i, x_i) < k/2$, Claim 3 implies that j is near to both $i-1$ and i. Then $n \geqq 4$ implies $j \in \{i-1, i\}$. Since $j \neq i$, we have $j = i-1$.

Choose $z_{i-1}(z_i)$ between w_i and w_{i-1} (w_{i+1}) such that $(z_{i-1}, z_i) \in E$ (such an edge exists, e.g. (v, u)) and $d_G(w_i, z_{i-1}) + d_G(w_i, z_i)$ is as large as possible. In the cycle $C'' = (w_1, \cdots, w_{i-1}, \cdots, z_{i-1}, z_i, \cdots, w_{i+1}, \cdots, w_1)$, by Lemma 2.5, there is a vertex w adjacent to z_{i-1} and z_i. Assume $w \in p_m$. By the choice of z_{i-1} and z_i, m is neither $i-1$ nor i. Hence m is not near to both $i-1$ and i. Assume, without loss of generality, that m is not near to i. By Claim 3,

$$d_G(x_i, z_i) = d_G(x_{i+1}, z_i) = d_G(x_m, w) = d_G(x_{m+1}, w) = k/2. \tag{2.8}$$

Thus $d_G(x_i, x_{i+1}) = k$ and z_i is equidistant from x_i and x_{i+1}. By (2.8),

$$k+1 \leqq d_G(x_i, x_m) \leqq d_G(x_i, z_{i-1}) + 1 + d_G(x_m, w) = d_G(x_i, z_{i-1}) + 1 + k/2$$

which implies

$$k/2 \leqq d_G(x_i, z_{i-1}). \tag{2.9}$$

Also

$$k+1 \leqq d_G(x_{i-1}, x_{i+1}) \leqq d_G(x_{i-1}, z_{i-1}) + 1 + d_G(x_{i+1}, z_i) = d_G(x_{i-1}, z_{i-1}) + 1 + k/2$$

which implies

$$(2.10) \qquad k/2 \leqq d_G(x_{i-1}, z_{i-1}).$$

Together (2.9) and (2.10) imply that $d_G(x_{i-1}, x_i) = k$ and $d_G(x_{i-1}, z_{i-1}) = d_G(x_i, z_{i-1}) = k/2$. Since this is true for all i, (1) and (2) of the lemma hold.

To prove (3), for each i consider the cycle $C_i = (z_{i-1}, \cdots$ along $p_{i-1} \cdots, w_i, \cdots$ along $p_i \cdots, z_i, z_{i-1})$. By Lemma 2.5, there is a vertex y_i of C_i that is adjacent to both z_{i-1} and z_i. This implies that $d_G(x_i, y_i) = k/2 - 1$.

Now $\{y_1, \cdots, y_n\}$ is a stable set in G, since if not, then there are $i \neq j$ such that $(y_i, y_j) \in E$, which would imply $d_G(x_i, x_j) \leqq k-1$. Finally, y_i is not adjacent to z_j if $j \notin \{i-1, i\}$, otherwise $d_G(x_i, x_j) \leqq k$ and (x_i, x_j) would be a chord of H. So $\{y_1, \cdots, y_n, z_1, \cdots, z_n\}$ induces an n-sun. □

Proof of Theorem 2.2 *from Lemma* 2.6.

(1⇒2). This is clear.

(2⇒3). Suppose G and G^2 are chordal. If G has an induced n-sun on $\{y_1, \cdots, y_n, z_1, \cdots, z_n\}$, then $C = (y_1, \cdots, y_n, y_1)$ is a cycle in G^2. So C has a chord (y_i, y_j), which implies that $d_G(y_i, y_j) = 2$ for some i and j that are not near to each other.

(3⇒1). Suppose G^k has a hole $H = (x_1, \cdots, x_n, x_1)$ of length $n \geqq 4$. Then (1) and (3) of Lemma 2.6 hold and there are i and j not near to each other such that $d_G(y_i, y_j) = 2$. Thus

$$d_G(x_i, x_j) \leqq d_G(x_i, y_i) + d_G(y_i, y_j) + d_G(y_j, x_j) \leqq k/2 - 1 + 2 + k/2 - 1 = k.$$

This means $(x_i, x_j) \in E^k$, which contradicts the fact that H is a hole in G^k. □

3. 3-sun-free chordal graphs—graphs with "minimum radius". The *diameter* of a graph G is defined by $d(G) = \max_{x,y \in V} d(x, y)$. The *radius* of G is $r(G) = \min_{x \in V} e(x)$, where $e(x) = \max_{y \in V} d(x, y)$, and a *center* is a vertex x such that $e(x) = r(G)$. For any connected graph G, $d(G)$ and $r(G)$ are finite and $r(G) \leqq d(G) \leqq 2r(G)$. In this section, we will prove that 3-sun-free chordal graphs are exactly those graphs that satisfy (P4) and that (P4) implies (P3).

LEMMA 3.1. *If G is chordal and $d(x, y) = k$, then* Bet (x, n, y) *is a clique for* $0 \leqq n \leqq k$.

Proof. Since Bet $(x, 0, y) = \{x\}$ and Bet $(x, k, y) = \{y\}$ are cliques, we can assume $1 \leqq n \leqq k-1$. In this case, x and y are not adjacent. For any path $p_{xy} = (x = x_0, x_1, \cdots, x_m = y)$ from x to y, choose i as large as possible such that $d(x_0, x_i) \leqq n$. Then $d(x_0, x_i) = n$, otherwise $d(x_0, x_i) \leqq n-1$ implies that $d(x_0, x_{i+1}) \leqq d(x_0, x_i) + 1 \leqq n$, which contradicts the choice of i. So p_{xy} contains some vertex x_i in Nbd (x, n). This proves that Nbd (x, n) is an x–y separator. For any z in Bet (x, n, y), a shortest path from x to y intersects Nbd (x, n) only at z. So any sub-separator of Nbd (x, n), in particular a minimal one, contains Bet (x, n, y). Hence, by Theorem 2.1, Bet (x, n, y) is a clique. □

The following lemma is a simple consequence of Laskar and Shier [1981, Lemma 1(d)].

LEMMA 3.2. *In a chordal graph G, if C is a clique and x is a vertex not in C such that $d(x, y) = k$ is a constant for all $y \in C$, then $\bigcap_{y \in C}$* Bet $(y, 1, x)$ *is not empty.*

LEMMA 3.3. *Suppose $G = (V, E)$ is a chordal graph on the six vertices $\{y_1, y_2, y_3, z_1, z_2, z_3\}$. If $\{z_1, z_2, z_3\}$ is a clique and $(y_i, z_j) \in E$ if and only if $i + 1 \neq j \pmod 3$, then $\{y_1, y_2, y_3\}$ is a stable set and hence G is a* 3-*sun.*

Proof. If y_1 were adjacent to y_2, then the cycle $(y_1, y_2, z_2, z_3, y_1)$ would be a hole, which is a contradiction. Similarly, y_3 is not adjacent to y_1 and y_2. □

LEMMA 3.4. *Suppose G is connected and chordal. If x is a center, $y \in \text{Nbd}(x, r(G))$, $u \in \text{Bet}(x, 1, y)$, and y' is such that $d(u, y') > d(x, y') \geqq r(G) - 1$, then $d(y, y') \geqq 2r(G) - 2$. If $d(y, y') = 2r(G) - 2$, then there exists $u' \in \text{Bet}(x, 1, y)$ such that $d(x, y') = d(u', y') = r(G) - 1$.*

Proof. Choose a shortest path p_{xy} from x to y containing u. Next, choose a shortest path $p_{yy'}(p_{xy'})$ from $y(x)$ to y' such that $p_{xy}(p_{xy'})$ intersects $p_{yy'}$ on a path from y to w (from y' to w'); see Fig. 3.1. Without loss of generality, we can assume $d(w, w')$ is as small as possible. If either w or w' is x, then $d(y, y') = d(y, x) + d(x, y') \geqq 2r(G) - 1$ and the lemma follows. Thus suppose that neither w nor w' is x. Let $v \in \text{Bet}(x, 1, y') \cap p_{xy'}$. (It is possible that $v = w'$.) u cannot be adjacent to v, otherwise $d(u, y') \leqq d(u, v) + d(v, y') = 1 + d(v, y') = d(x, y')$, which contradicts the assumption $d(u, y') > d(x, y')$. In the cycle $C = (x, v, \cdots, w', \cdots, w, \cdots, u, x)$, by Lemma 2.5, there is a vertex u' adjacent to both x and v. Since u is not adjacent to v, $u' \neq u$ and hence u' is strictly between w and w' as in Fig. 3.1. By the definition of p_{xy}, we have

$$(3.1) \qquad d(w, u') + 1 \geqq d(w, x), \quad \text{i.e.,} \quad d(y, u') + 1 \geqq d(y, x).$$

Since $d(u', w') < d(x, w')$ would imply that we can take the shortest path $(x, u', \cdots, w', \cdots, y')$ from x to y' to shorten $d(w, w')$, we also have

$$(3.2) \qquad d(u', w') \geqq d(x, w'), \quad \text{i.e.,} \quad d(u', y') \geqq d(x, y').$$

Together (3.1) and (3.2) imply that

$$(3.3) \qquad d(y, y') \geqq d(y, x) + d(x, y') - 1 \geqq 2r(G) - 2.$$

This proves the first part of Lemma 3.4.

If $d(y, y') = 2r(G) - 2$, then (3.1), (3.2), and (3.3) are equalities. Thus $u' \in \text{Bet}(x, 1, y)$ and $d(x, y') = d(u', y') = d(y, y') - d(y, u') = r(G) - 1$. □

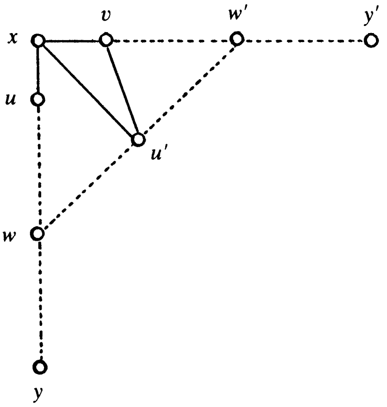


FIG. 3.1

Laskar and Shier [1982] proved that $d(G) \geqq 2r(G) - 3$ for any connected chordal graph. We give the following stronger result.

THEOREM 3.5. *$d(G) \geqq 2r(G) - 2$ for any connected chordal graph G.*

Proof. Choose a center x such that $|\text{Nbd}(x, r(G))|$ is as small as possible. Let $y \in \text{Nbd}(x, r(G))$ and $u \in \text{Bet}(x, 1, y)$. Consider the set S of all vertices w such that either $d(x, w) \leqq r(G) - 2$ or $d(x, w) \geqq d(u, w)$. Suppose $S = V$, then u is a center such that $|\text{Nbd}(u, r(G))| < |\text{Nbd}(x, r(G))|$, which is a contradiction. So there exists some

y' not in S. By the definition of S, $d(u, y') > d(x, y') \geqq r(G) - 1$. Now Theorem 3.5 follows from Lemma 3.4. □

THEOREM 3.6. *The following statements are equivalent for any graph G.*

(1) $r(H) = \lceil d(H)/2 \rceil$ *for every connected induced subgraph H of G.*

(2) *G is 3-sun-free chordal.*

Proof. ($1 \Rightarrow 2$). Suppose G has a hole H of length n; then H induces a subgraph with $r(H) = d(H) = \lfloor n/2 \rfloor$. Thus (1) implies $r(H) = d(H) = 1$ and $n = 3$, i.e. G is chordal. G does not have a 3-sun as an induced subgraph since r(3-sun) = d(3-sun) = 2.

($2 \Rightarrow 1$). Let $H = (V, E)$ be a connected induced subgraph of G. We will prove that $d(H) \geqq 2r(H) - 1$ by a method similar to the one used to prove Theorem 3.5.

Choose a center x such that $|\text{Nbd}(x, r(H))|$ is as small as possible. Let $y \in \text{Nbd}(x, r(H))$ and choose $u \in \text{Bet}(x, 1, y)$ such that the set $S(u) = \{w | d(u, w) \leqq d(x, w)$ or $d(x, w) \leqq r(H) - 2\}$ is as large as possible. Suppose $S(u) = V$, then u is a center such that $|\text{Nbd}(u, r(H))| < |\text{Nbd}(x, r(H))|$, which is impossible. So there is some $y' \notin S(u)$, i.e. $d(u, y') > d(x, y') \geqq r(H) - 1$. By Lemma 3.4., either $d(y, y') \geqq 2r(H) - 1$ and hence $d(H) \geqq 2r(H) - 1$ so that Theorem 3.6 is true, or else $d(y, y') = 2r(H) - 2$ and there exists some $u' \in \text{Bet}(x, 1, y)$ such that $d(x, y') = d(u', y') = r(H) - 1$. In the later case, Lemma 3.1 implies $(u, u') \in E$. By Lemma 3.2, there is $v \in \text{Bet}(u, 1, y) \cap \text{Bet}(u', 1, y)$ and $v' \in \text{Bet}(x, 1, y') \cap \text{Bet}(u', 1, y')$ as in Fig. 3.2. Note that $(x, v) \notin E$; also $(u, v') \notin E$, since $(u, v') \in E$ implies $d(u, y') \leqq d(x, y')$, which contradicts $y' \notin S(u)$.

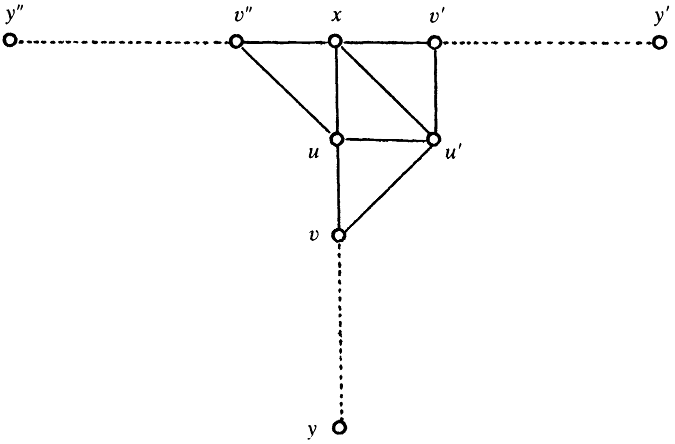


FIG. 3.2

Next we prove that $S(u') \supseteq S(u)$. Suppose not; then there is some $y'' \in S(u) \backslash S(u')$. Since $d(u', y'') > d(x, y'') \geqq r(H) - 1$, by Lemma 3.4, either $d(y, y'') \geqq 2r(H) - 1$ and hence $d(H) \geqq 2r(H) - 1$ so that Theorem 3.6 is true, or else $d(y, y'') = 2r(H) - 2$ and there is some $u'' \in \text{Bet}(x, 1, y)$ such that $d(u'', y'') = d(x, y'') = r(H) - 1$. Note that $d(u', y'') > r(H) - 1$ and $(u, u') \in E$ imply $d(u, y'') \geqq r(H) - 1$. On the other hand, $y'' \in S(u)$ implies that $d(u, y'') \leqq d(x, y'') = r(H) - 1$. So $d(u, y'') = d(x, y'') = r(H) - 1$. By Lemma 3.2, there is some $v'' \in \text{Bet}(x, 1, y'') \cap \text{Bet}(u, 1, y'')$. But $(v'', u') \notin E$, otherwise $d(u', y'') \leqq r(H) - 1$, which is impossible. So by Lemma 3.3, $\{x, u, u', v'', v, v'\}$ induces a 3-sun, which contradicts our assumption. This proves $S(u') \supseteq S(u)$. But $y' \in S(u') \backslash S(u)$, so that $|S(u')| > |S(u)|$, which contradicts the choice of u. So Theorem 3.6 holds. □

LEMMA 3.7. *If G is chordal and S is a maximal clique of G^k, then the induced subgraph G_S is connected and $d_G(x, y) = d_{G_S}(x, y)$ for all $x, y \in S$.*

Proof. We will prove that if x and y are in S, then all the vertices on any shortest path from x to y in G are in S, which implies the lemma. Suppose p_{xy} is a shortest path from x to y that contains a vertex z not in S. For all other paths from x to y in G, there exists at least one vertex not in p_{xy}. Let T be the set of all such vertices plus z, then T is an x–y separator. Choose a minimal x–y separator $T' \subseteq T$, then T' is a clique of G by Theorem 2.1. Also, $x \in T'$ since p_{xy} has exactly one vertex in T. For any vertex w in S suppose, without loss of generality, that w is not in the connected component of $G \backslash T'$ containing x. Choose a path $p_{xw} = (x, \cdots, u, \cdots, w)$ of length not greater than k from x to w in G, where $u \in T'$ (it is possible that $u = w$). Now $(z, u, \cdots, w)$ is a path from z to w of length not greater than k, so $d_G(z, w) \leqq k$ and $(z, w) \in E^k$. Thus $S \cup \{z\}$ is a clique of G^k, which contradicts the fact that S is a maximal clique in G^k. Thus all vertices of p_{xy} are in S. □

THEOREM 3.8. *Suppose G is 3-sun-free chordal and k a positive integer, then S is a clique in G^{2k} if and only if there is some x such that $d_G(x, y) \leqq k$ for all y in S.*

Proof. Suppose S is a clique in G^{2k}. Without loss of generality we can assume it is maximal. By Lemma 3.7, S induces a connected subgraph H such that $d_H(x, y) = d_G(x, y)$ for all $x, y \in S$. Hence $d(H) \leqq 2k$. Since G is 3-sun-free chordal, by Theorem 3.6, $r(H) = \lceil d(H)/2 \rceil \leqq k$. Thus there is some $x \in S$ such that $d_G(x, y) = d_H(x, y) \leqq k$ for all y in S.

The converse is obvious. □

4. Duality and algorithms for sun-free chordal graphs. From Corollary 2.3 and Theorem 3.8, we know that (P2) and (P3) of § 1 hold for any induced subgraph of a sun-free chordal graph. Thus we obtain the following duality between k-domination and $2k$-stability.

THEOREM 4.1. *If G is sun-free chordal, then $\delta_k(H) = \alpha_{2k}(H)$ for any positive integer k and any induced subgraph H of G.*

Farber [1981] has proved Theorem 4.1 for $k = 1$. He also gives several characterizations of sun-free chordal graphs. One of these yields a polynomial-time test to determine if a graph is sun-free chordal.

Examples of sun-free chordal graphs are total graphs of trees, line graphs of trees, directed path graphs, which include interval graphs, and powers of block graphs (Farber [1981]).

Because of (P2) and (P3), the k-domination problem on a sun-free chordal graph G is equivalent to the clique covering problem on the chordal graph G^{2k}. Hence it can be solved in polynomial time, by Gavril's [1972] clique covering algorithm for chordal graphs. The dominant step is the construction of G^{2k}, which takes $O(|V|^3)$ time.

Farber [1981], Kolen [1982] and Lubiw [1982] have given polynomial-time algorithms for the weighted 1-domination problem on sun-free chordal graphs. Chang [1982] observed that these algorithms could be used to solve the weighted k-domination problem on sun-free chordal graphs if powers of sun-free chordal graphs are sun-free chordal. Lubiw [1982] proved this result.

Because of (P1), the weighted k-stability problem on G can be reduced to the 1-stability problem on G^k. If G^k is chordal, we can use Frank's linear algorithm [1975] for the weighted 1-stability problem on chordal graphs to solve the weighted k-stability problem on G. Examples of such graphs are sun-free chordal graphs for even k and chordal graphs for odd k.

5. NP-completeness of the k-domination and k-stability problems. It is easy to see that the k-domination and k-stability problem are NP-complete on general graphs; see Garey and Johnson [1979] for the case of $k=1$. In this section, we will prove that they are NP-complete for bipartite graphs (except for the 1-stability problem), and for chordal graphs (except for the k-stability problem with odd k).

THEOREM 5.1. *For any fixed positive integer k, the k-domination problem is* NP-*complete for bipartite graphs.*

Proof. We prove the theorem for $k=1$ by giving a polynomial time reduction of the 1-domination problem on a general graph to one on a bipartite graph.

For any graph $G=(V,E)$ and $s\notin V$, construct the bipartite graph $G'=(V_1'\cup V_2', E')$, where $V_1'=V\cup\{s\}$, $V_2'=\{v'|v\in V_1'\}$, and $E'=\{(x,y')|x,y\in V$ and $d_G(x,y)\leqq 1\}\cup\{(v,s')|v\in V_1'\}$. An example of the transformation is given in Fig. 5.1.

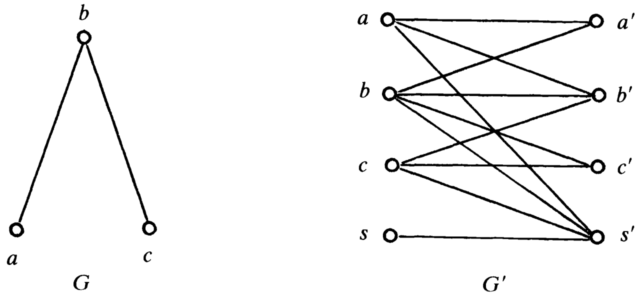


FIG. 5.1

We will prove that $\delta_1(G)+1=\delta_1(G')$. Suppose D is a 1-dominating set of G. Then $D\cup\{s'\}$ is a 1-dominating set of G' since s' dominates V_1' and D dominates V_2'. So $\delta_1(G)+1\geqq\delta_1(G')$.

Suppose D' is a minimum cardinality 1-dominating set of G'. D' must contain s or s' since s is adjacent only to s'. Without loss of generality, we can assume $s'\in D'$, otherwise we can use $D'\cup\{s'\}\backslash\{s\}$ instead of D'. Then $s'\in D'$ implies that if $v'\in D'$ for some $v\in V$, then $D'\cup\{v\}\backslash\{v'\}$ is also a 1-dominating set of G'. Therefore we can assume $D'=D\cup\{s'\}$ with $D\subseteq V$. For every $u\in V$, there is a $v\in D'$ such that $(v,u')\in E'$. Since $(s',u')\notin E'$, we have $v\in D$ and hence $d_G(u,v)\leqq 1$. This proves that D is a 1-dominating set of G, hence $\delta_1(G)+1\leqq\delta(G')$.

Since the 1-domination problem is NP-complete for general graphs, it is also NP-complete for bipartite graphs.

For general k, attach a path of length $k-1$ to each vertex in $\{v'|v\in V\}\cup\{s\}$ to obtain the bipartite graph G''. Then we can prove that $\delta_1(G)+1=\delta_k(G'')$ so that the k-domination problem on bipartite graphs is NP-complete. □

THEOREM 5.2. *For any fixed positive integer $k\geqq 2$, the k-stability problem is* NP-*complete for bipartite graphs.*

Proof. We will give a polynomial reduction of the 1-stability (vertex packing) problem on a general graph to a k-stability problem on a bipartite graph.

For any graph $G=(V,E)$, construct a graph $G'=(V',E')$ by replacing each edge $e=(u,v)$ of G by the tree T_e (see Fig. 5.2), in which $m=\lfloor k/2\rfloor\geqq 1$. Since every cycle in G' is of even length, G' is bipartite.

We will prove that $\alpha(G)+|E|=\alpha_k(G')$. Suppose S is a stable set in G, then $S\cup E$ is a k-stable set in G' since (i) for any $x\in V'$, $d_{G'}(e,x)\leqq k$ implies x is a vertex of $T_e\backslash\{u,v\}$, and (ii) $d_{G'}(u,v)\geqq 4m>k$ for any u, $v\in S$. Thus $\alpha(G)+|E|\leqq\alpha_k(G')$.

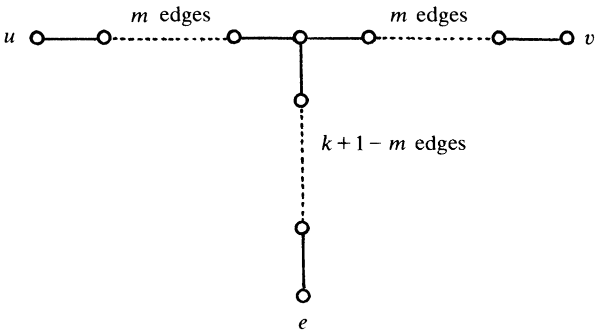


FIG. 5.2. *Tree T_e with u, v, e as leaves.*

Conversely, suppose S' is a maximum cardinality k-stable set in G'. For each $e=(u, v)\in E$, if S' contains no vertex of $T_e\backslash\{u, v\}$, then $S'\cup\{e\}$ is a k-stable set of G', which contradicts the optimality of S'. Furthermore, no two vertices of $T_e\backslash\{u, v\}$ can be in a k-stable set. So S' contains exactly one vertex of $T_e\backslash\{u, v\}$. Without loss of generality, we can assume it is e, i.e. $S'=S\cup E$ where $S\subseteq V$. Now S is stable in G, otherwise there are u, $v\in S$ and $(u, v)\in E$ so that $d_{G'}(u, v)=2m\leqq k$, which contradicts the k-stability of S'. So $\alpha(G)+|E|\geqq\alpha_k(G')$.

Since the 1-stability problem is NP-complete on general graphs, the k-stability problem is NP-complete for bipartite graphs. □

Booth [1980] proved the NP-completeness of the 1-domination problem on chordal graphs. In fact, his proof showed that the 1-domination problem is NP-complete on the subclass of chordal graphs called split graphs. A graph is *split* if its vertex set can be partitioned into a stable set and a clique. Booth and Johnson [1982] also proved that the 1-domination problem is NP-complete on undirected path graphs, which is another subclass of chordal graphs. More generally, we have:

THEOREM 5.3. *The k-domination and $2k$-stability problems are* NP-*complete for chordal graphs of diameter* $2k+1$.

Proof. We will give a polynomial reduction of the 1-domination problem on general graphs to the k-domination problem on chordal graphs.

For any graph $G=(V, E)$, consider the chordal, in fact split, graph $G'=(V\cup V', E)$ with $V'=\{v'|v\in V\}$ and $E'=\{(u, v)|u, v\in V \text{ and } u\neq v\}\cup\{(u, v')|u, v\in V$ with $d_G(u, v)\leqq 1\}$. An example of the transformation is given in Fig. 5.3. Now attach a path $(v'=v_1, v_2, \cdots, v_k)$ of length $k-1$ to each vertex v' in G' to get the chordal graph G''. Note that $d(G')=3$ and $d(G'')=2k+1$ unless $d(G)\leqq 2$.

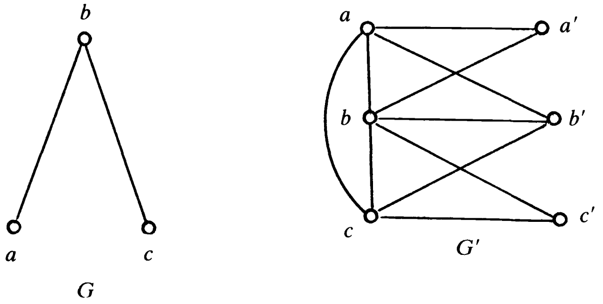


FIG. 5.3

We will prove that $\delta_1(G)=\delta_k(G'')$. Any 1-dominating set D of G is a k-dominating set of G'' since for any $u\in V$, there is some $v\in D$ such that $d_G(u,v)\leqq 1$ and so $d_{G''}(u,v)\leqq 1$ and $d_{G''}(u_i,v)=i\leqq k$ for $1\leqq i\leqq k$. Hence $\delta_1(G)\geqq\delta_k(G'')$.

If D'' is a minimum cardinality k-dominating set of G'', then $D=\{v\in V|v$ or $v_i\in D''\}$ is also a minimum cardinality k-dominating set of G'' since v dominates every vertex dominated by v_i in G'' and $|D|\leqq|D''|$. For any $u\in V$, there is some $v\in D$ such that $d_{G''}(u_k,v)\leqq k$ and so $d_G(u,v)\leqq 1$. Thus D is a 1-dominating set of G. This proves $\delta_1(G)\leqq\delta_k(G'')$ and completes the proof for the k-domination problem.

The proof for the $2k$-stability problem is similar. □

6. Conclusions. We have studied duality between the k-domination and $2k$-stability problems for graphs satisfying (P2) and (P4) of § 1 and have characterized these graphs. However, condition (P2) is unnecessarily restrictive since we only require G^k to be perfect rather than chordal to obtain the duality of Theorem 4.1. In a sequel to this paper (Chang and Nemhauser [1982a]), we will establish this duality for chordal graphs without 3-suns and complements of 3-suns. We will also prove $\alpha_2(G)=\delta_1(G)$ for odd-sun-free chordal graphs and that the "strong perfect graph conjecture" implies Theorem 4.1 for odd sun-free chordal graphs.

Acknowledgment. We are grateful to the referees for providing several helpful suggestions.

REFERENCES

H. L. ABBOTT AND A. C. LIU, (1979), *Bounds for the covering number of a graph*, Discrete Math., 25, pp. 281–284.

R. BALAKRISHNAN AND P. PALURAJA, (1982), *Graphs whose squares are chordal*, Australian J. Math., to appear.

C. BERGE, (1973), *Graphs and Hypergraphs*, American Elsevier, New York.

K. S. BOOTH, (1980), *Dominating sets in chordal graphs*, CS-80-34, Univ. Waterloo, Waterloo, Ontario.

K. S. BOOTH AND J. H. JOHNSON, (1982), *Dominating sets in chordal graphs*, SIAM J. Comput., 11, pp. 191–199.

P. BUNEMAN, (1974), *A characterization of rigid circuit graphs*, Discrete Math., 9, pp. 205–212.

G. J. CHANG, (1982), *k-domination and graph covering problems*, Ph.D. thesis, School of OR & IE, Cornell University, Ithaca, NY.

G. J. CHANG AND G. L. NEMHAUSER, (1982), *R-domination on block graphs*, Oper. Res. Lett., 1, pp. 214–218.

———, (1982a), *Covering, packing and generalized perfection*, Tech. Rep. No. 551, School of OR & IE, Cornell University, Ithaca, NY.

E. J. COCKAYNE, (1978), *Domination of undirected graphs—A survey*, in Theory and Application of Graphs, Lecture Notes in Mathematics 642, Springer, Berlin, pp. 141–147.

E. J. COCKAYNE, S. GOODMAN AND S. T. HEDETNIEMI, (1975), *A linear algorithm for the domination number of a tree*, Inform. Proc. Letters, 4, pp. 41–44.

G. A. DIRAC, (1961), *On rigid circuit graphs*, Abh. Math. Sem. Univ. Hamburg, 25, pp. 71–76.

P. DUCHET, (1982), *Classical perfect graphs—An introduction with emphasis on triangulated and interval graphs*, to appear in Topics on Perfect Graphs, C. Berge and V. Chvatal, eds., North-Holland, Amsterdam.

M. FARBER, (1981), *Applications of linear programming duality to problems involving independence and domination*, TR 81-13, Dept. of Computer Science, Simon Fraser Univ., Burnaby, British Columbia, Canada.

A. FRANK, (1975), *Some polynomial algorithms for certain graphs and hypergraphs*, Proc. 5th British Comb. Conf., pp. 211–226.

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

M. R. GAREY AND D. S. JOHNSON, (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

F. GAVRIL, (1972), *Algorithms for minimum covering, maximum clique, minimum covering by cliques, and maximum independent set of chordal graphs*, SIAM J. Comput. 1, pp. 180–187.

———, (1974), *The intersection graphs of subtrees in trees are exactly the chordal graphs*, J. Comb. Theory, B16, pp. 47–56.

M. C. GOLUMBIC, (1980), *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York.

C. JORDAN, (1869), *Sur les assemblages de lignes*, J. Reine Angew. Math., 70, pp. 185–190.

O. KARIV AND S. L. HAKIMI, (1979), *An algorithmic approach to network location problems* I. *The p-centers*, SIAM J. Appl. Math., 37, pp. 513–538.

A. W. J. KOLEN, (1982), *Location problems on trees and the rectilinear plane*, Ph.D. thesis, University of Amsterdam, Amsterdam.

D. KÖNIG, (1950), *Theorie der Endlichen und Unendlichen Graphen*, Chelsea Publ., New York.

R. LASKAR AND D. SHIER, (1980), *On chordal graphs*, Proc. Eleventh Southeastern Conf. on Comb., Graph Theory and Computing, Utilitas Math., Winnipeg, pp. 579–588.

———, (1981), *Construction of (r, d)-invariant chordal graphs*, Congressus Numerantium, 33, pp. 155–165.

———, (1982), *On powers and centers of chordal graphs*, Disc. Appl. Math., to appear.

C. LIU, (1968), *Introduction to Combinatorial Mathematics*, McGraw-Hill, New York.

A. LUBIW, (1982), *Γ-free matrices*, M.S. thesis, Dept. Combinatorics and Optimization, Univ. Waterloo, Waterloo, Ontario.

A. MEIR AND J. W. MOON, (1975), *Relation between packing and covering of a tree*, Pacific J. Math., 61, pp. 225–233.

K. S. NATARAJAN AND L. J. WHITE, (1978), *Optimum domination in weighted trees*, Inform. Proc. Letters, 7, pp. 261–265.

P. J. SLATER, (1976), *R-domination in graphs*, J. Assoc. Comp. Mach., 23, pp. 446–450.

J. R. WALTER, (1972), *Representation of rigid cycle graphs*, Ph.D. thesis, Wayne State Univ., Detroit.

———, (1978), *Representation of chordal graphs as subtrees of a tree*, J. Graph Theory, 2, pp. 265–267.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

006

ON THE CONTROLLABILITY OF MATRIX PAIRS (A, K) WITH K POSITIVE SEMIDEFINITE*

DAVID CARLSON†, B. N. DATTA‡ AND HANS SCHNEIDER§

Dedicated to Emilie V. Haynsworth

Abstract. The controllability of matrix pairs (A, K) is studied when K is positive semi-definite, and in particular when K is in the range of the Lyapunov map determined by A. This extends previous work of Chen, Wimmer, Carlson and Loewy, and Coppel.

Key words. controllability, Lyapunov matrix maps

AMS 1975 subject classifications. 15A24, 15A18

1. Introduction.

This note is devoted to the study of the controllability of (A, K), where $A \in C^{n,n}$, $K \in H_n$ (the set of hermitian matrices in $C^{n,n}$), and K is positive semidefinite (which we shall write as $K \geqq 0$). A well-known result, proved independently by Chen [5] and Wimmer [11], states:

THEOREM 1. *Let $A \in C^{n,n}$, and suppose that $K = AH + HA^* \geqq 0$ for some H, $K \in H_n$. If (A, K) is controllable, then A has no eigenvalues on the imaginary axis and H is nonsingular (and, in fact, the numbers of eigenvalues of A with positive and negative real parts equal respectively the numbers of positive and negative eigenvalues of H).*

Using Theorem 1, Wimmer extended previous results in the damping of certain quadratic differential equations involved in linear vibration problems.

An example [3, p. 240] shows that the converse of Theorem 1 is false. However, working independently of Chen and Wimmer, Carlson and Loewy [3] established a converse under an additional hypothesis:

THEOREM 2. *Let $A \in C^{n,n}$, such that $\lambda + \bar{\mu} \neq 0$ for all eigenvalues λ, μ of A. Suppose that $K = AH + HA^* \geqq 0$ for some H, $K \in H_n$. Then the following are equivalent:*

(i) *(A, K) is controllable.*

(ii) *H is nonsingular.*

The question thus arose as to the role of the additional hypothesis of Theorem 2 in a more complete converse of Theorem 1. We answer this question by proving a result (Theorem 4) which will yield, under $K = AH + HA^* \geqq 0$, a condition equivalent to the controllability of (A, K) in terms of the spectrum of A and the nonsingularity of a matrix $\hat{H}$ determined by A and H. The matrix $\hat{H}$ is obtained from H via projections associated with an A-modal decomposition of C^n; see § 2 for definitions. Our proof of this result will use Theorem 1 and a result in [3] preliminary to Theorem 2; the result itself contains Theorem 2 as a special case.

As a consequence of Theorem 4 we will be able to discuss special cases (like that in Theorem 2) in which $\hat{H}$ may be replaced by H, that is, for which (i) and (ii) above are equivalent. This clarifies (see also [7]) Coppel's discussion in [6] of the relationship between dichotomies for linear differential equations and Lyapunov functions in the constant-coefficient case. Coppel's work, along with that of Chen, Wimmer, and Carlson and Loewy, has motivated our investigations.

* Received by the editors December 31, 1982, and in revised form June 8, 1983.

† Mathematical Sciences Department, San Diego State University, San Diego, California 92182. Permanent address, Mathematics Department, Oregon State University, Corvallis, Oregon 97331.

‡ Mathematical Sciences Department, Northern Illinois University, DeKalb, Illinois 60115.

§ Mathematics Department, University of Wisconsin, Madison, Wisconsin 53706. The research of this author was partly supported by the National Science Foundation under grant MCS 80-26132.

2. Definitions. So that our decompositions depend only on the spaces involved and not particular choices of bases for the spaces, we will set our results in an equivalent but seemingly more abstract setting. Let V be a finite-dimensional inner product space, and let $L(V)$, $H(V)$ be respectively the sets of linear operators and self-adjoint linear operators on V. $K \in H(V)$ is positive semi-definite iff $(x, Kx) \geqq 0$ for all $x \in V$.

Let $A \in L(V)$ have spectrum $\sigma(A) = \{\lambda_1, \cdots, \lambda_n\}$; let $\delta(A)$ be the number of eigenvalues λ_i which are imaginary, and let $\Delta(A) = \prod_{i,j=1}^{n} (\lambda_i + \bar{\lambda}_j)$. Evidently $\Delta(A) \neq 0$ is equivalent to $\sigma(A) \cap \sigma(-A^*) = 0$, where A^* is the adjoint of A, and $\delta(A) = 0$ is equivalent to $\sigma(A) \cap iR = 0$. Thus $\Delta(A) \neq 0$ implies that $\delta(A) = 0$; the converse is false. We denote the kernel and image of A by Ker A and Im A respectively, and the rank of A by $\rho(A)$.

Let $A, B \in L(V)$; the *control space* of (A, B) is $C(A, B) = \sum_{r=0}^{\infty} \operatorname{Im} A^r B$, the smallest A-invariant space containing Im B. Note that $C(A, B)$ depends only on A and Im B and that (because of the Cayley–Hamilton Theorem),

$$C(A, B) = \sum_{r=0}^{n-1} \operatorname{Im} A^r B.$$

The pair (A, B) is said to be *controllable* if $C(A, B) = V$.

For $A \in L(V)$, we may decompose V (generally in a number of ways) as $V = V_1 \oplus \cdots \oplus V_p$, so that each V_j is A-invariant, and so that the restrictions $A|V_j$ of A to distinct V_j have disjoint spectra. Following Wonham [12, p. 18], we call such decompositions *A-modal*. In the finest A-modal decomposition of V the V_j are the generalized eigenspaces of the distinct eigenvalues of A. We call this the *A-spectral decomposition* of V. Another natural A-modal decomposition is obtained by choosing $V_1 = V_+$, $V_2 = V_-$, and $V_3 = V_0$, the direct sums of the generalized eigenspaces of the eigenvalues of A with, respectively, positive, negative, and zero real parts. We call this the *A-inertial decomposition* of V.

If $V = V_1 \oplus \cdots \oplus V_p$ is an A-modal decomposition of V, for $j = 1, \cdots, p$ we let E_j denote the projection in $L(V)$ onto V_j which annihilates $\sum_{i \neq j} V_i$. It is well-known that $\sum_{j=1}^{p} E_j = I$, that $E_i E_j = 0$, $i \neq j$, and that each E_j is a polynomial in A, cf. [8, p. 221]. Also, $V = W_1 \oplus \cdots \oplus W_p$, where each W_j is the range in V of the corresponding projection E_j^* in $L(V)$. For $j = 1, \cdots, p$, as V_j is A-invariant, we may set

$$A_j = E_j A E_j = A E_j = E_j A, \tag{1}$$

so that

$$A = \sum_{j=1}^{p} A_j, \tag{2}$$

and for $K \in H(V)$, we set

$$K_{jj} = E_j K E_j^*, \tag{3}$$

$$\hat{K} = \sum_{j=1}^{p} K_{jj}. \tag{4}$$

Note that $\rho(\hat{K}) = \sum_{j=1}^{p} \rho(K_{jj})$.

The restrictions of the linear operations A and A_j to V_j are equal, and the restrictions to W_j of the Hermitian forms induced by K and K_{jj} are equal: if $x, y \in W_j$,

$$(y, K_{jj}x) = (y, E_j K E_j^* x) = (E_j^* y, K E_j^* x) = (y, Kx).$$

3. Results.

LEMMA 1. *Let $A \in L(V)$, and suppose that $V = V_1 \oplus \cdots \oplus V_p$ is an A-modal decomposition. If $K \in H(V)$, with $K \geqq 0$, then $C(A, \hat{K}) = C(A, K)$.*

Proof. We observe that

$$C(A, \hat{K}) = C(A_1, K_{11}) \oplus \cdots \oplus C(A_p, K_{pp}) \qquad j = 1, \cdots, p, \tag{5}$$

since $A^r\hat{K} = \sum_{j=1}^p A_j^r K_{jj}$, and $\mathrm{Im}\, A^r K_{jj} \subseteq V_j$, $r = 0, \cdots, n-1$; it follows that $\mathrm{Im}\, A^r\hat{K} = \sum_{j=1}^r \mathrm{Im}\, A_j^r K_{jj}$. Thus to prove $C(A, K) \supseteq C(A, \hat{K})$, it is sufficient to show that $C(A, K) \supseteq C(A_j, K_{jj})$, $j = 1, \cdots, p$. Since E_j is a polynomial in A, we have

$$\mathrm{Im}\, (A_j^r K_{jj}) = \mathrm{Im}\, (E_j A^r E_j K E_j^*) \subseteq \mathrm{Im}\, (E_j A^r E_j K) \subseteq C(A, K)$$

and the inclusion follows. (We have not used $K \geqq 0$ here.)

To prove $C(A, K) \subseteq C(A, \hat{K})$, we first note the easily-proved result that $K \geqq 0$ implies that $\mathrm{Ker}\, \hat{K} \subseteq \mathrm{Ker}\, K$. It follows that $\mathrm{Im}\, K \subseteq \mathrm{Im}\, \hat{K}$ and hence $\mathrm{Im}\, A^r K \subseteq \mathrm{Im}\, A^r \hat{K}$, $r = 1, \cdots, n-1$. The result follows. □

THEOREM 3. *Let $A \in L(V)$, and suppose that $V = V_1 \oplus \cdots \oplus V_p$ is an A-modal decomposition of V. Suppose that $K \in H(V)$, with $K \geqq 0$. Then the following are equivalent*:

(i) *(A, K) is controllable.*

(ii) *$C(A_j, K_{jj}) = V_j$, $j = 1, \cdots, p$.*

(iii) *$(A\hat{K})$ is controllable.*

(iv) *$(x, Kx) > 0$ for every eigenvector x of A^*.*

Proof. The equivalence of (i), (ii), and (iii) follows immediately from (5) and Lemma 1. The equivalence of (i) and (iv) is a special case of [2, Lemma 3]. □

We remark that a related result which holds for all $K \in L(V)$ is known, cf. [12, p. 45, Exercise 1.5].

The equivalence of (i) and (iv) was noted in [3] (under the unnecessary assumption that $\Delta A \neq 0$); it is in fact merely a rephrasing of Hautus' criterion for controllability (cf. [11]) in the case that $K \geqq 0$. We cannot drop the condition $K \geqq 0$ from either Lemma 1 or Theorem 3: let $V = C^2 = V_1 \oplus V_2$, where V_1, V_2 are the coordinate subspaces, and let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix};$$

then (A, K) is controllable, but $C(A, K) = \{0\}$ since $\hat{K} = 0$.

In Theorem 3 we considered the controllability of pairs (A, K) where the only restriction on K is $K \geqq 0$. We shall now assume that $K = AH + HA^* \geqq 0$, where $H \in H(V)$. We note that the mapping $H \to AH + HA^*$ of $H(V)$ into itself is onto $H(V)$ if and only if $\Delta(A) \neq 0$.

Before stating Lemma 2, we must take care of a trivial but awkward technicality which we require to relate our results to Theorems 1 and 2. If $K = AH + HA^*$, note that $K_{jj} = A_j H_{jj} + H_{jj} A_j^*$ and indeed, for any $c \in C$, $K_{jj} = B_j H_{jj} + H_{jj} B_j^*$, where $B_j = A_j + c(\sum_{i \neq j} E_i)$. Then $\sigma(B_j) = \sigma(A|V_j) \cup \{c\}$. Hence, if $\Delta(A|V_j) \neq 0$, $j = 1, \cdots, p$, we may choose $c \in R$ so that $\Delta(B_j) \neq 0$, $j = 1, \cdots, p$.

LEMMA 2. *Let $A \in L(V)$, and let $V = V_1 \oplus \cdots \oplus V_p$ be an A-modal decomposition with $\Delta(A|V_j) \neq 0$, $j = 1, \cdots, p$. Let H, $K \in H_n$ with $K = AH + HA^* \geqq 0$. Then $C(A, K) = \mathrm{Im}\, \hat{H}$.*

Proof. For $j = 1, \cdots, p$, since $\Delta(A|V_j) \neq 0$, we choose $c \in R$ so that $B_j = A_j + c \sum_{i \neq j} E_i$ has $\Delta(B_j) \neq 0$. Also $C(B_j, K_{jj}) = C(A_j, K_{jj})$ and

$$B_j H_{jj} + H_{jj} B_j^* = A_j H_{jj} + H_{jj} A_j^* = K_{jj} \geqq 0.$$

By Corollary 2 of [3], then,

$$\operatorname{Im} H_{jj} = C(B_j, K_{jj}) = C(A_j, K_{jj}),$$

and the lemma follows by Lemma 1 and (5). □

THEOREM 4. *Let $A \in L(V)$ with $\delta(A) = 0$. Suppose that $V = V_1 \oplus \cdots \oplus V_p$ is an A-modal decomposition and that, for each $j = 1, \cdots, p$, $\Delta(A|V_j) \neq 0$. Let $K = AH + HA^* \geqq 0$ for $H, K \in H(V)$. The following are equivalent*:

(i) *(AK) is controllable.*

(ii) *$\hat{H}$ is nonsingular.*

(iii) *$(x, Hx) \neq 0$ for each eigenvector x of A^*.*

(iv) *$(x, \hat{H}x) \neq 0$ for each eigenvector x of A^*.*

(v) *H is nonsingular and $(A^*, H^{-1}K)$ is controllable.*

Proof. We note first that $\delta(A) = 0$ guarantees that there exists an A-modal decomposition $V = V_1 \oplus \cdots \oplus V_p$ for which $\Delta(A|V_j) \neq 0$, $j = 1, \cdots, p$.

The equivalence of (i) and (ii) follows immediately from Lemma 2.

To show that (i) and (iii) are equivalent, note that for any $x \in V$ for which $A^*x = \lambda x$,

$$(x, Kx) = (x, (AH + HA^*)x) = (A^*x, Hx) + (x, H(A^*x)) = (\bar{\lambda} + \lambda)(x, Hx),$$

and use condition (iv) from Theorem 3. To show that (iii) and (iv) are equivalent, we observe that if $A^*x = \lambda x$, then $x \in W_j$ for some j, $1 \leqq j \leqq p$, where W_j is defined in § 2. Hence $(x, Hx) = (x, H_{jj}x) = (x, \hat{H}x)$.

To show that (i) and (v) are equivalent, suppose that H is nonsingular. Then

$$A^*H^{-1} + H^{-1}A = H^{-1}KH^{-1},$$

and the equivalence follows easily from [1, Thm. 4]. □

We state the special case of A-inertial decomposition as a

COROLLARY 1. *Let $A \in L(V)$ and let $V = V_+ \oplus V_- \oplus V_0$ be the A-inertial decomposition of V. Suppose $K = AH + HA^* \geqq 0$ for some $H, K \in H(V)$. Then (A, K) is controllable if and only if $\delta(A) = 0$ and $\hat{H}$ is nonsingular.*

Proof. If $\delta(A) = 0$, then $V = V_+ \oplus V_-$ is an A-modal decomposition with $\Delta(A|V_+) \neq 0$, $\Delta(A|V_-) \neq 0$, and Theorem 4 applies.

If $\delta(A) \neq 0$, then by Theorem 1, (A, K) is not controllable. □

As an example, let $V = C^2$ and let

$$A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad K = AH + HA^* = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \geqq 0.$$

Here $\delta(A) = \delta(H) = 0$, yet $(A, \hat{K})$ is not controllable. We have

$$V_+ = \left\langle \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\rangle, \quad V_- = \left\langle \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\rangle, \quad E_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix},$$

and

$$\hat{H} = E_1 H E_1^* + E_2 H E_2^* = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

which is singular.

Finally, we observe that if $V = V_1 \oplus \cdots \oplus V_p$ is any A-modal decomposition and each V_j is also H-invariant (in particular, this is true if $p = 1$ or if A and H commute), then H commutes with all E_j (cf. [8, p. 221]) and

$$\hat{H}=\sum_{j=1}^{p} E_jHE_j^*=H\left(\sum_{j=1}^{p} E_jE_j^*\right).$$

It is easily shown that $\sum_{j=1}^{p} E_jE_j^*$ is nonsingular, so that H and $\hat{H}$ are singular or nonsingular together, and we may replace $\hat{H}$ by H in (ii) of Theorem 4. If, for example, all eigenvalues of A are known to have negative real part, then (A, K) is controllable if and only if H is nonsingular. This result is stated in [7].

REFERENCES

[1] DAVID CARLSON AND B. N. DATTA, *The Lyapunov matrix equation $SA+A^*S=S^*B^*BS$*, Lin. Alg. and Appl., 28 (1979), pp. 43–52.

[2] DAVID CARLSON AND RICHARD HILL, *Generalized controllability and inertia theory*, Lin. Alg. and Appl., 15 (1976), pp. 177–187.

[3] DAVID CARLSON AND RAPHAEL LOEWY, *On ranges of Lyapunov transformations*, Lin. Alg. and Appl., 8 (1974), pp. 237–248.

[4] DAVID CARLSON AND HANS SCHNEIDER, *Inertia theorems for matrices: the semidefinite case*, J. Math. Anal. Appl., 6 (1963), pp. 430–446.

[5] C. T. CHEN, *A generalization of the inertia theorem*, this Journal, 25 (1973), pp. 158–161.

[6] W. A. COPPEL, *Dichotomies in Stability Theory*, Lecture Notes in Mathematics 629, Springer-Verlag, Berlin, 1978.

[7] ———, *Dichotomies and Lyapunov functions*, J. Differential Equations, to appear.

[8] KENNETH HOFFMAN AND RAY KUNZE, *Linear Algebra*, second ed., Prentice-Hall, Englewood Cliffs, NJ, 1971.

[9] A. OSTROWSKI AND HANS SCHNEIDER, *Some theorems on the inertia of general matrices*, J. Math. Anal. Appl., 4 (1962), pp. 72–84.

[10] O. TAUSSKY, *A generalization of a theorem of Lyapunov*, J. Soc. Ind. Appl. Math., 9 (1961), pp. 640–643.

[11] HARALD WIMMER, *Inertia theorems for matrices, controllability, and linear vibrations*, Lin. Alg. and Appl., 8 (1974), pp. 337–343.

[12] W. MURRAY WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer, New York, 1979.

OPTICAL SPECTRA FROM CHEMICAL TITRATION: AN ANALYSIS BY SVD*

RICHARD I. SHRAGER†

Abstract. Given an unknown chemical mixture to which a known substance is being added, a chemist may wish to determine a) how many substances in the mixture are reacting, b) the physical (e.g. thermodynamic) properties of each, and c) what each substance is, or at least some physical identifier of it (e.g. an optical spectrum). We will describe a method using singular value decomposition (SVD) that has been applied to a variety of such problems, ranging from the analysis of simple inorganic mixtures to the examination of mitochondrial membranes of mammalian cells. A simulated chemical example will be used to illustrate the method, and to test the effect of matrix condition on the results. Flexibility of the method and techniques of noise detection will also be discussed.

AMS(MOS) subject classifications. 15A18, 15A90, 80A15

Introduction. Singular value decomposition (SVD) can be a powerful tool in the detection and characterization of chemical transitions in titration experiments [1]. Such experiments, fundamental to chemical practice, start with a substance or mixture to which a control substance (titrant) is added step by step. Gradually, the substances in the mixture change from some initial state through possible intermediate states to a final state, with the fraction of each substance in each state being governed by the concentration of titrant. After each addition of titrant, allowing sufficient time for the mixture to equilibrate, a spectrum of the mixture is taken, e.g. optical absorbance from 300 to 700 nanometers (nm) wavelength in steps of 2 nm. Other types of spectra can also be used, separately or in combination. However, they must all behave linearly: the spectrum of a mixture must be the sum of spectra of the individual species (a given substance in a given state) and the amplitude of a spectrum of a species must be proportional to the concentration of that species.

The object of our calculations is two fold:

1. Describe the law by which the titrant converts each species to its next state (i.e., the transition curves, e.g. see next section).
2. Compute the difference spectrum between the species being converted and the species being produced in each transition. Such differences of spectra help to identify the substances involved.

In mathematical terms, the goal is a matrix decomposition. The measured optical spectra are stored in successive columns of the matrix A, so that the matrix element a_{ij} is the optical absorbance of the mixture at the ith wavelength in the jth spectrum. The desired decomposition is:

$$A = DF^T + E,$$

where each column of D is a difference spectrum associated with one of the transitions, the corresponding column of F is the appropriate transition curve, and E is the matrix of experimental errors. By convention, the final column D is the spectrum that would appear before any transition occurred, i.e. a base spectrum, and the final column of F is all one's, indicating that the base spectrum is applied to all columns of A before

* Received by the editors July 29, 1983. This paper was presented as an invited paper at the Minisymposium on Numerical and Matrix Methods with Physical Applications held in conjunction with the 1983 SIAM National Meeting in Denver, Colorado.

† National Institutes of Health, Division of Computer Research and Technology, Bethesda, Maryland 20205.

the difference spectra are imposed. The general decomposition is now completely specified. In subsequent sections, the kth column of a matrix A will be denoted A col k. The symbol H denotes a matrix, while the symbol $[H^+]$ denotes hydrogen ion concentration. This notation conforms with reference [1].

A chemical example. Before explaining how SVD is used to deduce D and F, some examples may clarify the meaning that D and F hold for chemists. A pH indicator is a substance with two states. A site on the molecule is either protonated (occupied by a proton) or unprotonated, and transitions between the two states are observed through marked changes in color. The concentration of protons or hydrogen ions, denoted $[H^+]$, is related to pH by the definition $[H^+]=10^{-\mathrm{pH}}$ or $\mathrm{pH}=-\log_{10}[H^+]$. That is, the higher pH is, the fewer protons there are per volume of solution, consequently there are also fewer protonated sites. The transition curve for fraction of sites unprotonated as a function of pH is:

$$f(\mathrm{pH}; \mathrm{pK}) = 1/(1+10^{\mathrm{pK}-\mathrm{pH}})$$

where pK is the value of pH at which half the sites of a given indicator are saturated. Each pH indicator has its own pK. Curves of this form are known to chemists as Henderson–Hasselbach (H–H) curves. If the experiment involves, say, three pH indicators, with pK's denoted pK_1, pK_2, and pK_3, then the kth column of F is $f(\mathrm{pH}, \mathrm{pK}_k)$ plotted for every pH in the experiment. The kth column of D is the difference between spectra, pure unprotonated minus pure protonated, of the kth indicator in the mixture.

In this example, the base spectrum in the fourth column of D is the sum of the three pure protonated spectra, which is uninformative, since the individual pure protonated spectra cannot be deduced. In many examples, the base spectrum is further complicated by extraneous materials which absorb light but which do not undergo transitions. In other words, the base spectrum is a burden, which raises the effective rank of A without adding available information. Since the work required of the user (i.e. the curve-fitting described in the next section) is proportional to the effective rank of A, it is current practice to subtract a reference spectrum from all the columns of A. Usually this reference is the initial spectrum in A or the average of all spectra in A. The effective rank of A is thus lowered by one, because D column 4 has now been replaced by a linear combination of D columns 1, 2, and 3, neglecting noise.

SVD. The singular value decomposition (SVD) of A is given by $A = USV^T$, where A is m by n, U is m by n unitary, S is n by n diagonal, and V is n by n unitary. The theory of this decomposition is discussed in [1], and an Algol program is given in [2]. A FORTRAN version is available from G. H. Golub, Computer Science Department, Stanford University, Stanford, California. The fundamental relation for our purposes is:

$$A = DF^T + E = USV^T.$$

Without noise, the example of the previous section, having three transitions, would produce A of rank 3, assuming that the three difference spectra are linearly independent. Therefore, only the first three singular values on the main diagonal of S will be positive, and we need retain only the first three columns of U and V, and the upper left 3 by 3 of S. We denote the truncated matrices as $\bar{U}$, $\bar{S}$, and $\bar{V}$, where

$$DF^T = \bar{U}\bar{S}\bar{V}^T$$

holds exactly.

The columns of $\bar{V}$ are linear combinations of the columns of F. But $F \operatorname{col} K = f(\mathrm{pH}; \mathrm{pK}_k)$, where only pK_k is unknown. Therefore F can be deduced by a series of curve-fitting operations:

$$\bar{V} \operatorname{col} i = h_{i,4}(F \operatorname{col} 4) + \sum_{j=1}^{3} h_{i,j} f(\mathrm{pH}; \mathrm{pK}_k)$$

where the h's and pK's are parameters to be estimated. The h's form a matrix H which satisfies the relation

$$\bar{V}^T = HF^T$$

from which it follows directly that:

$$D = \bar{U}\bar{S}H.$$

The curve-fitting phase of this procedure can sometimes require considerable time and skill. Each column of $\bar{V}$ must be fitted, and there must be agreement as to which pK's appear in more than one column of $\bar{V}$. Since the presence of noise will introduce errors in pK estimates, especially in columns of $\bar{V}$ associated with small singular values, the decision as to whether a pK from $\bar{V} \operatorname{col} 2$ is effectively equal to a pK from $\bar{V} \operatorname{col} 3$ may not always be easy. Sometimes techniques involving constrained fitting or simultaneous fitting of the columns of $\bar{V}$ in question will be required. In ambiguous cases, when more than one hypothesis about the number of pK's becomes workable, an experiment of a different design is required to resolve the issue.

Flexibility. The matrix A can be augmented or pruned whenever an advantage is perceived in resolving the required parameters. Several different kinds of spectrum can be included. Extraneous ranges of wavelength or pH can be removed. If more detail is required in certain regions of wavelength or pH, denser data for those regions can be included. (Equal spacing of the independent variables is not required.) If the curve-fits yield ambiguous results for some pK's, but consistent results for others, the spectra for the reliable pK's can be extracted independently of the others by the relation:

$$D \operatorname{col} j = \bar{U}\bar{S}(H \operatorname{col} j).$$

Flexibility of input naturally leads to the question, "What is a workable experimental design?." In an experiment with e.g. Gaussian absorbance peaks and Henderson–Hasselback transition curves, a rule of thumb is: several (10 or more) values of wavelength (or pH) in the vicinity of each peak (or H–H transition). This type of design produces A matrices with dimensions 10's by 10's. The size of A can sometimes be limited by processing the data in readily distinguishable regions (i.e. a submatrix at a time), thus reducing an experiment with hundreds of points in each direction to manageable size.

Finally, there is also flexibility of the choice of model(s) for fitting the various columns of $\bar{V}$. These models can stem from any physically derivable relation between the concentration of titrant and the concentrations of the various species. The Henderson–Hasselback model is only one such model. Regardless of the models chosen, each transition will usually appear in more than one column of $\bar{V}$, thus requiring agreement in all common parameters except the scale factors $h_{i,j}$.

Noise. While the matrix DF^T is typically of low rank, the matrix $DF^T + E$ is full rank (full rank minus one if a reference spectrum has been subtracted from the columns of A). Therefore, all the singular values will be positive, and the truncation of the matrices U, S, and V will involve some statistical decisions. Assume the variance of each element of A to be σ^2, and choose r such that

$$\text{sums}\,(r+1) \leqq mn\sigma^2 < \text{sums}\,(r) \quad \text{where sums}\,(j) = \sum_{i=j}^{n} s_{i,i}^2.$$

Then the matrix $\bar{U}\bar{S}\bar{V}^T$, where only the first r columns of U, S, and V (and rows of S) have been retained, differs from A by no more than is attributable to noise. The theoretical justification for this choice given by [1, Thm. 2] and [3, formulas 14–17]. However, statistical variation alone would caution against trusting such a sharp threshold. In addition, the assumption of uniform variance in A hardly ever holds. For example, variance tends to increase with optical absorbance. If variance were uniform within each row of A, a diagonal scaling matrix Z would be chosen such that ZA had uniform variance, but it is quite common that contrast of variance within rows is as great as contrast within columns. One hesitates to use scaling on the right, because the resulting $\bar{V}$ would no longer exhibit the proper transition curves.

To avoid complete reliance on an estimate of noise level to determine the effective rank of A, the first order autocorrelations of the columns of U and V are also used. Since the columns of U and V are already normalized, the autocorrelations of U col j and V col j respectively are approximated by:

$$\text{ac}\, u(j) = \sum_{i=2}^{m} u_{i-1,j}u_{i,j}, \qquad \text{ac}\, v(j) = \sum_{i-2}^{n} v_{i-1,j}v_{i,j}.$$

If the experiment is thorough, e.g. several wavelengths on each absorbance peak and several pH's on each transition, then sequential absorbance values in the signal will be highly correlated in both the wavelength and the pH directions. Those columns of U and V that represent signal should also exhibit high autocorrelation. In practice, an autocorrelation below 0.6 seems sufficient to reject corresponding columns of U and V as noise. Notice that no justification can be made for this test if the experiment is not thorough.

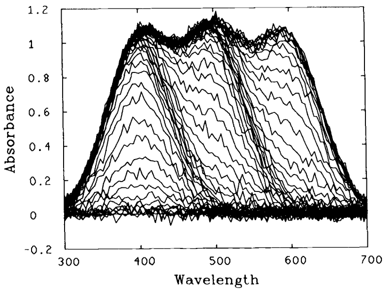
In summary, U col j, V col j, and $s_{j,j}$ are removed from the system if at least one of these conditions holds:

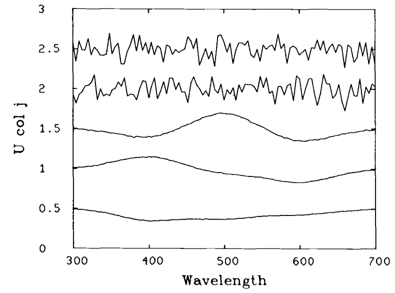
$$\text{sums}\,(j) \leqq mn\sigma^2, \quad \text{ac}\, u(j) < 0.6, \quad \text{ac}\, v(j) < 0.6.$$

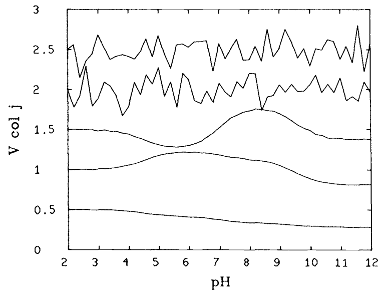
Tests. Three laboratory examples are given in [3]. Figure 1 in [3] with the accompanying text is especially informative because the three pH indicators in that example were known in advance, providing a standard for checking the results. The example was not trivial, because the difference spectra of the three indicators were similar, and two of the pK's were close (about 0.5 pH units apart). Either of these conditions is enough to produce singular values close to the noise level, and nearby pK's cause poor conditioning of the curve-fitting process.

In this section, we will create three sets of data from the same artificial model: three indicators, each with a single Gaussian peak in its difference spectrum, and each with an H–H transition. All Gaussians have half-widths of 50 nm, with midpoints denoted m_i:

format	example 1	example 2	example 3
m_1 pk_1	400 4.5	425 5.5	450 6
m_2 pk_2	500 7	500 7	500 7
m_3 pk_3	600 9.5	575 8.5	550 8

1)


2)


3)


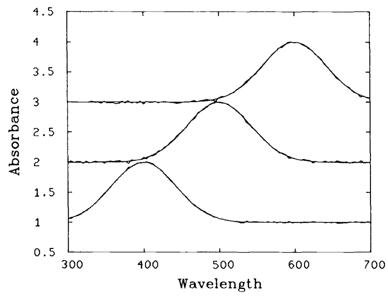
4)


FIG. 1. *Data of the first example, from* pH 2 (*zero-plus-noise*) *to* pH 12 (*three fully-developed peaks*).

FIG. 2. *The first five columns of U from example* 1. *Each U* col *j is raised by* 0.5*j units for visual clarity.*

FIG. 3. *The first five columns of V from example* 1. *Each V* col *j is raised by* 0.5*j units for visual clarity.*

FIG. 4. *Theoretical* (*noiseless*) *and reconstructed difference spectra for the three transitions of example* 1. *Spectra for the jth transition have been raised by j units for visual clarity.*

The noise level (standard deviation of each point) is 0.02 in all cases. With each successive example, the peak midpoints and pK's are moved closer to observe the effects of deteriorating condition on the estimated difference spectra.

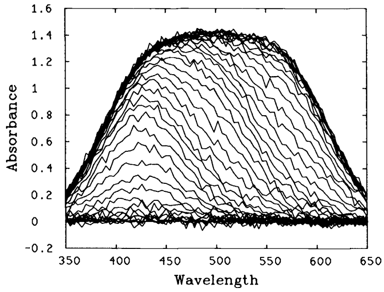
Example 1 is an ideal case with well-separated peaks and pK's (Fig. 1). In fact, the transition curves could be estimated quite well by plotting absorbance versus pH at 400, 500, and 600 nm respectively, since, at those wavelengths, only one peak dominates, and each peak undergoes only one transition. From the SVD of data in Fig. 1, plots of U col j versus wavelength are shown in Fig. 2, with V col j versus pH in Fig. 3. Notice the comparative smoothness of columns 1, 2, and 3 in both U and V. The singular values and autocorrelations of the first five components are:

	1	2	3	4	5
singular values	39.28	10.69	4.62	.323	.302
ac (U col j)	.999	.997	.992	.026	−.127
ac (V col j)	.976	.976	.971	−.045	−.233

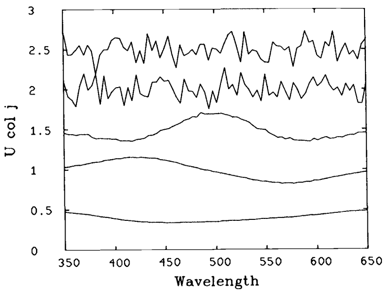
Subsequent singular values decrease gradually, and subsequent autocorrelations are all small. It is clear that the matrix A is effectively rank 3. Curve-fits of the first three columns of V produce pK's that are in error by at most .03 pH unit. Resulting estimated spectra are shown in Fig. 4 with the true Gaussians superimposed.

In example 2 (Fig. 5), it is no longer possible for the chemist to separate peaks visually, nor is it easy to select wavelengths that exhibit "separate" transitions. Absorbance at 500 nm, for example, is influenced noticeably by all three transitions. There

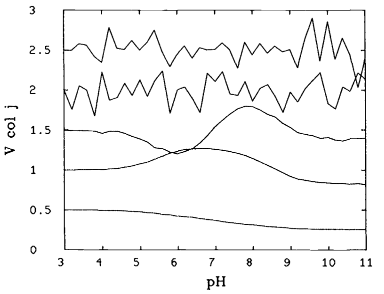
5)



6)



7)



8)

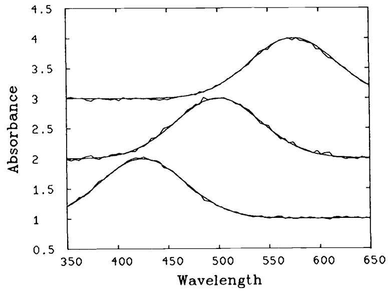


FIG. 5. *Data of the second example, from* pH 3 (*zero-plus-noise*) *to* pH 11.
FIG. 6. *U from example* 2 *as in Fig.* 2.
FIG. 7. *V from example* 2 *as in Fig.* 3.
FIG. 8. *Difference spectra from example* 2 *as in Fig.* 4.

is no wavelength that isolates the transition at pK 7, and the best hope for isolating the pK 5.5 and pK 8.5 transitions is at the tails of the data, where signal-to-noise ratio is at its worst. Yet, for SVD, resolution is still quite good. The first five columns of U col j and V col j are shown in Figs. 6 and 7. It is still visually obvious that only the first three columns of U and V represent signal. The singular values and autocorrelations are:

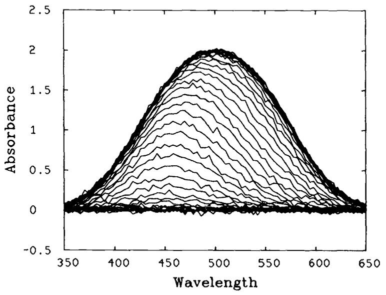
	1	2	3	4	5
singular values	39.43	8.01	2.02	.271	.266
ac (U col j)	.999	.996	.984	−.050	.003
ac (V col j)	.970	.973	.954	−.132	−.111

When fitting each of the first three columns of V to the sum of three H–H curves, the maximum error in any pK is about .064. It is not difficult to conclude that, of the nine pK's generated by the curve-fits, only three are distinct. Figure 8 shows good agreement between the true and estimated spectra, but with noticeable increase in noise level.

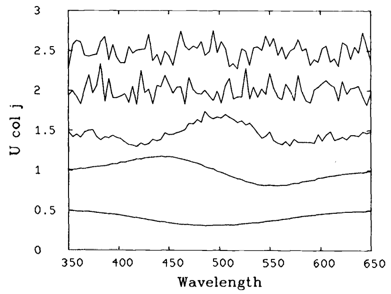
In example 3 (Fig. 9), the limits of resolution are being reached. In Figs. 10 and 11, the noise levels in U col 3 and V col 3 are considerably higher than in the previous examples, because they are now associated with a small singular value:

	1	2	3	4	5
singular values	45.57	5.59	.676	.269	.266
ac (U col j)	.999	.995	.917	−.170	.114
ac (V col j)	.970	.974	.910	−.199	−.070

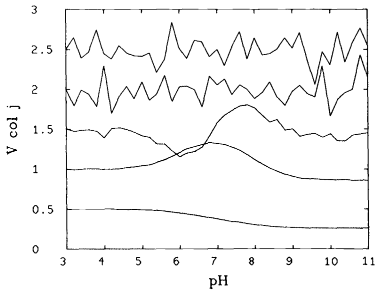
9)



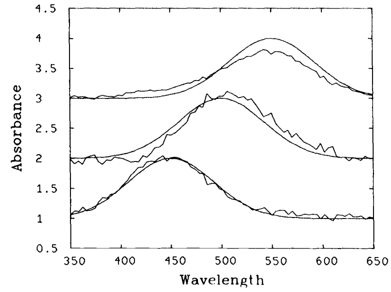
10)



11)



12)



13)

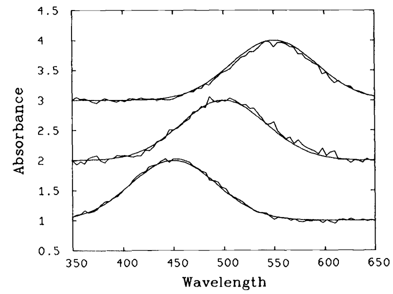


FIG. 9. *Data of the third example, from* pH 2 (*zero-plus-noise*) *to* pH 11.

FIG. 10. *U from example* 3 *as in Fig.* 2.

FIG. 11. *V from example* 3 *as in Fig.* 3.

FIG. 12. *Difference spectra from example* 3 *as in Fig.* 4. *The generating amplitudes* $h_{i,j}$ *were accepted regardless of discrepancies in the corresponding* pK'*s.*

FIG. 13. *As in Fig.* 12, *permitting no discrepancies in the* pK'*s. V* col 2 *and V* col 3 *were fit with* pK'*s fixed at their values from V* col 1.

Notice, however, that the autocorrelations of the third component clearly indicate signal, as does our visual impression. The serious problem is the closeness of the pK's, only one unit apart, making the curve-fitting results sensitive to noise. The pK's, which should ideally be 6, 7, and 8 in the first three columns of *V*, come out like this:

	pK1	pK2	pK3
V col 1	6.02	7.04	8.04
V col 2	5.98	7.69	8.11
V col 3	6.08	6.73	8.08

When the *h*'s from these fits are used to generate estimated spectra, the results are qualitative at best (Fig. 12). It has actually become difficult to assert that the above pK 2 results represent only one transition.

Before giving up on Example 3, consider one further possibility: under the hypothesis that there are only three transitions, it is reasonable to constrain the corresponding pK's to be exactly equal. For example, when the pK's of V col 2 and V col 3 were held fixed at the values generated in V col 1, allowing only the h's to vary, the resulting estimated spectra (Fig. 13) were once again convincing. However, it is difficult to conceive of a reliable strategy for improving robustness. Certainly, pK's from less noisy columns of are not always the most reliable. Reliability also depends, for example, on the amplitude of the transition in V col j. Ultimately, when the pK's get too close, any attempt at resolving all of them will be futile.

REFERENCES

[1] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (1965), pp. 205–224.

[2] G. H. Golub and C. Reinsch, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420.

[3] R. I. Shrager and R. W. Hendler, *Titration of individual components in a mixture with resolution of difference spectra*, pK*'s and redox transitions*, Anal. Chem., 54 (1982), pp. 1147–1152.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

008

CHARACTERS OF ELEMENTS OF FINITE ORDER IN LIE GROUPS*

R. V. MOODY† AND J. PATERA‡

Abstract. In this paper we use the theory of elements of finite order (EFO) as a new and very effective tool for discrete methods in simple Lie groups. The EFO provide a systematic way of discretely approximating the group. Their character values allow us to systematically determine information about Lie groups and their representations for groups well beyond the range of standard methods.

We discuss the theory of EFO, the use of algebraic number fields to single out finite classes of them, and methods of explicitly determining such classes. We introduce an algorithm for effectively computing their character values which utilizes double coset decompositions in the Weyl group and a fast algorithm for determining weight space multiplicities which we developed earlier. The methods are uniform for all simple Lie groups.

We briefly discuss a number of applications of this work and finish with a number of tables (including some for E_6) of EFO and their character values.

AMS (MOS) subject classifications. Primary 22E46; secondary 17B20, 22E40, 2004.

1. Introduction. The theory of simple Lie groups is of major importance in mathematics and physics. The need for ever more subtle information about these groups and their representations and the increasing importance of higher rank Lie groups in physics (see for example the major report of Slansky [Sla]) has led us to look for new and more effective discrete methods for handling such groups.

In practice a Lie group G always appears in connection with one or more representations. One of the most valuable tools at our disposal is the character of the representation. The character is an analytic function on G containing potentially all the information about the representation. Unfortunately, it is not always easy to compute with the characters. However, one rarely needs the complete information about a representation for a specific task. The idea which motivates our approach is to replace the character by a set of its values at suitably chosen discrete points. To handle the characters discretely (for instance, from the point of view of computation) we need an increasing system of finite sets $F_1 \subseteq F_2 \subseteq, \cdots$ of group elements containing representatives of different conjugacy classes of G whose union is dense in G. For this purpose the emerging theory [Ka], [Ko], [DJ], [CQ], [MPS] of elements of finite order (EFO) turns out to be perfectly suited. We take F_n to be carefully selected EFO representing the conjugacy classes of elements of G of order $\leqq n$. These classes are quite easy to describe and are particularly suitable for applications in that their character values are algebraic integers in cyclotomic fields and hence are *exactly* computable.

This paper is concerned with EFO and with *effective* methods for computing their character values. The general applicability of character tables for finite groups is well known and documented (e.g. [Mi]). The potential applicability of EFO and their character values is probably equally great although largely unexplored. As a simple illustration, consider the problem, which frequently occurs in applications, of determining the irreducible constituents of the tensor product of two irreducible representations

* Received by the editors February 1, 1983, and in revised form August 31, 1983. This work was supported by operating grants and a Collaborative Special Project Grant from the Natural Sciences and Engineering Research Council of Canada.

† Centre de Recherches de Mathématiques Appliquées, Université de Montréal, Montréal, Québec, Canada H3C-3J7. On leave from Department of Mathematics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

‡ California Institute of Technology, Pasadena, California 91125. On leave from Centre de Recherches de Mathématiqués Appliquées, Université de Montréal, Montréal, Québec, Canada.

ϕ_1 and ϕ_2: $\phi_1 \otimes \phi_2 = \bigoplus_{i=3}^{n} \phi_i$. In principle this problem is solved by any of several well-known formulas (Racah, Steinberg). In practice these formulas are hopelessly inefficient and quite impossible to use for higher rank groups. On the other hand if s_i is the character of ϕ_i then we have $s_1 s_2 = \sum_{i=3}^{n} s_i$. With some knowledge of the s_i which are actually possible in the decomposition (usually readily available), the evaluation of the characters of EFO reduces the problem to the solution of a system of linear equations [MP2].

EFO are closely connected to the existence of finite subgroups of Lie groups. In physics, for example, they are being used in Monte Carlo methods in lattice gauge theories (see [Sta], for instance). For some purely mathematical developments see [Me], [Sl]. The characters give basic information about the ways in which such groups lie in Lie groups. As an example in § 9 we identify the conjugacy classes of EFO belonging to PSL(2, 13) in G_2. Section 9 contains in addition a number of other applications of our work.

We turn now to a more explicit description of our methods.

Let G be a compact simply connected simple Lie group and $\mathfrak{g}$ its Lie algebra. We fix, once and for all, a maximal torus T of G with Lie algebra $\mathfrak{t}^1$. Define $\mathfrak{t} = i\mathfrak{t}^1 \subset \mathfrak{t}^1_{\mathbb{C}}$ (here as elsewhere in the paper, the subscript $\mathbb{C}$ attached to the name of real vector space indicates the complexification of that space). We have the exponential map

$$\exp 2\pi i(\cdot): \mathfrak{t} \to T. \tag{1.1}$$

Throughout the paper all representations of G will be assumed to be finite dimensional and analytic. If $\pi: G \to GL(V)$ is such a representtion, then V decomposes into weight spaces relative to T:

$$V = \bigoplus_{\lambda \in \Omega} V^{\lambda}, \tag{1.2}$$

where $\Omega \subset \mathfrak{t}^*$ is the weight system of V and the V^{λ}, $\lambda \in \Omega$, are the corresponding weight spaces. In particular the adjoint representation leads to the root space decomposition

$$\mathfrak{g}_{\mathbb{C}} = \mathfrak{t}_{\mathbb{C}} \oplus \bigoplus_{\alpha \in \Delta} \mathfrak{g}^{\alpha}_{\mathbb{C}}, \tag{1.3}$$

where Δ is the root system of $\mathfrak{g}$ relative to T.

We choose a basis

$$\Pi = \{\alpha_1, \cdots, \alpha_l\} \tag{1.4}$$

of simple roots in Δ once and for all, and let Δ^+ denote the corresponding system of positive roots.

Let $\langle \cdot, \cdot \rangle: \mathfrak{t}^* \times \mathfrak{t} \to \mathbb{R}$ be the standard pairing and denote by $(\cdot, \cdot): \mathfrak{t}^* \times \mathfrak{t}^* \to \mathbb{R}$ the transpose of the Killing form. The Cartan matrix A of G is then defined by $A_{ij} = 2(\alpha_i, \alpha_j)/(\alpha_j, \alpha_j)$. The kernel of the exponential mapping in (1.1) is the *coroot lattice* $Q^{\wedge}$. Its $\mathbb{Z}$-dual in $\mathfrak{t}^*$ is the *weight lattice* P in which all weight systems of representations of G must lie. In particular the *root lattice* Q, which is by definition the $\mathbb{Z}$-span of the roots, is a sublattice of P. The index of Q in P is the finite number $\det(A)$, also called the *index of connection.* It is a well-known fact that the weight system of any irreducible representation of G lies entirely in one coset of P/Q. This results in the partitioning of the irreducible representations into $[P:Q]$ classes called *congruence classes* [LP].

The $\mathbb{Z}$-dual of Q in $\mathfrak{t}$ is the *coweight lattice*, $P^\wedge$. The coweight and coroot lattices may be considered as the weight and root lattices for the "dual" group $G^\wedge$ defined by the transpose of the Cartan matrix A.

Let $\alpha_1^\wedge, \cdots, \alpha_l^\wedge$ be the simple coroots defined by $\langle \phi, \alpha_i^\wedge \rangle = 2(\phi, a_i)/(\alpha_i, \alpha_i)$ for all $\phi \in \mathfrak{t}^*$. Then $\{\alpha_1^\wedge, \cdots, \alpha_l^\wedge\}$ is a $\mathbb{Z}$-basis of $Q^\wedge$. The basis $\{\omega_1, \cdots, \omega_l\}$ of P dual to this consists of the *fundamental weights* in P (relative to Π).

Let (V, π) be a representation of G. The character of G afforded by V is the mapping denoted ch_V

$$x \mapsto \mathrm{tr}_V\,(\pi(x)). \tag{1.5}$$

Every element of G is conjugate to an element of T, and for $x = \exp 2\pi i \mathbf{x} \in T$, $\mathbf{x} \in \mathfrak{t}$

$$\mathrm{ch}_V\,(x) = \sum_\lambda \dim V^\lambda\, e^{2\pi i \langle \lambda, \mathbf{x} \rangle}, \tag{1.6}$$

where λ runs over the weight system Ω of V.

Our primary aim is to present the methods we have evolved for computing such character values when x is an EFO. All the simple Lie groups are put on an equal footing and we do not exploit the special properties of any particular group in the computation of its characters. Indeed the theory of simple Lie groups is now extremely uniform and there is no reason why the algorithms should not be equally so. Apart from the aesthetic principle involved, there is a considerable economy in programming in such an approach. The only other extensive computation of Lie group characters we are aware of is that of J. Conway and L. Queen for E_8 [*CQ*]. This work, which we found very useful, was never intended for anything other than than E_8.

An important point in our work is the full exploitation of the Weyl group $W := N/T$ ($:=$ means that the right-hand side defines the left-hand side), where N is the normalizer of T in G. As is well known, weight systems and weight space multiplicities are W-invariant and vast amounts of computing can be dispensed with by utilizing this inherent symmetry. This is absolutely essential as size is a dominating consideration when computing in any but the lowest rank Lie algebras. For example, E_6 is relatively modest and yet $|W| = 51{,}840$ and there are only 9 irreducible representations of dimension $<10{,}000$ in congruence class 0 and only 6 in each of the remaining two classes. The E_6 character tables at the end of this paper were computed in a total of 434 seconds[1] (excluding the computation of the multiplicities, which was a small fraction of this by comparison (see Table 2)). Actually we find that the complexity of our algorithm depends more on the structures of the stabilizers in W of the weights and the EFO than the dimensions of the representations.

To explain this in more detail, recall that W acts naturally on $\mathfrak{t}$ where it stabilizes the sets $P^\wedge$ and $Q^\wedge$. By transpose action it acts on $\mathfrak{t}^*$ where it stabilizes the sets P, Q, Δ, and every weight system Ω. For each $\alpha \in \Delta$ there is a coroot $\alpha^\wedge \in \mathfrak{t}$ such that the symmetry in α

$$r_\alpha : \phi \mapsto \phi - \langle \phi, \alpha^\wedge \rangle \alpha, \qquad \phi \in \mathfrak{t}^*, \tag{1.7}$$

is in W. Furthermore, W is generated by the symmetries $r_i := r_{\alpha_i}$, $i = 1, \cdots, l$, in the simple roots (each $\alpha_i^\wedge$ being the coroot defined before). An element λ of P (respectively $\mathbf{x} \in P^\wedge$) is called *dominant* if $\langle \lambda, \alpha_i^\wedge \rangle \geqq 0$ (respectively $\langle \alpha_i, \mathbf{x} \rangle \geqq 0$) for each $i = 1, \cdots, l$. Every W-orbit in P or $P^\wedge$ has a unique dominant element.

[1] CDC Cyber 835, Centre de Calcul, Université de Montréal. All programs are written in Pascal.

Let $\mathbf{S} := \{1, \cdots, l\}$ and for each subset J of $\mathbf{S}$ let W_J be the group generated by the r_j, $j \in J$. Such subgroups are called *parabolic subgroups* of W. For dominant $\lambda \in \mathrm{t}^*$ (respectively $\mathbf{x} \in \mathrm{t}$) the stabilizer $\mathrm{Stab}_W(\lambda)$ of λ (respectively, $\mathrm{Stab}_W(\mathbf{x})$ of $\mathbf{x}$) in W is W_J, where $J = \{i \in \mathbf{S} \mid \langle \lambda, \alpha_i^{\wedge} \rangle = 0\}$ (respectively, $J = \{i \in \mathbf{S} \mid \langle \alpha_i, \mathbf{x} \rangle = 0\}$) [Bo].

Let (V, π) be an irreducible representation of G and let Ω be the weight system of V relative to T. Let $\lambda_1, \cdots, \lambda_k$ be the dominant weights in V and $D_1, \cdots, D_k$ their multiplicities: $D_j = \dim V^{\lambda_j}$. Then for $x = \exp 2\pi i\mathbf{x}$, $\mathbf{x} \in \mathrm{t}$, we have

$$\text{(1.8)} \qquad \mathrm{ch}_V(x) = \sum_{j=1}^{k} D_j \sum_{\mu \in W\lambda_j} e^{2\pi i \langle \mu, \mathbf{x} \rangle},$$

where $W\lambda_j = \{w\lambda_j \mid w \in W\}$. Let $W_{(j)}$ denote $\mathrm{Stab}_W(\lambda_j)$. Each left coset $wW_{(j)}$ has a unique element of minimal length (with respect to the generators $r_1, \cdots, r_l$). Denote the set of these minimal respresentatives by $W^{(j)}$. Then

$$\text{(1.9)} \qquad \mathrm{ch}_V(x) = \sum_{j=1}^{k} D_j \sum_{w \in W^{(j)}} e^{2\pi i \langle w\lambda_j, \mathbf{x} \rangle}.$$

The advantage of (1.8) or (1.9) is that it is possible to determine $\lambda_1, \cdots, \lambda_k; D_1, \cdots, D_k$ efficiently without having to compute the entire weight system.

Let $W_K = \mathrm{Stab}_W(\mathbf{x})$. Then, since $\langle w\lambda_j, \mathbf{x} \rangle = \langle \lambda_j, w^{-1}\mathbf{x} \rangle$, elements $W^{(j)}$ in the same right coset $W_K\mu$ determine the same exponent in (1.9). Let ${}^K W^{(j)}$ denote the set of unique [Bo, Chap. IV, Ex. 3] minimal coset representatives for $W_K \backslash W / W_{(j)}$. For each $w \in {}^K W^{(j)}$ denote by $n(w, K, (j))$ the number of left cosets of $W_{(j)}$ which are covered by $W_K w W_{(j)}$. Then

$$\text{(1.10)} \qquad \mathrm{ch}_V(x) = \sum_{j=1}^{k} D_j \sum_{w \in {}^K W^{(j)}} n(w, K, (j))\, e^{2\pi i \langle w\lambda_j, \mathbf{x} \rangle}.$$

For a regular element (one whose stabilizer in W is trivial), (1.10) offers nothing over (1.9). However, in higher rank groups elements with nontrivial stabilizers are abundant and for them the computation in (1.9) is reduced by a factor comparable with the order of the stabilizer. A simple example will illustrate the point.

Consider the E_6 representation whose highest weight label is ${}_{1\,0\,\overset{0}{0}\,0\,1}$ of dimension 650. There are three dominant weights in the system heading orbits of size 270, 72, and 1 according to Table 1.

If we now compute the character of the element

$$\begin{array}{c} 0 \\ 0 \\ \mathbf{s} = [00100] \quad \text{of order } 3 \end{array}$$

(see § 4 for notation) then the number of double cosets in (1.10) is given by the last column of the table. Thus the sum of (1.9) with 343 terms is reduced to one with 29

TABLE 1

dominant weight	multiplicity	left cosets	double cosets
$1\ 0\ \overset{0}{0}\ 0\ 1$	1	270	19
$0\ 0\ \overset{1}{0}\ 0\ 0$	5	72	9
$0\ 0\ \overset{0}{0}\ 0\ 0$	20	1	1

terms. As we will see, the calculation of double coset representatives is actually very fast once the single coset representatives are known. In 5 of § 9 we discuss some additional collapsing of (1.10) which occurs when the 0th component of the coordinate of the EFO vanishes.

Using either (1.9) or (1.10) we are faced with three problems:

(I) Compute $\lambda_1, \cdots, \lambda_k$; $D_1, \cdots, D_k$.

(II) Compute the minimal single or double coset representatives for parabolic subgroups of W.

(III) Determine suitable elements of finite order at which to compute the character values.

We have already dealt with (I) in [MP1]. For completeness and because it is central to our work, we briefly describe it in § 2. We deal with (II) in § 3. In this paper we have taken (III) to mean that the character values should appear in some uncomplicated algebraic number field, usually the rational numbers $\mathbb{Q}$.

The theory of elements of finite order is less well known than other parts of the theory and we found it necessary to fill in several gaps before we could undertake any computation. In § 4, § 5, § 6 we sketch out this theory, including a number of results which we discovered in the process of our computations. This includes a discussion of rationalilty questions and some results about regular EFO.

The natural inclusions between simple Lie groups lead to identification between elements in various groups. In the case of rational elements some of these relations are at first sight quite amazing. In § 7 we show how identifications are made and give a number of examples.

Section 8 is a brief description of the computational plan for computing characters and § 9 consists of some additional remarks on various aspects of characters and their applications. Finally there is a set of Tables with information and some sample computations of characters. This includes a G_2 character table including the relevant information about PSL (2, 13) in G_2 and two tables of E_6 characters on the rational elements of order $\leqq 8$.

As a final remark let us point out that a different approach to the determination of character values has been given in [MPS], where for the Lie groups of types A_1, A_2, A_3, B_2, G_2 we have constructed the explicit character generating functions and determined their specializations at all the rational EFO of these groups. This approach gives the character information in a complete and very compact form but appears to be impractical for much higher ranks.

Some results of this paper were already announced in [MPS].

Note added in proof. In [MP2] we indicate a more efficient method for computing characters of EFO. This is based on techniques discussed here but in addition exploits the existence of various finite subgroups of T.

2. Determination of the dominant weights and their multiplicities. (Details of the mathematics of this section appear in [MP1]. The tables of [BMP] should provide the multiplicities for most of the cases of practical interest.)

Let (V, π) be the irreducible representation with highest weight Λ. Then Λ lies in the set P^{++} of dominant elements in the weight lattice. Let Ω be the weight system of V and $\Omega^{++} = \Omega \cap P^{++}$.

Define L_k, $k = 0, 1, 2, \cdots$, inductively by $L_0 = \{\Lambda\}$,

$$L_k = \left\{ \gamma \in P^{++} - \bigcup_{i=1}^{k-1} L_i \,\middle|\, \gamma = \lambda - \beta, \lambda \in L_{k-1}, \beta \in \Delta^+ \right\}. \tag{2.1}$$

Then

$$\Omega^{++}=\bigcup_{k=0}^{\infty} L_k. \tag{2.2}$$

This provides an effective simple procedure for inductively constructing Ω^{++}. To compute multiplicities $D_\lambda=\dim V^\lambda$, $\lambda\in\Omega^{++}$, we use the following modified Freudenthal formula: Let $\mathrm{Stab}_W(\lambda)=W_J$, where

$$J=\{i\mid\langle\lambda,\alpha_i^\wedge\rangle=0\}=\{i\mid(\lambda,\alpha_i)=0\}.$$

Let $\hat{W}_J=\langle W_J,-1_V\rangle$ be the group generated by W_J and -1_V. $\hat{W}_J$ decomposes Δ into orbits O_i, $i=1,\cdots,n$, and each orbit has a unique representative $\xi_i=\sum n_{ij}\omega_j$, where $n_{ij}\geqq 0$ for all $j\in J$. In terms of these

$$\sum_{i=1}^{n}|O_i|\sum_{p=1}^{\infty}(\lambda+p\xi_i,\xi_i)D_{\lambda+p\xi_i}=(C_\Lambda-C_\lambda)D_\lambda, \tag{2.3}$$

where for $\mu\in P$

$$C_\mu:=(\mu+\rho,\mu+\rho)-(\rho,\rho),\qquad \rho=\tfrac{1}{2}\sum_{\alpha\in\Delta^+}\alpha.$$

In practice after its computation, Ω^{++} is ordered according to the *level*; that is according to $\langle\lambda,\rho^\wedge\rangle$, $\lambda\in\Omega^{++}$, where $\langle\alpha_i,\rho^\wedge\rangle=1$ for all i. Then the D_λ are computed by (2.3) in decreasing level. This is an example of the situation mentioned before: for increasing rank, the weights most usually encountered have large stabilizers and considerable savings result by using (2.3). For example, in E_6 the "first" dominant weight with a nontrivial stabilizer is given by ρ and the corresponding irreducible representation is of dimension 2^{36}.

In general $\lambda+p\xi_i$ is not dominant and has to be reflected (by at most $|\Delta^+|$ reflections in the r_i, $i\in\mathbf{S}$) before its multiplicity can be looked up. Even with this, the algorithm appears to be quite fast. Table 2 lists some timings. In each case the times are *total* times in seconds for the first (by level) 60 irreducible representations of a single congruence class. More precisely, since timings between different congruence classes of the same algebra are within 12% of the average, we have listed the average times (Table 2).

TABLE 2

Rank type	A	B	C	D	E
5	64.8	120.9	132.1	89.2	—
6	75.7	145.6	149.3	110.7	121.4
7	89.0	167.1	197.4	134.9	155.4
8	94.0	194.6	202.4	156.6	240.7

If one wishes to use (2.3) for a hand computation of multiplicities, the determination of the ξ_i becomes tiresome. There are, however, only finitely many of ξ_i for any group. Tables of them for all the rank $\leqq 8$ simple Lie groups appear in [BMP]. For the machine computation of weight multiplicities these tables are sufficiently large that it is more efficient to simply compute the ξ_i as they are needed.

3. The Weyl group. In this section we describe an algorithm which can be used for determining the elements of W, the minimal left coset representatives for a parabolic subgroup W_J, or the minimal double coset representatives for a pair of parabolic subgroups W_K, W_J. We are grateful to N. Iwahori for suggesting this beautiful and simple procedure. For our purposes its important virtue is that one is not required to compute the entire Weyl group (which is generally out of the question for reasons of size) in order to compute single or double cosets.

For $\lambda \in P$ write $\lambda = \sum d_i(\lambda)\omega_i$, where $\omega_1, \cdots, \omega_l$ are the fundamental weights ($\langle \omega_i, \alpha_j \rangle = \delta_{ij}$). Let $\Lambda \in P^{++}$ and let $J = \{j \in \mathbf{S} \mid d_j(\Lambda) = 0\}$. Then $W_J = \mathrm{Stab}_W(\Lambda)$ and W/W_J is in 1–1 correspondence with $W\Lambda$ by $w\Lambda \leftrightarrow wW_J$. We construct $W\Lambda$ as the nodes of the coset graph $\Sigma(\Lambda)$ with edges "coloured" by the set $\mathbf{S}$, where distinct nodes $w\Lambda$ and $w'\Lambda$ are joined by an i-edge if and only if $r_i w\Lambda = w'\Lambda$. Note that Σ is connected by assumption and has $|\Sigma|$ nodes where

$$|\Sigma| = |W|/|W_J|. \tag{3.1}$$

Define the depth of a node λ to be the number of edges in a minimal path from the head node, Λ. We use the adjective "up" and "down" in the obvious sense with respect to the depth. The depth of λ is the minimum value of the length $l(w)$ of w as w runs over all elements in W such that $\lambda = w\Lambda$.

LEMMA 3.1. *For $\lambda = \sum d_j(\lambda)\omega_j = w\Lambda$ (w reduced) and $i \in \mathbf{S}$*

(i) $d_i(\lambda) > 0 \Rightarrow l(r_i w) > l(w)$,

(ii) $d_i(\lambda) < 0 \Rightarrow l(r_i w) < l(w)$,

(iii) $d_i(\lambda) = 0 \Rightarrow r_i\lambda = \lambda$.

Proof.

$$l(r_i w) > l(w) \Leftrightarrow w^{-1}\alpha_{\hat{i}} \varepsilon \Delta^+ \Leftrightarrow \langle \omega_k, w^{-1}\alpha_{\hat{i}} \rangle > 0 \quad \text{for some } k.$$

Thus

$$\begin{aligned} d_i(\lambda) > 0 &\Rightarrow \langle \lambda, \alpha_{\hat{i}} \rangle > 0 \\ &\Rightarrow \langle \Lambda, w^{-1}\alpha_{\hat{i}} \rangle > 0 \\ &\Rightarrow \langle \omega_k, w^{-1}\alpha_{\hat{i}} \rangle > 0 \quad \text{for some } k \text{ since } \Lambda \text{ is dominant} \\ &\Rightarrow l(r_i w) > l(w). \end{aligned}$$

The argument for $d_i(\lambda) < 0$ is similar. The last statement is obvious. □

We construct Σ by increasing depth starting from Λ. The lemma tells us how to determine the nodes adjacent to, and of greater depth than λ. The edges of any minimal path from Λ to $w\Lambda$ give a reduced expression for the minimal coset representative of wW_J. The implementation of the coset graph construction is by an array, indexed by the depth, of linked records. Each record corresponds to a node $w\Lambda$ of the graph and basically contains the vector $w\Lambda$ and the sequence of edges from Λ by which it was formed. The latter is, of course, equivalent to a reduced expression for the minimal word in the coset wW_J. During execution the vector λ in each constructed node is searched for possible descendants ($d_i(\lambda) > 0$) and any new ones inserted. For the purpose of computing characters it is in fact unnecessary to know the Weyl group elements, the node weights being sufficient. However, we do need the node weights, and it is not sufficient to construct the "simpler" coset graph $\Sigma(\Lambda')$ based on $\Lambda' = \sum_{d_j(\lambda) \neq 0} \omega_j$.

Now suppose that $I \subset \mathbf{S}$ and we wish to determine the minimal double coset representatives of $W_I \backslash W / W_J$ We have:

PROPOSITION 3.2. *A node of Σ represents a minimal double coset representative of $W_I \backslash W / W_J$ if and only if it has no i-edges ($i \in I$) upward from it.*

Proof. Let λ be a node of Σ of depth $d > 0$. If there is an i-edge $(i \in I)$ upwards from λ then there is a minimal path from Λ through $r_i\lambda$ and then through λ defining Weyl group elements r_iw and w respectively with $l(r_iw) < l(w)$. Thus λ does not represent a minimal double coset of $W_I \backslash W / W_J$.

Now suppose that $\lambda = w_0\Lambda$ and there is no i-edge $(i \in I)$ upwards from λ. Then by Lemma 3.1, $d_i(\lambda) \geqq 0$ for all $i \in I$. Let $V_I = \sum_{i \in I} \mathbb{R}\alpha_i$ and let $\pi\colon V \to V_I$ be orthogonal projection. The mapping π is W_I-equivariant and $V_I \cap \ker \pi = 0$, so π is injective on V_I-cosets, and hence on W_I-orbits. Thus there is in the W_I-orbit of λ a *unique* element λ' which projects onto a dominant weight of V_I relative to $\{\alpha_i \mid i \in I\}$. Seeing as $(\pi(v), \alpha_i) = (v, \alpha_i)$ for $i \in I$, this tells us that there is in the W_I-orbit of λ a unique element λ' with $(\lambda', \alpha_i) \geqq 0$ for $i \in I$; that is, with no i-edges up from it. This is then λ itself and shows that it represents the unique minimal double coset representative.

Thus a simple search through the graph enables one to pick out the minimal double coset representatives. If $\lambda = \sum d_j(\lambda)\omega_j$ has no i-edge upward from it then the corresponding double coset has

$$|W_I|\,|W_J|/|W_{I_0}| \tag{3.2}$$

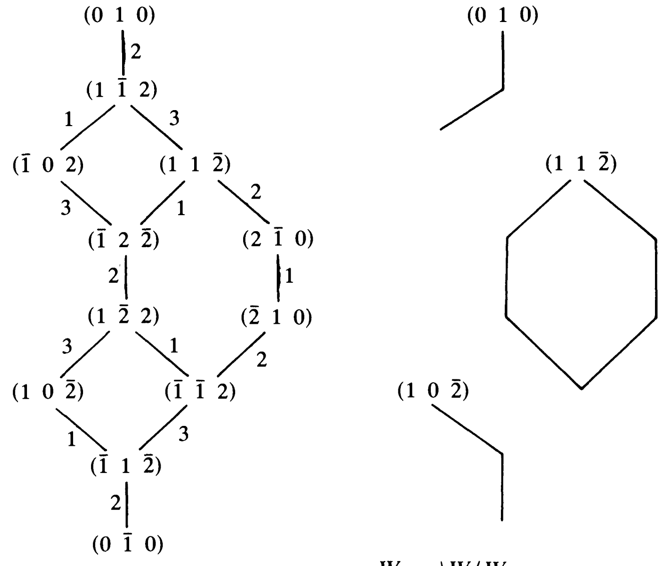
elements, where $I_0 = \{i \in I \mid d_i(\lambda) = 0\}$, and consists of

$$n(\lambda, I, J) = |W_I|/|W_{I_0}| \tag{3.3}$$

left W_J-cosets. An example of the single and double coset graphs appears in Fig. 1.

B_3 Cartan matrix $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -2 \\ 0 & -1 & 2 \end{pmatrix}$

$J = \{1, 3\}$; $\Lambda = (0, 1, 0)$



$W_{\{1\ 2\}} \backslash W / W_{\{1\ 3\}}$
$J = \{1, 3\}$; $I = \{1, 2\}$
double coset representatives
1, 32, 32132

FIG. 1. *An example of single and double coset graphs of the representation* $\Lambda = (0, 1, 0)$ *of dimension* 21 *of the group* B_3.

4. Elements of finite order. The extended Coxeter–Dynkin diagram $\tilde{\Gamma}$ associated with the simply connected simple Lie group G is the standard Coxeter–Dynkin diagram Γ of G extended by a single node associated with the highest long root $\xi = \sum_{i=1}^{l} n_i\alpha_i$. Thus there are l nodes associated with the simple system of roots Π and a 0th node attached to Γ according to the geometry of $-\xi$ relative to the elements of Π [Bo]. We call $n_0 := 1, n_1, \cdots, n_l$ the *numerical marks* of $\tilde{\Gamma}$. For convenience of the reader and to pin down the labelling (which is based on that of Dynkin [Dy, Table 26]) we have included a list of extended Coxeter–Dynkin diagrams in Table 6 (at the end of the article). Following J. Conway we call the nodes whose marks are 1 the *tips* of $\tilde{\Gamma}$.

There are numerous interpretations of $\tilde{\Gamma}$, the one concerning us being probably the earliest which states that it describes the geometry of the fundamental region $\mathbf{F}$ of the affine Weyl group $\tilde{W}$ acting on t. We define $\tilde{W}$ as the semi-direct product $W \ltimes Q^\wedge$ where the action of W on $Q^\wedge$ is given by restriction of its action on t. $\tilde{W}$ acts on t according to

$$(4.1) \qquad (w, \mathbf{q}) : \mathbf{x} \mapsto w\mathbf{x} + \mathbf{q}.$$

Then

$$(4.2) \qquad \mathbf{F} = \{\mathbf{x} \in \mathrm{t} \,|\, \langle \alpha_i, \mathbf{x}\rangle \geqq 0,\ i \in \mathbf{S}, \langle \xi, \mathbf{x}\rangle \leqq 1\}$$

which is simplex in t. $\tilde{W}$ is generated by the Euclidean reflections in the faces of $\mathbf{F}$. For future reference we record these reflections.

$$(4.3) \qquad \begin{aligned} r_i &: \mathbf{x} \mapsto \mathbf{x} - \langle \alpha_i, \mathbf{x}\rangle \alpha_i^\wedge, \qquad i \in \mathbf{S}, \\ r_0 &: \mathbf{x} \mapsto \mathbf{x} - (\langle -\xi, \mathbf{x}\rangle + 1)\alpha_0^\wedge, \end{aligned}$$

where $\alpha_0^\wedge \in$ t is defined by

$$(4.4) \qquad \langle \alpha_i, \alpha_0^\wedge \rangle = -2(\alpha_i, \xi)/(\xi, \xi), \qquad i \in \mathbf{S}.$$

For every element $x \in G$ there is a unique $\mathbf{x} \in \mathbf{F}$ for which x is conjugate to $\exp 2\pi i\mathbf{x}$. In particular each element x of finite order N is represented by a point $\mathbf{s}$ in $\mathbf{F}$. Following V. Kac [Ka1] we assign $\mathbf{s}$ a set of weighted barycentric coordinates $[s_0, s_1, \cdots, s_l]$ of nonnegative integers with $\gcd(s_0, s_1, \cdots, s_l) = 1$ according to the prescription: Let the order of x in Ad (G) (the Ad-*order*) be M. Then

$$(4.5) \qquad \langle \alpha_i, \mathbf{s}\rangle = s_i/M, \quad i \in \mathbf{S}, \qquad M = \sum_{i=0}^{l} n_i s_i.$$

The exact order N of x is related to M in the following way. Define $w_i^\wedge \in$ t by $\langle \alpha_j, \omega_i^\wedge \rangle = \delta_{ji}$ so that $\omega_1^\wedge, \cdots, \omega_l^\wedge$ is a basis for the dual weight lattice $P^\wedge$. Then $\mathbf{s} = (1/M)\sum_{i=1}^{l} s_i\omega_i^\wedge$ and $M\mathbf{s}$ clearly lies in $P^\wedge$. If C is the order of $M\mathbf{s}$ modulo $Q^\wedge$ then $N = MC$. Column C of Table 6, which was derived from the results of [*LP*], gives these values explicitly. As an example consider the elements $[s_0, s_1, s_2, s_3] = [1, 1, 2, 1]$, $[1, 1, 1, 2]$, $[2, 1, 1, 1]$, and $[1, 2, 1, 1]$ respectively in A_3. The values g of $\sum_{i=0}^{4} is_i$ modulo $4(=l+1)$ determine the coweight class modulo the coroot lattice to which these elements belong. These are 0, 1, 2 and 3 respectively. The full orders of the corresponding EFO are $(\sum n_i s_i)C = 5C$ where $C = 4/\gcd(4, g) = 1, 4, 2, 4$ respectively. As a matter of notation, if $q_0, \cdots, q_l$ is an $(l+1)$-tuple of nonnegative rational numbers (not all zeros), we shall write $[q_0, \cdots, q_l]^{\boldsymbol{\cdot}}$ to indicate the EFO whose coordinates are the q_i so scaled that they are nonnegative integers without common factors.

In the sequel we will often refer to a point $\mathbf{s}$ of $\mathbf{F}$ being an EFO by which we mean that $\exp 2\pi i\mathbf{s}$ is such. For instance a point $\mathbf{s}$ of $\mathbf{F}$ is in the center Z of G (meaning that $\exp 2\pi i\mathbf{s}$ is) if and only if $\mathbf{s}$ is at a vertex of $\mathbf{F}$ whose weight is 1—that is $\mathbf{s}=[0,\cdots,1,\cdots,0]$ corresponding to a tip of $\tilde{\Gamma}$. This follows directly from noting that $\mathbf{s}$ is in the center if and only $\langle\alpha_i,\mathbf{s}\rangle\in\mathbb{Z}$ for $i\in\mathbf{S}$.

The center also may be identified with $P^\wedge/Q^\wedge$ by $\mathbf{z}\in P^\wedge\to\exp 2\pi i\mathbf{z}$. This gives rise to an action of Z on $\mathbf{F}$, and hence on $\tilde{\Gamma}$, through

$$P^\wedge/Q^\wedge\times \mathrm{t}/\tilde{W}\to \mathrm{t}/\tilde{W}$$

induced through

$$(\mathbf{z},\mathbf{x})\mapsto\mathbf{z}+\mathbf{x},\qquad \mathbf{z}\in P^\wedge,\quad \mathbf{x}\in\mathrm{t}. \tag{4.6}$$

This action is clearly simply transitive on the tips (since $\mathbf{z}\in P^\wedge$ sends $\mathbf{0}$ to $\mathbf{z}$) and determines the decomposition

$$\mathrm{Aut}\,(\tilde{\Gamma})\simeq Z\rtimes\mathrm{Aut}\,(\Gamma).$$

Two EFO, $\mathbf{s}$ and $\mathbf{s}'$, are conjugate in $\mathrm{Ad}\,(G)$ if and only if their entries differ by a permutation induced by an element of Z. A set of generators for Z as a permutation group on the tips is also given in Table 4. Thus for example in C_3, $[1\ \ 1\ \ 3\ \ 2]$ and $[2\ \ 3\ \ 1\ \ 1]$ are Ad-conjugate and their orders in G are 11 and 22 respectively.

In this regard we mention a formula of D. Djoković [Dj]: for positive integers k with $\gcd(k,|W|)=1$ the number of conjugacy classes of EFO in G with order dividing k is

$$\prod_{i=1}^{l}\frac{m_i+k}{m_i+1}, \tag{4.7}$$

where $m_1,\cdots,m_l$ are the exponents of W.

On the other hand we can determine the generating function of the number $g(k)$ of conjugacy classes of EFO in G with order dividing k, for *all* k. Except in the cases D_2, $r=2,3,\cdots$, let $c_0,\cdots,c_l$ be determined as the coefficients of the s_i in the linear expression $\sum c_is_i$ appearing in the definition of C in Table 4. In the cases E_8, F_4, B_2, the values of the c_i have not been defined and are irrelevant. Let X and t be two variables with $X^{|Z|}=1$. Define

$$F(t)=\prod_{i=0}^{l}(1-X^{c_i}t^{n_i})^{-1}=\sum_{k=0}^{\infty}a_k(X)t^k, \tag{4.8}$$

where $a_k(X)=\sum_{j=0}^{|Z|-1}a_k^{(j)}X^j$. In the case $D_l=D_{2r}$ we introduce three variables X, Y, t with $X^2=Y^2=1$. Define

$$\begin{aligned}F(t)&=(1-t)^{-1}(1-Xt)^{-1}(1-t^2)^{1-r}(1-Xt^2)^{2-r}(1-XYt)^{-1}(1-Yt)^{-1}\\&=\sum_{k=0}^{\infty}a_k(X,Y)t^k,\end{aligned} \tag{4.9}$$

where

$$a_k(X,Y)=\sum_{0\leqq i,j\leqq 1}a_k^{(i,j)}X^iY^j.$$

PROPOSITION 4.1.

$$g(k)=\begin{cases}a_k^{(0)} & \textit{if } G \textit{ is not of type } D_{2r},\\ a_k^{(0,0)} & \textit{if } G \textit{ is of type } D_{2r}.\end{cases}$$

Proof. Suppose that G is not of type D_{2r}. The coefficient of X^0t^n in $f(t)$ is the number of $(l+1)$-tuples $(s_0, \cdots, s_l)$ such that

$$(4.10) \qquad \sum n_is_i = n, \qquad \sum c_is_i \equiv 0 \bmod |Z|.$$

For each such $(l+1)$-tuple let $g=\gcd(s_0, \cdots, s_l)$ and let $t_i = s_i/g$. Then $[t_0, \cdots, t_l]$ labels an EFO whose exact order is $m = (\sum n_it_i)C$ where C is the order of $\sum c_it_i \bmod |Z|$. Then $C \mid g$ and $m \mid n$.

Conversely suppose $[t_0, \cdots, t_l]$ is the label of an EFO x satisfying $x^n=1$ and exact order $m=(\sum n_it_i)C$, where C is the order of $\sum c_it_i \bmod |Z|$. Then $[t_0Cn/m, \cdots, t_lCn/m]$ satisfies (4.10).

The argument for D_{2r} is similar. □

Exact formulas for $g(k)$ are easy to work out. For example for C_l

$$(4.11) \qquad g(k) = \binom{l+\lfloor k/2 \rfloor}{\lfloor k/2 \rfloor},$$

where $\lfloor \cdot \rfloor$ is the greatest integer function. For others see [Dj2].

Although the Djoković formula is of limited applicability because of the restriction on k, its striking form suggests searching for a deeper relationship between the marks n_i and m_i.

An explicit generating function for the terms $a_k^{(0)}$ and $a_k^{(0,0)}$ can be deduced by using the character group $Z^\wedge$ of Z. Namely,

$$\sum_{k=0}^{\infty} g(k)t^k = \frac{1}{|Z^\wedge|} \sum_{\chi \in Z^\wedge} \prod_{i=0}^{l} (1-\chi(\omega_i^\wedge)t^{n_i})^{-1},$$

where $\omega_i^\wedge$ is considered as an element of Z by reduction modulo $Q^\wedge$ and the isomorphism of $P^\wedge/Q^\wedge$ and Z. In this form the formula is due to D. Djoković [Dj2] which he derived after seeing Proposition 4.1.

5. *K*-rationality. If $x \in G$ is of finite order N then for all representations (V, π), $\mathrm{ch}_V(x)$ lies in the cyclotomic field L_N generated by the Nth roots of 1. Suppose that K is a number field (finite extension of $\mathbb{Q}$). We say that x is *K-rational* (resp. *real*) if $\mathrm{ch}_V(X) \in K$ (resp. $\mathbb{R}$) for all representations (V, π).

Let $\mathrm{Gal}(L_N/\mathbb{Q})$ be the Galois group of L_N over $\mathbb{Q}$ and let H be the subgroup $\mathrm{Gal}(L_N/K \cap L_N)$. Let ω be a primitive Nth root of 1 and let

$$(5.1) \qquad \sigma_k : \omega \mapsto \omega^k, \qquad \gcd(k, N) = 1$$

be a typical element of H. We let H act on G by

$$(5.2) \qquad \sigma_k : x \mapsto x^k.$$

For x and $y \in G$ (resp. $x, y \in T$) let $x \sim_G y$ (resp. $x \sim_W y$) denote that x and y are G-conjugate (resp. W-conjugate). The following is the Lie group version of a well-known theorem of finite group theory.

PROPOSITION 5.1. *An EFO $x \in G$ of order N is K-rational if and only if $x \sim_G x^k$ for all k such that $\sigma_k \in H$.*

Proof. We may assume that $x \in T$. Let (V, π) be a representation of G. If the eigenvalues of $\pi(x)$ are $\varepsilon_1, \cdots, \varepsilon_n$ then those of $\pi(x^k)$ are $\varepsilon_1^k, \cdots, \varepsilon_n^k$. If $x \sim x^k$ whenever $\sigma_k \in H$ then $\sum \varepsilon_i = \mathrm{ch}_V(x) = \mathrm{ch}_V(x^k) = \sum \varepsilon_i^k$, whence $\mathrm{ch}_V(x) \in K$. Conversely, $\mathrm{ch}_V(x)^{\sigma_k} = \mathrm{ch}_V(x) \mapsto \mathrm{ch}_V(x^k) = \mathrm{ch}_V(x)$. However the characters are well known to generate the entire ring $\mathbb{C}[T]^W$ of W-invariant polynomial functions on T [St]. Since $\mathbb{C}[T]^W$ separates different W-orbits, $x^k \sim_W x$, $x^k \sim_G x$. □

Let us say that the *degree* of an EFO x is the degree over $\mathbb{Q}$ of the field $\mathbb{Q}\{x\}$ which is generated by the set of all its character values.

PROPOSITION 5.2. *For each positive integer k, the number of conjugacy classes of* EFO *of degree k is finite.*

Proof. Let (V, π) be a faithful finite dimensional representation of G and let $x \in G$ be EFO of order N and degree k. Let $K = \mathbb{Q}\{x\}$. We show that N is bounded by an integer depending on $k = \dim(K/\mathbb{Q})$ and $\dim V$.

The Galois group of $L_N/K \cap L_N$ is naturally isomorphic to a subgroup S of the group U_N of units of $\mathbf{Z}/N\mathbf{Z}$. Let $k_N = \dim_{\mathbb{Q}}(K \cap L_N)$ so that $\phi(N)/|S| = k_N$, where ϕ is the Euler function. Let M be a divisor of N. Under reduction mod M, U_N is mapped onto U_M and S onto some subgroup S_M of U_M with

$$k_N \geqq |U_M|/|S_M| = \phi(M)/|S_M|. \tag{5.3}$$

In particular S_M has at least $\phi(M)/k_N$ elements. Let ε be an eigenvalue of $\pi(x)$ and let the order of ε be M, $M|N$. Since $x \sim_G x^j$ for $j \in S$, we obtain at least $|S_M|$ eigenvalues from the powers of ε and $\phi(M)/k_N \leqq \dim V$. Thus $\phi(M) \leqq k \dim V$ and this bounds M, and then N, which is the least common multiple of such M. □

As far as K-rationality is concerned we need only consider subfields of cyclotomic fields, that is to say abelian extensions of $\mathbb{Q}$. If L_N is a cyclotomic field containing K then determination $\mathrm{Gal}(L_N/K)$ is sufficient to determine whether or not a given $\mathbf{s} = [s_0, \cdots, s_l]$ is K-rational. The most straightforward procedure is to use Proposition 5.1. The "powers" of $\mathbf{s}$ (that is the points $k\mathbf{s} = [ks_0, \cdots, ks_l]$, $k = 1, 2, 3, \cdots$) are brought back to $\mathbf{F}$ by a sequence of reflections. Specifically, if $\mathbf{x}$ is the result of some intermediate step and $\mathbf{x} \notin \mathbf{F}$ then one of the inequalities

$$\langle \alpha_i, \mathbf{x}\rangle \geqq 0, \quad i = 1, \cdots, l, \qquad \langle \xi, \mathbf{x}\rangle \leqq 1 \tag{5.4}$$

fails, and the corresponding reflection brings $\mathbf{x}$ closer to $\mathbf{F}$. In the case that $K = \mathbb{Q}(\sqrt{d})$, d a square free integer, $K \subset L_N$ if and only if $d|N$ or $4d|N$ according as $d \equiv 1$ or $d \not\equiv 1 \bmod 4$. Thus for example, only $\mathbb{Q}(\sqrt{-3})$, $\mathbb{Q}(\sqrt{-1})$, and $\mathbb{Q}(\sqrt{3})$ occur as quadratic subfields of L_{12}. $H := \mathrm{Gal}(L_{12}/\mathbb{Q}) \simeq 1, 5, 7, 11 \bmod 12$. With $\omega = e^{2\pi i/12}$ and $\sigma_k \in H$ defined by (5.1), $k = 1, 5, 7, 11$, the fixing subgroups of the quadratic fields are $\langle\sigma_7\rangle$, $\langle\sigma_5\rangle$, $\langle\sigma_{11}\rangle$ respectively. For $x \in G$ of order 12

$$\begin{aligned} x \text{ is: } & \mathbb{Q}\text{-rational} && \Leftrightarrow \; x \sim x^5 \sim x^7 \sim x^{11} \\ & \mathbb{Q}(\sqrt{-3})\text{-rational} && \Leftrightarrow \; x \sim x^7 \\ & \mathbb{Q}(\sqrt{-1})\text{-rational} && \Leftrightarrow \; x \sim x^5 \\ & \mathbb{Q}(\sqrt{3})\text{-rational} && \Leftrightarrow \; x \sim x^{11}. \end{aligned}$$

It is natural to prefer EFO with the simplest properties whenever one has the choice. Most often these are the rational EFO because their character values are integers. However, for groups with complex valued characters, i.e., A_k, D_{2k+1}, $k \geqq 2$, and E_6, the character values of rational elements do not distinguish pairs of contragredient representations. Then, in addition to the rational EFO, the preference should be given to the *Gaussian* ones. We say that an EFO is Gaussian if its character values are Gaussian integers ($\mathbb{Z}[\sqrt{-1}]$). An EFO is Gaussian if and only if $x \sim x^k$ whenever $\gcd(k, N) = 1$ and $k \equiv 1 \bmod 4$. The order N of a Gaussian EFO satisfies always $N \equiv 0 \bmod 4$. Properly Gaussian exist, of course, only when G admits non-selfcontragradient representations. The Gaussian EFO in A_2 and A_3 together with all their characters are found in [MPS].

As the order of $\mathbf{x}$ increases the number of reflections required to transfer the powers of $\mathbf{x}$ to $\mathbf{F}$ becomes prohibitive. For example

$$\begin{matrix} 1 \\ [00010101] \end{matrix}$$

required 716 reflections to determine that it was rational, a fairly typical number for EFO of order 12 in E_8. A more efficient way to compute this type of information is to calculate the character values $\mathrm{ch}_i(s)$ of $s = \exp(2\pi i \mathbf{s})$ on the set of fundamental representations corresponding to the "ends" of the Coxeter–Dynkin diagram [MP2]. Thus in E_6, for example, $\mathbf{s}$ is rational if and only if $\mathrm{ch}_i(s) \in \mathbb{Z}$ for $i = 1, 6$ (since $\mathrm{ch}_5(s) = \overline{\mathrm{ch}_1(s)}$).

The irreducible G-representation with highest weight λ is real if and only if $-w_0\lambda = \lambda$ where w_0 is the opposite involution. Correspondingly:

Proposition 5.3. *The* EFO $\mathbf{s}$ *is real if and only if* $-w_0\mathbf{s} = \mathbf{s}$.

Proof. Let (V, π) be a representation of G and let

$$\mathrm{ch}_V(s) = \sum \dim V^{\mu} \exp 2\pi i \langle \mu, \mathbf{s} \rangle, \tag{5.5}$$

where $V = \bigoplus V^{\mu}$ is the weight space decomposition of V and $s = \exp(2\pi i \mathbf{s})$. Then the reality of $\mathrm{ch}_V(s)$ is equivalent to

$$\sum \dim V^{\mu} \exp(-2\pi i \langle \mu, \mathbf{s} \rangle) = \sum \dim V^{\mu} \exp(2\pi i \langle \mu, \mathbf{s} \rangle),$$

which in turn is equivalent to

$$\sum \dim V^{\mu} \exp(2\pi i \langle \mu, -w_0\mathbf{s} \rangle) = \sum \dim V^{\mu} \exp(2\pi i \langle \mu, \mathbf{s} \rangle)$$

since $\dim V^{w_0\mu} = \dim V^{\mu}$. This is true for all representations of G if and only if $\mathbf{s} \sim_W -w_0\mathbf{s}$ which, since both are in the fundamental region $\mathbf{F}$, is equivalent to $\mathbf{s} = -w_0\mathbf{s}$.

This condition makes it obvious from the label whether or not the element is real. For example, in type A_l, $\mathbf{s} = [s_0, s_1, \cdots, s_l]$ is real if and only if $s_i = s_{l-i}$, $i = 1, 2, \cdots, l$. On the other hand for types B_l, C_l, D_{2n}, E_7, E_8, F_4, and G_2 all representations, and in particular all elements of finite order, are real.

The generating reflections $r_1, \cdots, r_l$ of W "appear" in G as elements of order 4 defined by

$$R_i = \exp e_{-\alpha_i} \exp -e_{\alpha_i} \exp e_{-\alpha_i},$$

where $\{e_{\alpha_i}, e_{-\alpha_i}, [e_{\alpha_i} e_{-\alpha_i}]\}$ is a standard $\mathfrak{sl}_2$-triplet for the root pair $\pm\alpha_i$, $i = 1, \cdots, l$. The R_i are rational and fall into 1 or 2 conjugacy classes depending on whether or not there are 1 or 2 root lengths. Table 3 identifies these classes for the algebras of type A, B, C, and G_2.

6. Regular elements. An element $x \in G$ is called *regular* if the centralizer of $\mathrm{Ad}(x)$ in $\mathrm{Ad}(G)$ has dimension $l(= \mathrm{rank}\, G)$. Now $s = \exp 2\pi i \mathbf{s}$ is regular if and only if $\mathbf{s}$ is off every hyperplane $\langle \alpha, \mathbf{x} \rangle = n$, $n \in \mathbb{Z}$, $\alpha \in \Delta$. If $\mathbf{s} = [s_0, s_1, \cdots, s_l]$ this is simply equivalent to having $s_i \neq 0$ for each i. It is then obvious that there is a unique class of elements of minimal Ad-order, namely the *Kostant principal elements* $[1, \cdots, 1]$ with Ad-order the *Coxeter number* $h = \sum_{i=0}^{l} n_i$. Since $\langle \alpha_i, [1, \cdots, 1] \rangle = 1/h$, $i = 1, \cdots, l$, $h[1, \cdots, 1]$ is the half sum $\rho^{\wedge}$ of the positive dual roots and hence the full order of the Kostant elements is h or $2h$ according as $\rho^{\wedge} \in Q^{\wedge}$ or not.

Similarly the only regular elements of Ad-order $h+1$ are the *extended principal* elements [Ka2] formed by permuting the labels of $[2, 1, \cdots, 1]$, where 2 is over a tip. These are clearly all Ad-conjugate. Since $\gcd(h+1, |Z|) = 1$ in all cases, there are always extended principal elements of full order $h+1$.

TABLE 3
Conjugacy classes of the elements R_i of § 5.

		$R_{\alpha_{\text{long}}}$	$R_{\alpha_{\text{short}}}$
A_1		[1 1]	
A_l	$l \geqq 2$	[2 1 0 ⋯ 0 1]	
B_l	$l \geqq 3$	[2 0 1 0 ⋯ 0]	[1 1 0 ⋯ 0 0]
C_l	$l \geqq 2$	[2 1 0 ⋯ 0]	[2 0 1 0 ⋯ 0]
G_2		[2 1 0]	[1 0 1]

From the condition for rationality given in § 5 it is clear that the Kostant elements and extended elements of order $h+1$ are rational. In fact as was shown by Kostant and Kac respectively, they only have character values 0, ±1 [Ko], [Ka2]. Strangely enough no one seems to have noted the following elementary result.

PROPOSITION 6.1. *For $x \in G$ of finite order the set of character values* $\mathrm{ch}_V(x)$, *as* (V, π) *runs through all irreducible representations, is finite if and only if x is regular.*

Proof. If $x = \exp 2\pi i \mathbf{x}$ is regular of finite order N then by Weyl's character formula the character values of x on the irreducible representation of highest weight Λ is

$$\sum_{w \in W} (-1)^{l(w)} e^{2\pi i \langle w(\Lambda+\rho), \mathbf{x}\rangle} / D(\mathbf{x}), \tag{6.1}$$

where $D(\mathbf{x}) \neq 0$. Since the exponential summands are Nth roots of 1 there are obviously only finitely many values for the numerator.

The converse is proved by Kac [Ka2] where it is only applied to an argument involving elements with character values 0, ±1.

7. Relations between subgroups. 1. The various inclusions of simple Lie groups amongst each other naturally allow one to consider the relationship between the conjugacy class of a given EFO in one group with its class as seen in an over-group.

In principle, explicit relationships between EFO of a group G and those of a subgroup $\tilde{G}$ can be built using the projection matrices discussed in [NP]. For maximal $\tilde{G}$ in G these are available in [McPS]. Relative to suitable maximal tori $\tilde{T}$ and T, the inclusion

$$i: \tilde{G} \to G \tag{7.1}$$

determines

$$\begin{aligned} &i: \tilde{T} \to T, \\ &di: \underline{\tilde{t}} \to \underline{t}, \\ &di|_{\tilde{Q}^{\wedge}}: \tilde{Q}^{\wedge} \to Q^{\wedge}. \end{aligned} \tag{7.2}$$

The transpose of the last mapping

$$(di)^{\circ}: P \to \tilde{P} \tag{7.3}$$

is the projection mapping whose matrix relative to suitable choices of fundamental weights is the projection matrix. If the matrix of di is written relative to the bases of fundamental coweights for $\underset{\sim}{\tilde{t}}$ and $\underset{\sim}{t}$ respectively, then one obtains an explicit label transformation for EFO. In general the resulting EFO is not in the fundamental region and has to be transformed there by a sequence of reflections which depends on the labels themselves. For this reason the method is not particularly suitable for human computation.

At least for the classical groups, it is easier to use specific standard matrix representations where the EFO are easy to write down and their labels easy to read. The proof of Proposition 7.1 and the subsequent remarks illustrate the method.

Table 4 lists some specific relationships of this type for $A_{2l-1} \subset A_{2l}$, $A_{2l} \subset A_{2l+1}$, $C_l \subset A_{2l}$, $B_{l-1} \subset B_l$, $C_{l-1} \subset C_l$. As an example of this we can follow the rational elements of A_1 through A_2, A_3, C_2 and C_3, as shown in Table 5.

TABLE 4
Identification of EFO *in various group-subgroup pairs.*

A_{2l}	(real EFO)	A_{2l-1}
$[s_0\ s_1\ s_2 \cdots \tfrac{1}{2}s_l\ \tfrac{1}{2}s_l \cdots s_2\ s_1]^{\cdot} \leftrightarrow [s_0\ s_1\ s_2 \cdots s_l \cdots s_2\ s_1]^{\cdot}$		

A_{2l+1}	(real EFO)	A_{2l}
$[s_0\ s_1 \cdots s_l\ 0\ s_{l+1} \cdots s_{2l}] \leftrightarrow [s_0\ s_1 \cdots s_{2l}]$		

A_{2l-1}	C_l
$[s_0\ s_1 \cdots s_{l-1}\ s_l\ s_{l-1} \cdots s_1] \leftrightarrow [s_0\ s_1 \cdots s_l]$	

C_l	C_l
$[s_0\ s_1 \cdots s_{l-2}\ \tfrac{1}{2}(s_{l-1})\ 0]^{\cdot} \leftrightarrow [s_0\ s_1 \cdots s_{l-1}]$	

B_l	B_{l-1}
$[s_0\ s_1 \cdots s_{l-1}\ 0] \leftrightarrow [s_0\ s_1 \cdots s_{l-1}]$	

TABLE 5

order	A_1	A_2	A_3	C_2	C_3
1	[1 0]	[1 0 0]	[1 0 0 0]	[1 0 0]	[1 0 0 0]
2	[0 1]	[0 1 1]	[0 1 0 1]	[0 1 0]	[0 1 0 0]
3	[1 2]	[1 1 1]	[1 1 0 1]	[1 1 0]	[1 1 0 0]
4	[1 1]	[2 1 1]	[2 1 0 1]	[2 1 0]	[2 1 0 0]
6	[2 1]	[4 1 1]	[4 1 0 1]	[4 1 0]	[4 1 0 0]

PROPOSITION 7.1. *Consider the inclusions*

$$C_l \subset A_{2l-1} \subset A_{2l}$$

determined by the natural representations of

$$USp(2l) \subset SU(2l) \subset SU(2l+1).$$

These inclusions determine bijective mappings between the conjugacy classes of rational (*resp. real*) EFO *of each order in each of these groups.*

Proof. Suppose that (V, π) is an irreducible representation of G. An element $x \in G$ of finite order N is rational if and only if $x \sim x^k$ for all k such that $\gcd(k, N) = 1$. This is equivalent to saying that primitive mth roots of unity amongst the eigenvalues of $\pi(x)$ occur in complete sets. Except for $m = 1, 2$, the value of Euler function $\phi(m)$ is even and the primitive mth roots of unity partition into distinct pairs ε, ε^{-1}. Given that $\pi(x)$ has determinant 1, the number of eigenvalues -1 must be even.

Now take $G = SU(n)$ and (V, π) the natural n-dimensional representation. By the usual conjugacy of $SU(n)$, one has

$$\begin{gathered}\pi(x) \sim \operatorname{diag}\{\exp 2\pi i k_1/N, \cdots, \exp 2\pi i k_n/N\}, \\ \sum_{i=1}^{n} k_i = 0, \quad k_1 \geqq \cdots \geqq k_n, \quad k_1 - k_n \leqq N.\end{gathered} \tag{7.4}$$

According to the remarks above, if x is also rational, then

$$k_i + k_{n+1-i} = 0. \tag{7.5}$$

We call the above diagonal matrix the canonical form of $\pi(x)$. It is indeed uniquely specified by the conjugacy class, as is seen by the way in which it determines the Kac coordinates below. If $n = 2p+1$ is odd then k_{p+1} is 0 and we have a direct identification with an element of $SU(2p)$. Since π is faithful this explains the bijection between the rational elements of these two groups.

In the case when $n = 2l$ is even, the canonical form is symplectic (relative to a symplectic basis $\nu_1, \cdots, \nu_{2l}$ with $(\nu_i, \nu_{2l+1-j}) = \delta_{ij}$). Symplectic elements of finite order (rational or not) may be conjugated to the canonical form (7.4), (7.5). Since the natural symplectic representation of compact C_l is faithful we have the second bijection of rational elements of Proposition 7.1.

The weight system of the natural n-dimensional representation of $SU(n)$ is

$$\omega_1, \omega_1 - \alpha_1, \cdots, \omega_1 - \alpha_1 - \cdots - \alpha_n.$$

For an element s of order N written in the canonical form (7.4) define

$$s_i' := (k_i - k_{i+1})/N, \qquad i = 1, 2, \cdots, n-1.$$

The least positive integer M such that $s_i := M s_i' \in \mathbb{Z}$ for each i is the Ad-order of s and

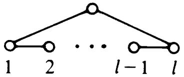
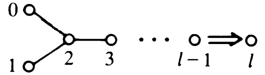
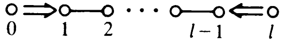
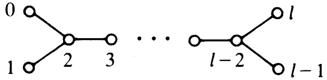
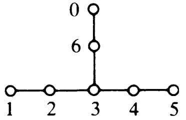
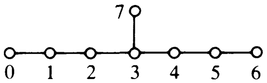
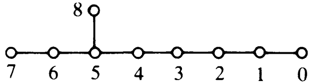
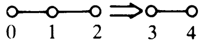
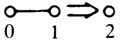
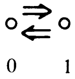
$$\mathbf{s} = [s_0, s_1, \cdots, s_{n-1}],$$

where $s_0 = M - \sum_{i=1}^{n-1} s_i$.

8. Computations of character tables. The previous sections have described the algorithms involved in the various tasks of selecting elements of finite order and evaluating the sums (1.9), (1.10). Supposing that $s_1, s_2, \cdots, s_p$ is a list of the elements of finite order for which the characters are to be evaluated, there remains the question of an efficient way in which to compute a character table. Of the various ways in which one can order the irreducible representations (or what is equivalent, their highest

TABLE 6

Numbering of the nodes of extended Dynkin diagrams; numerical marks associated with the nodes; action of the center Z or the Lie group on the tips of the diagram; the ratio C of the full order N to the adjoint order M of an EFO $[s_0, s_1, \cdots, s_l]$.

	$\tilde{\Gamma}$ and indexing	Numerical marks	Action of Z as a permutation on the tips		C
A_l $l \geqq 2$		$(1, 1, \cdots, 1)$	$(0\ 1\ 2\ \cdots\ l)$		$l+1/\gcd(l+1, \sum_{i=0}^{l} is_i)$
B_l $l \geqq 3$		$(1, 1, 2, \cdots, 2)$	$(0\ 1)$		$2/\gcd(2, \sum_{i \geqq 0} s_{2i+1})$
C_l $l \geqq 2$		$(1, 2, \cdots, 2, 1)$	$(0\ l)$		$2/\gcd(2, s_l)$
D_l $l \geqq 4$		$(1, 1, 2, \cdots, 2, 1, 1)$	$(0\ l\ 1\ l-1)$ $(0\ 1)(l-1\ l)$ $(0\ l)(1\ l-1)$	odd even	$4/\gcd(4, \sigma)$ $2/\gcd(2, s_{l-1}+s_l, \sigma/2)$ where $\sigma = 2(s_1+s_3+\cdots)+(l-2)s_{l-1}+ls_l$
E_6		$(1, 1, 2, 3, 2, 1, 2)$	$(0\ 1\ 5)$		$3/\gcd(3, s_1-s_2+s_4-s_5)$
E_7		$(1, 2, 3, 4, 3, 2, 1, 2)$	$(0\ 6)$		$2/\gcd(2, s_4+s_6+s_7)$
E_8		$(1, 2, 3, 4, 5, 6, 4, 2, 3)$	—		1
F_4		$(1, 2, 3, 4, 2)$	—		1
G_2		$(1, 2, 3)$	—		1
		$(1, 1)$	$(0\ 1)$		$2/\gcd(2, s_1)$

weights) the best for our purposes is the one which arises naturally out of the weight space multiplicity algorithm, that is, by increasing level. If $\lambda_1, \cdots, \lambda_q$ is the complete set of dominant weights in increasing level up to a certain point then the dominant weights which occur in the weight space decomposition of λ_k always lie amongst $\lambda_1, \cdots, \lambda_k$, $k=1,2,\cdots,q$. In addition the weight system of any irreducible module always lies in a single congruence class P modulo Q and hence we normally take $\lambda_1, \cdots, \lambda_q$ to be in a single class. The weight space multiplicities for the irreducible representations with these highest weights are then computed.

For a given pair (λ_j, s_k) we take $W_{(j)} = \mathrm{Stab}_W(\lambda_j)$ and $W_k = \mathrm{Stab}_W(s_k)$ and use the double coset algorithm to determine a list consisting of the pairs

$$\{(\langle w\lambda_j, s_k\rangle, n(w, k, (j))) \mid w \in {}^kW^{(j)}\}$$

(see § 3 for notation). Notice that the graph nodes selected by the double coset algorithm *are* the weights $w\lambda_j$. Hence it is not necessary to carry the actual Weyl group elements $w \in W^{(j)}$ in this process. Summing over the list we obtain the *orbit sums*

$$\sum(\lambda_j, s_k) := \sum n(w, k, (j))\, e^{2\pi i\langle w\lambda_j, s_k\rangle}.$$

These orbit sums $\sum(\lambda_j, s_k)$ are the primary constituents of the sums (1.10) which can now be computed all together. Note that by [Bo, Chap. VI, 3.4] the s_k is K-rational if and only if all orbit sums $\sum(\lambda, s_k)$ lie in K. Examples of character tables are found in Tables 7–10. For further methods in computing characters see [MP2].

9. Remarks.

1. Using the method of character generators on the low rank simple Lie groups, we have found, with R. T. Sharp, that the elements whose character values are restricted to 0 and ±1 are quite abundant among the regular rational elements. More precisely, the following is true [MPS].

Among the elements of finite order in the simple Lie groups of types A_1, A_2, A_3, B_2, and G_2 each of the regular rational elements has only three distinct character values, 0 and ±1, on irreducible representations of the Lie groups, with the exception of the following ones:

[411] in A_2	character values $\pm3, \pm2, \pm1, 0$
[6141], **[4161]** in A_3	character values $\pm4, \pm3, \pm2, \pm1, 0$
[112], **[641]**, **[461]** in B_2	character values $\pm2, \pm1, 0$
[141] in G_2	character values $\pm2, \pm1, 0.$

2. *Signatures.* In the case where one of the fundamental representations of G carries an invariant symmetric or hermitian bilinear form for some real form G' of $G_{\mathbb{C}}$ the characters can be used to determine the signatures of the induced forms on the various representations of G'. For example consider the groups $SU(p,q)$ which are defined as subgroups of $SL_n(\mathbb{C})$ by

$$\{X \in SL_n(\mathbb{C}) \mid AXA^{-1} = X^*\},$$

where X^* is the inverse conjugate transpose of X and

$$A = \mathrm{diag}\,\{\overset{p\text{-times}}{1, 1, \cdots, 1}, -1, \cdots, -1\}.$$

By definition, in its natural representation, each of these groups leave invariant a hermitian form of signature

$$p - q = \mathrm{tr}\, A.$$

TABLE 7

Relations between rational elements of A_2, C_2, A_3, and A_4; character values of rational elements of A_4 in the 13 *lowest representations. The Conway notation used in column* 6 *gives the following: The order of the element and a letter-name labeling its conjugacy class, subsequent letters designate the conjugacy classes of the prime powers of the element dividing its order.*

Rational finite order element of the nonregular A_1 in A_2	Rational finite order element of the regular A_1 in A_2	Rational finite order element in A_2	Rational finite order element in C_2	Rational finite order element in A_3	Name of class	Class of prime power, prime dividing order	Rational finite order element in A_4	Character on A_4 representations												
								(0005)	(0012)	(0110)	(1002)	(0004)	(0020)	(0101)	(0011)	(0003)	(1001)	(0002)	(0010)	(0001)
—	—	—	[614]	[6141]	12*C*	*BA*	[61221]	1	−1	5	0	−1	3	4	3	−1	3	1	3	2
—	—	—	[416]	[4161]	12*B*	*BB*	[41331]	−1	−1	1	0	1	1	0	−1	1	−1	−1	1	0
—	—	—	[121]	[1212]	12*A*	*EC*	[24114]	2	0	0	1	2	2	−2	0	2	0	2	−1	1
—	—	—	[211]	[2111]	10*A*	*AB*	[42112]	1	2	0	2	2	0	2	2	2	3	2	2	2
—	—	—	[111]	[1111]	8*A*	*C*	[22112]	0	1	−1	0	0	0	−1	0	1	0	1	0	1
[21]	—	—	[201]	[2010]	6*E*	*BB*	[40110]	−4	8	16	9	−2	10	12	10	2	8	4	5	3
—	[21]	[411]	[410]	[4101]	6*D*	*AA*	[41001]	27	39	33	32	21	21	24	21	15	15	9	7	4
—	—	—	[212]	[2121]	6*C*	*BA*	[21111]	0	0	0	−1	0	0	0	0	0	0	0	1	1
—	—	—	[014]	[0141]	6*B*	*AA*	[01221]	−3	3	−3	−4	3	3	0	−3	−3	3	3	1	−2
—	—	—	[011]	[0111]	6*A*	*AB*	[02112]	−1	−1	1	0	1	1	0	1	−1	−1	1	−1	0
—	—	—	[112]	[1121]	5*A*		[11111]	1	0	0	0	0	0	0	0	0	−1	0	0	0
[11]	—	—	[101]	[1010]	4*C*	*B*	[20110]	2	−3	3	−2	2	2	1	0	−1	0	−1	2	1
—	[11]	[211]	[210]	[2101]	4*B*	*A*	[21001]	14	11	7	12	10	4	9	8	7	8	5	4	3
—	—	—	[012]	[0121]	4*A*	*A*	[01111]	−2	−1	−1	0	2	0	1	0	−1	0	1	0	−1
[12]	—	—	[102]	[1020]	3*B*		[10110]	0	0	0	1	−2	2	0	−2	2	0	0	1	−1
—	[12]	[111]	[110]	[1101]	3*A*		[11001]	9	3	−3	4	7	−1	0	1	5	3	3	1	2
[01]	—	—	[001]	[0010]	2*B*		[00110]	−34	17	−5	−18	22	10	−3	−8	−13	8	7	2	−3
—	[01]	[011]	[010]	[0101]	2*A*		[01001]	6	−3	3	2	6	6	−3	0	3	0	3	−2	1
[10]	[10]	[100]	[100]	[1000]	1*A*		[10000]	126	105	75	70	70	50	45	40	35	24	15	10	5

TABLE 8

All rational EFO *of order* $\leqq 8$ *in* E_6, *Conway notation* (cf. *Table* 7), *and characters in the* 12 *lowest representations of the congruence class* 0.

CHARACTERS OF ALL RATIONAL ELEMENTS OF ORDER $\leqslant$ 8 OF E6   1 2 $\overset{6}{3}$ 4 5

	RATIONAL	REPRESENTATIONS OF E6 CLASS 0											
	ELEMENTS	(000001)	(100010)	(001000)	(000002)	(000110)	(000030)	(100011)	(010100)	(200100)	(001001)	(200020)	(000003)
B	8H (0 1 1 0 1 1 1)	0	0	-1	0	0	-1	1	0	1	0	-1	0
A	8G (0 0 1 0 1 0 2)	-2	2	-3	6	0	1	-3	6	-3	0	5	-10
E	8F (3 0 0 1 0 0 1)	16	64	111	64	160	73	449	432	487	320	495	64
E	8E (1 1 0 1 0 1 1)	0	0	-1	0	0	1	1	0	-1	0	-1	0
C	8D (1 0 0 1 0 0 2)	0	8	-17	8	0	19	1	0	-19	0	63	8
A	8C (2 1 0 0 0 1 2)	6	18	5	6	16	21	29	-34	25	-32	61	22
A	8B (0 2 1 0 1 2 0)	6	2	5	22	-16	-19	29	-34	-15	32	61	38
D	8A (2 1 1 0 1 1 0)	2	-2	1	2	0	1	-3	2	1	0	1	2
	7B (2 0 0 1 0 0 1)	8	27	20	8	35	28	64	0	71	-37	117	1
	7A (1 1 1 0 1 1 0)	1	-1	-1	1	0	0	1	0	1	-2	-2	1
BB	6K (0 2 0 0 0 2 1)	7	1	10	30	-18	-19	24	-38	-24	72	81	67
CA	6J (2 0 1 0 1 0 0)	5	10	9	0	8	-2	0	13	0	-8	0	-6
AA	6I (4 1 0 0 0 1 0)	38	217	681	507	1112	463	4869	7630	7419	9712	6912	3459
BB	6H (1 0 0 1 0 0 1)	1	7	-8	0	0	11	0	-8	0	0	27	7
AB	6G (0 0 1 0 1 0 1)	-2	1	1	3	0	-1	-3	-2	3	0	0	-5
CB	6F (3 0 0 1 0 0 0)	25	118	289	192	432	170	1536	2041	1920	2160	1728	598
AA	6E (0 1 1 0 1 1 0)	2	1	-3	3	-4	-5	9	-2	3	-8	0	3
BB	6D (4 0 0 0 0 0 1)	55	385	1450	1134	2646	1253	13608	24346	25920	33480	26433	12643
AB	6C (2 1 0 0 0 1 1)	10	25	37	27	36	11	81	46	27	72	0	43
CB	6B (1 1 0 1 0 1 0)	1	-2	1	0	0	2	0	1	0	0	0	-2
BA	6A (0 1 0 0 0 1 2)	-1	1	-6	6	2	1	0	10	-12	-8	9	3
	5A (1 1 0 0 0 1 1)	3	0	0	5	-1	3	-1	-5	0	0	-7	8
B	4E (1 0 0 1 0 0 0)	2	6	1	-6	0	3	-3	2	3	0	9	-6
A	4D (0 0 1 0 1 0 0)	-2	2	-3	6	0	-1	-3	6	3	0	-3	-10
B	4C (2 0 0 0 0 0 1)	34	190	545	386	896	399	3709	5474	5615	6400	5481	2050
B	4B (0 1 0 0 0 1 1)	2	-2	1	2	0	-1	-3	2	-1	0	9	2
A	4A (2 1 0 0 0 1 0)	14	34	77	70	64	7	189	182	27	384	-83	246
	3C (0 0 0 1 0 0 0)	-3	2	9	0	-8	6	0	5	0	-24	0	18
	3B (1 0 0 0 0 0 1)	15	65	90	54	154	105	432	266	540	120	729	99
	3A (1 1 0 0 0 1 0)	6	-7	9	27	-8	15	-27	14	27	48	0	99
	2B (0 0 0 0 0 0 1)	-2	10	-35	30	0	35	-3	70	-105	0	189	-50
	2A (0 1 0 0 0 1 0)	14	10	45	126	-64	-101	189	-266	-225	640	621	750
	1A (1 0 0 0 0 0 0)	78	650	2925	2430	5824	3003	34749	70070	78975	105600	85293	43758

If either p or q is even they can be interchanged without harm if necessary so that q is even and $A \in SU(n)$. Assuming $q \neq 0$ the canonical form (7.4) of the element A leads to the coordinates

$$[0\ \ 0\ \cdot\cdot\cdot\cdot\ 0\ \ 1\ \ 0\ \cdot\cdot\cdot\cdot\ 0\ \ 1\ \ 0\ \cdot\cdot\cdot\cdot\ 0]$$

where the two 1's are in the $q/2$ and $l-q/2+1$ positions. Here $l+1=n$. The character of A in the natural representation is the signature $p-q$ and it follows that its character on every representation is the corresponding signature.

If p and q are odd the situation is more complicated. Let $\xi = e^{\pi i/n}$ so that det $\xi A = 1$. Then ξA has the canonical form (7.4)

$$\xi A = \text{diag}\,\{\exp \pi i k_1/n, \cdots, \exp \pi i k_n/n\},$$

TABLE 9

All rational EFO *of order* ≦8 *in* E_6, *Conway notation* (cf. *Table* 7), *and characters in the* 12 *lowest representations of the congruence class* 2 (*and also class* 1 *through diagram symmetry*).

CHARACTERS OF ALL RATIONAL ELEMENTS OF ORDER ⩽ 8 OF E6

$1\ 2\ \overset{6}{3}\ 4\ 5$

	RATIONAL	REPRESENTATIONS OF E6 CLASS 2													
	ELEMENTS	(000010)	(010000)	(200000)	(000011)	(100100)	(100020)	(010001)	(200001)	(001010)	(000012)	(000200)	(110010)	(000120)	(300010)
B	8H (0 1 1 0 1 1 1)	-1	1	1	0	-1	0	0	-1	1	0	0	0	0	1
A	8G (0 0 1 0 1 0 2)	1	-3	3	0	-1	2	6	-5	-5	0	6	0	-2	3
E	8F (3 0 0 1 0 0 1)	9	39	33	96	207	192	240	249	417	240	192	672	336	375
E	8E (1 1 0 1 0 1 1)	1	-1	1	0	-1	0	0	1	1	0	0	0	0	-1
C	8D (1 0 0 1 0 0 2)	3	-3	9	0	-9	24	0	3	-39	24	24	0	0	45
A	8C (2 1 0 0 0 1 2)	5	9	11	16	15	34	-2	31	-13	16	-10	16	38	63
A	8B (0 2 1 0 1 2 0)	-3	1	11	-16	15	-14	-2	39	-13	-48	38	-16	38	7
D	8A (2 1 1 0 1 1 0)	1	1	-1	0	-1	-2	2	-1	-1	0	2	0	2	3
	7B (2 0 0 1 0 0 1)	6	15	15	27	42	57	15	48	21	6	0	75	63	105
	7A (1 1 1 0 1 1 0)	-1	1	1	-1	0	1	1	-1	0	-1	0	-2	0	0
BB	6K (0 2 0 0 0 2 1)	-3	0	12	-18	21	-15	-12	51	-21	-60	54	-18	33	-6
CA	6J (2 0 1 0 1 0 0)	4	8	4	8	12	4	4	-4	4	-12	8	8	-4	-8
AA	6I (4 1 0 0 0 1 0)	16	128	112	464	1440	1312	2704	2576	6160	4944	3584	10832	4832	4720
BB	6H (1 0 0 1 0 0 1)	3	0	6	0	-3	15	-6	3	-15	6	0	0	3	24
AB	6G (0 0 1 0 1 0 1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CB	6F (3 0 0 1 0 0 0)	12	72	60	216	564	492	876	804	1740	1212	936	2808	1236	1176
AA	6E (0 1 1 0 1 1 0)	-2	2	4	-4	0	-2	4	2	4	-6	2	-4	8	4
BB	6D (4 0 0 0 0 0 1)	21	216	204	918	3381	3345	7140	7419	18795	15756	11718	37206	17409	18690
AB	6C (2 1 0 0 0 1 1)	6	18	12	36	48	30	60	42	60	66	18	36	24	12
CB	6B (1 1 0 1 0 1 0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BA	6A (0 1 0 0 0 1 2)	1	-4	4	2	-3	1	4	-1	-5	0	14	2	-7	-2
	5A (1 1 0 0 0 1 1)	2	1	1	3	-4	-3	0	5	0	7	-2	-5	4	0
B	4E (1 0 0 1 0 0 0)	3	3	3	0	3	6	-6	-3	3	-12	-6	0	-6	9
A	4D (0 0 1 0 1 0 0)	-1	3	-1	0	-5	2	-2	5	7	-4	6	0	-10	-3
B	4C (2 0 0 0 0 0 1)	15	111	99	384	1155	1086	2010	2001	4515	3396	2562	8064	3738	3885
B	4B (0 1 0 0 0 1 1)	-1	-1	3	0	3	-2	-6	1	3	4	2	0	-6	-3
A	4A (2 1 0 0 0 1 0)	7	27	15	64	91	34	174	77	183	252	70	64	6	-43
	3C (0 0 0 1 0 0 0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3B (1 0 0 0 0 0 1)	9	36	36	90	189	225	180	279	315	216	126	666	441	630
	3A (1 1 0 0 0 1 0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2B (0 0 0 0 0 0 1)	3	-9	15	0	-21	42	30	-15	-105	60	126	0	-42	105
	2A (0 1 0 0 0 1 0)	-5	-1	31	-64	75	-86	-50	329	-185	-516	350	-64	230	145
	1A (1 0 0 0 0 0 0)	27	351	351	1728	7371	7722	17550	19305	51975	46332	34398	112320	54054	61425

where

$$(k_1,\cdots,k_n)=(\underbrace{n+1,\cdots,n+1}_{\frac{q-1}{2}},\underbrace{1,\cdots,1}_{p},\underbrace{1-n,\cdots,1-n}_{\frac{q+1}{2}}).$$

The corresponding element of t is

$$s=[\underbrace{0,\cdots,0}_{\frac{q-1}{2}},1,\underbrace{0,\cdots,0}_{p-1},1,\underbrace{0,\cdots,0}_{\frac{q-1}{2}}].$$

The element $\xi 1$ lies in $SU^{\pm 1}(n)$ (unitary matrices of determinant ±1) where it is central. Its square, $\xi^2 1 \in SU(n)$, has coordinates $[0,\cdots,0,1]$. If (V,π) is any irreducible representation of $SU(n)$ then $\xi^2 1$ acts as a scalar (phase factor) γ on V and the representation can be extended to one of $SU^{\pm}(n)$ by choosing ξ to act as one of the

TABLE 10

List of all rational EFO *in G_2, corresponding Conway notation* (cf. *Table* 7), *and characters in the lowest representations. Characters of the two conjugacy classes of quadratic elements of $PSL(2, 13)$ of order* 13 *in G_2.*

CHARACTERS OF ALL RATIONAL ELEMENTS OF FINITE ORDER OF G2 1⇛2

	RATIONAL	REPRESENTATIONS OF G2 CLASS 0																			
	ELEMENTS	(00)	(01)	(10)	(02)	(11)	(03)	(20)	(12)	(04)	(21)	(13)	(30)	(05)	(22)	(14)	(06)	(31)	(23)	(15)	(40)
	1A (1 0 0)	1	7	14	27	64	77	77	189	182	286	448	273	378	729	924	714	896	1547	1728	748
	2A (0 1 0)	1	-1	-2	3	0	-3	5	-3	6	-2	0	-7	-6	9	-4	10	0	-5	0	12
	3A (1 1 0)	1	1	-1	0	-2	2	-1	0	2	1	-2	3	0	0	-3	3	2	-1	0	-2
	3B (0 0 1)	1	-2	5	0	-8	5	14	0	-7	-20	16	30	0	0	-21	12	-40	35	0	55
A	4A (2 1 0)	1	3	2	3	0	1	-3	-3	2	-6	0	-3	6	-3	4	6	0	-1	0	4
A	4B (1 0 1)	1	-1	2	-1	0	1	1	-3	2	2	0	1	-2	-3	4	-2	0	-1	0	4
AA	6A (4 1 0)	1	5	7	12	18	18	11	24	18	13	18	-1	12	0	5	7	-18	-17	0	-18
BA	6B (3 0 1)	1	2	1	0	0	-3	2	0	-3	4	0	2	0	0	-1	4	0	-5	0	3
AA	6C (1 1 1)	1	-1	1	0	0	0	-1	0	0	1	0	-1	0	0	-1	1	0	1	0	0
	7A (2 1 1)	1	0	0	-1	1	0	0	0	0	-1	0	0	0	1	0	0	0	0	-1	-1
B	8A (3 1 1)	1	1	0	-1	0	-1	1	1	0	0	0	-1	0	-1	0	0	0	1	0	0
A	8B (1 2 1)	1	-1	0	1	0	-1	-1	1	0	0	0	1	0	-1	0	0	0	1	0	0
BA	12A (3 3 1)	1	0	-1	0	0	1	0	0	-1	0	0	0	0	0	1	0	0	-1	0	1
AB	12B (1 4 1)	1	-1	-1	2	0	-2	1	0	2	-1	0	1	-2	0	1	1	0	-1	0	-2
	13A (6 1 2)		a	1	1	-1	-1	-1	$\bar{a}$	0	0	$-\bar{a}$	0	1	1	1	-1	-1	0	-1	$\bar{a}$
	13B (1 2 3)		$\bar{a}$	1	1	-1	-1	-1	a	0	0	$-a$	0	1	1	1	-1	-1	0	-1	a

$a = 1/2 + \sqrt{13}/2$; $\bar{a} = 1/2 - \sqrt{13}/2$

square roots $\gamma^{1/2}$ of γ. The signature of the invariant hermitian form on V (if any) induced from the canonical form associated with $SU(p, q)$ is, up to sign, $\mathrm{tr}_V(A) = \mathrm{tr}_V(\xi A)/\gamma^{1/2} = \mathrm{ch}_V(\xi A)/\gamma^{1/2}$. Thus the signature is expressed entirely in terms of the characters of the EFO $[1, 0, \cdots, 0]$, $[0, \cdots, 0, 1]$, and $\mathbf{s}$:

$$\text{signature} = \mathrm{ch}_V(\xi A)/(\mathrm{ch}_V(\xi^2 1)/\mathrm{ch}_V(1))^{1/2}.$$

Explicit formulas have been established for $SU(p, q)$ signatures in [PS].

3. *Congruence classes.* The weight system of an irreducible module V always lies in a single coset of the weight lattice P modulo the root lattice Q. This is the *congruence class* $[LP]$ of V. If $z = \exp 2\pi i\mathbf{z}$ is a central element of G then it acts on V by $e^{2\pi i\langle\Lambda,\mathbf{z}\rangle}$ where Λ is the highest weight of V and this value appears as $\mathrm{ch}_V(z)/\mathrm{ch}_V(1)$. Since $Z \simeq P^\wedge/Q^\wedge$ via $z \to z \bmod Q^\wedge$, Z separates the various congruence classes and one may read the congruence class from the character values $\mathrm{ch}(z)$, $z \in Z$. Of course, except in the cases D_{2n}, a single generator of $P^\wedge/Q^\wedge$ will suffice. Recall that in § 4 the central elements of G were identified.

4. *Gradings and* EFO. Let M be a positive integer. A mod M *grading* on a complex Lie algebra $\mathfrak{g}$ is a decompositions

$$\mathfrak{g} = \bigoplus_{i \in \mathbb{Z}/M\mathbb{Z}} \mathfrak{g}_i \quad \text{with } [\mathfrak{g}^i, \mathfrak{g}^j] \subset \mathfrak{g}^{i+j}. \tag{9.1}$$

Here it is often convenient to view i as an integer rather than a congruence class modulo M. If ξ is a primitive Mth root of unity then the linear mapping

$$\phi\colon \mathfrak{g} \mapsto \mathfrak{g}, \qquad \phi|\mathfrak{g}^i = \xi^i$$

is an automorphism of $\mathfrak{g}$ and is of order M if $\gcd\{i|\mathfrak{g}^i \neq 0\} = 1$. In [Ka1] V. Kac gave a beautiful classification of all such gradings on semi-simple Lie algebras. The part of this that is relevant here is the classification of the gradings on a simple Lie algebra $\mathfrak{g}$ for which ϕ is an inner automorphism. In this case ϕ is clearly induced by an EFO of the corresponding simply connected compact group G. Precisely, such a ϕ always point-wise fixes a Cartan subalgebra and there is a choice $\Pi = \{\alpha_1, \cdots, \alpha_l\}$ of the base of the corresponding root system Δ such that the root space $\mathfrak{g}^{\alpha_i}$ lies in $\mathfrak{g}^{s_i}$ where $0 \leqq s_i < M$. Furthermore letting $n_0, n_1, \cdots, n_l$ be the numerical marks, the equation $\sum n_i s_i = M$ defines s_0 and $\mathbf{s} = [s_0, s_1, \cdots, s_l]$ is an EFO. This element induces ϕ.

The element $\mathbf{s}$ induces on any $\mathfrak{g}$-module V a grading compatible with that on $\mathfrak{g}$:

$$V = \bigoplus_{j\in \mathbb{Z}/N\mathbb{Z}} V^j, \qquad \mathfrak{g}^i \cdot V^j \subset V^{i+j} \tag{9.2}$$

by using the eigenspaces of $s = \exp 2\pi i\mathbf{s}$ as it acts on V. Here N is the full order of s.

If s induces a grading on V then specific information like the dimensions of the homogeneous subspaces of V of a given degree are evidently computable from the character value of s and its powers. Indeed

$$\mathrm{ch}_V(s) = \sum \dim V^j e^{2\pi ij/N}, \qquad \dim V^k = 1/N \sum_{j=0}^{N-1} \mathrm{ch}_V(s^j)\, e^{-2\pi ijk/N}. \tag{9.3}$$

5. *Additional Weyl group symmetry.* In computing an orbit sum $\sum(\Lambda, s)$ we take into account the repetition of terms due to the presence of the stabilizing group W_K (see § 1). If $\mathbf{s} = [0, s_1, \cdots, s_l]$ then it is easy to see that $\langle \xi, \mathbf{s}\rangle = 1$ and $\langle r_\xi w\Lambda, \mathbf{s}\rangle = \langle w\Lambda, \mathbf{s}\rangle \bmod \mathbb{Z}$ for all $w \in W$. Here ξ is the highest root of Δ^+. This determines "linking" of various W_K-orbits of the coset graph Γ. In principle it is easy to detect when linking occurs. Indeed, let $\xi^\wedge$ denote the highest short root of the dual root system, so that $r_\xi\colon \mu \mapsto \mu - \langle \mu, \xi^\wedge\rangle \xi$. Let λ^- denote the deepest element in some W_K-orbit of Γ and let λ be another element of the same orbit. Then λ links to a higher element if and only if $\langle \lambda, \xi^\wedge\rangle < 0$. When this happens $\langle \lambda^-, \xi^\wedge\rangle < 0$ and λ^- links to a higher element. Furthermore, $r_\xi\lambda^- = \lambda^- - \langle\lambda^-, \xi^\wedge\rangle\xi$ cannot be in the same W_K-orbit as λ^- since $\xi \notin \sum_{i\in K} \mathbb{Z}\alpha_i$. Thus simply determining the sign of $\langle \lambda^-, \xi^\wedge\rangle$ is sufficient to detect that an orbit is linked to at least one other orbit (in general more than one). However, from the point of view of computing, the time required to locate the elements λ^- and to keep track of the various linkings seems to offset the savings which otherwise would be afforded. For this reason we have not utilized this symmetry in our work.

6. *The finite group PSL* (2, 13) *can be embedded in* G_2 [Me]. Using our Table 10 and a character table *PSL* (2, 13) (e.g. [McK]) it is easy to identify the G_2-conjugacy classes of *PSL*(2, 13):

	PSL (2, 13)	G_2
order of conjugacy class	2	[0 1 0]
	3	[1 1 0]
	6	[1 1 1]
	7	[2 1 1]
	13	[6 1 2], [1 2 3]

The three conjugacy classes of elements of order 7 in $PSL(2, 13)$ are G_2-conjugate. There are exactly two embeddings of PSL (2, 13) in G_2 (we are grateful to A. Meurman for showing us a proof of this fact) which are related by the outer automorphism of PSL (2, 13). These interchange the two elements of order 13.

Acknowledgments. We would like to express our appreciation to W. McKay for her valuable part in the development of the programs for this work and for her cheery disposition through all its various vicissitudes. Also we would like to thank J. McKay, L. Solomon, N. Iwahori, T. Yokonuma, and D. Djoković for their useful and friendly advice, and M. Bremner for the timings of Table 2. We are also grateful for the hospitality of the Aspen Center for Physics where this work was finished.

Finally, we would like to thank the referees of this paper, whose critical remarks were most useful.

REFERENCES

[Bo] N. Bourbaki, *Groupes et algèbres de Lie. Ch.* IV, V, VI, Hermann, Paris, 1968.

[BMP] M. Bremner, R. V. Moody and J. Patera, *Tables of Dominant Weight Multiplicities for Simple Lie Algebras of Ranks* $\leqq 8$, Marcel Dekker, New York, to be published.

[CQ] J. Conway and L. Queen, *Computing the character table of a Lie group* (to appear in the proceedings of the conference on finite groups, Montreal, 1982).

[Dj1] D. Djoković, *On conjugacy classes of elements of finite order in compact or complex semi-simple Lie groups*, Proc. AMS, 80 (1980), pp. 181–184.

[Dj2] ———, *On conjugacy classes of elements of finite order in complex semisimple Lie groups*, to appear.

[Dy] E. B. Dynkin, *Maximal subgroups of the classical groups*, Trudy. Moskov. Mat. Osc. 1, 39–166 (1952); Transl. Amer. Math. Soc. Ser. 2, Vol. 6 (1957); Supplement #23.

[Ka1] V. Kac, *Automorphisms of finite order of semi-simple Lie algebras*, Funct. Anal. and Appl., 3 (1969), pp. 252–254.

[Ka2] ———, *Simple Lie groups and the Legendre symbol*, in Algebra, Carbondale (1980), Lecture Notes in Mathematics 848, Springer-Verlag, New York, 1981, pp. 110–123.

[Ko] B. Kostant, *On Macdonald's η-function formula, the Laplacian and generalized exponents*, Adv. Math., 20 (176), pp. 179–212.

[LP] F. Lemire and J. Patera, *Congruence number, a generalization of $SU(3)$ triality*, J. Math. Phys., 21 (1980), pp. 2026–2027.

[McK] J. McKay, *The non-abelian simple groups G, $|G| < 10^6$-character tables*. Comm. Alg., 7 (1979), pp. 1407–1445.

[McPS] W. McKay, M. Patera and D. Sankoff, *The computation of branching rules for representations of semisimple Lie algebras*, in Computers in Non-associative Rings and Algebras, R. Beck and B. Kolman, eds., Academic Press, New York, 1977.

[Me] A. Meurman, *An embedding of $PSL(2,13)$ in $G_2(\mathbb{C})$*, in Lie Algebras and Related Topics, (New Brunswick) (1981), Lecture Notes in Mathematics 933, Springer-Verlag, 1982, pp. 157-165.

[Mi] W. Miller, Jr., *Symmetry Groups and Their Applications*, Academic Press, New York, 1972.

[MP1] R. V. Moody and J. Patera, *Fast recursion formula for weight multiplicities*, Bull. AMS, 7 (1982), pp. 237–242.

[MP2] ———, *Computing character decompositions of class functions on compact semi-simple Lie groups*, to appear.

[MPi] R. V. Moody and A. Pianzola, *λ-mappings between representation rings of Lie algebras*, Canad. J. Math., 35 (1983), pp. 898–960.

[MPS] R. V. Moody, J. Patera and R. T. Sharp, *Character generators for elements of finite order in simple Lie groups A_1, A_2, A_3, B_2, and G_2*, J. Math. Phys., 24 (1983), pp. 2387–2397.

[PS] J. Patera and R. T. Sharp, *Signatures of all finite representations of $SU(p, q)$*, J. Math. Phys., 25 (1984), to appear.

[PSS] J. Patera, R. T. Sharp and R. Slansky, *On a new relation between semisimple Lie algebras*, J. Math. Phys., 21 (1980), pp. 2335–2341.

[Sla] R. Slansky, *Group theory for unified model building*, Physics Reports, Vol. 79, 1981.

[Sl] P. SLODOWY, *Simple singularities and simple algebraic groups*, Lecture Notes in Mathematics 815, Springer-Verlag, New York, 1980).

[Sta] J. STACK, *Heavy quark potential in $SU(2)$ lattice gauge theory*, Phys. Rev. D, 27 (1983), pp. 412–420.

[St] R. STEINBERG, *Conjugacy Classes in Algebraic Groups*, Lecture Notes in Mathematics 366, Springer-Verlag, New York, 1974.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

009

ON OPERATOR AND FORMAL SUM METHODS FOR GRAPH ENUMERATION PROBLEMS*

C. J. LIU† AND YUTZE CHOW‡

Abstract. This article investigates some computational aspects of the operator approaches recently introduced by the authors for graph enumeration problems involving forests, subgraphs with fixed cyclomatic numbers and matchings.

AMS(MOS) subject classifications. Primary 05C05, 05C30; secondary 05A15

Introduction. "Trees", i.e. a connected graph without any cycle, have to be considered to be one of the most important concepts in graph theory. Besides its pure graph-theoretical aspect, its usefulness in applications ranges from electric network theory in engineering to particle and solid-state theory in physics. The first natural question about "trees" is undoubtedly that of how many *spanning trees* are there, as subgraphs, in a given graph (labelled, say). This elementary question was, as well-known, answered long ago (in 1847) by G. Kirchhoff. The next questions one may ask, for a (labelled) graph G, are:

(i) What is the number of spanning m-forests in G?

(ii) What is the number of spanning subgraphs, with a given cyclomatic number, in G?

By an *m-forest* in G we mean a spanning subgraph consisting of m connected components each being a tree. By the *cyclomatic number*, for a connected linear graph, we mean the number

$$|E(G)|-|V(G)|+1$$

where $E(G)$ and $V(G)$ are, respectively, the sets of all edges and vertices in G. $|S|$ denotes the *cardinality* of the set S.

Recently we have introduced an operator approach to the two questions raised above. Indeed, the operator method has successfully resolved the first question above. The m-forest enumeration, for any graph and any m, can now be carried out by a straightforward computation [1]. As to the second question, it was recently resolved too by the operator approach, for any *planar* graph [2]. In both references [1] and [2], however, the formalisms and the results were formally presented without due concern for the computational aspect of the problems. The purpose of this article is to pursue this particular phase of our operator methods. In particular, we want to show the graph-theoretical meanings at different stages of the operator manipulations and also the possibility of reducing the computational complexity through the recurrent use of certain functions we introduced.

Before going into the details of the planned exposition we shall give an example showing directly the advantage of the use of operators. Consider the long-standing enumeration problem of "Hamilton cycles." Let $\mathbf{K}=\|k_{ij}\|$ be the Kirchhoff matrix of

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, Wisconsin 53233.

‡ Department of Physics, University of Wisconsin, Milwaukee, Wisconsin 53201. The research of this author was partially supported by the Graduate School, University of Wisconsin.

a connected graph. Define operators $x_1, \cdots, x_n$ $(n = |V(G)|)$ satisfying

$$x_i x_j = x_j x_i, \qquad x_i^2 = 0, \qquad j, i = 1, \cdots, n.$$

Let $k'_{ij} \equiv k_{ij} x_j$ and $k'_{ii} = -\sum_{j \neq i} k'_{ij}$. Then the total number of Hamiltonian cycles H is given by the relation

$$\left(\prod_{j=1}^{n} x_j\right) H = \frac{1}{2} k'_{ii} \det \mathbf{K}'(i), \qquad i = 1, \cdots, n$$

where $\mathbf{K}'(i)$ is the ith principal minor of $\mathbf{K}' \equiv \|k'_{ij}\|$. One can be easily convinced by going through some examples that the above operator expression usually cuts down drastically the computation that would otherwise be involved in the ordinary (operator-free) formalism, e.g.

$$H = \frac{1}{2} \sum_{\{P\}} a_{ip_1} a_{p_1 p_2} \cdots a_{p_{n-1} i} \qquad i = 1, \cdots, n$$

where $P \equiv (p_1, \cdots, p_{n-1})$ is a permutation of $\{1, \cdots, i-1, i+1, \cdots, n\}$ and a_{ij} are the entries of the adjacency matrix $\mathbf{A}$ of the graph.

Our operator method, often considered to be rather abstract by some graph theorists, was originally developed entirely from a problem-solving viewpoint: in tackling the problems in a particular way we found certain steps and abstractions to be indispensable in order to make progress. It is, therefore, a fact that the operator method, though it requires certain abstraction, was not artificially created for the sake of abstraction itself. It should also be mentioned that though we found it necessary to introduce certain operators in tackling the problems, yet the final formal solutions *can* nevertheless be expressed in a form free of any operator.

To define these operators it is necessary to define the vector space on which they operate. In fact, these vector spaces are spaces of *formal sums*, with real coefficients, over certain sets related to the given graph. It also turns out that the operators possess very simple commutation relations that effectively simplify the explicit calculations. The final stage of the operator method involves a real-valued linear function on the formal vector space. This linear function plays an important role in the whole approach; whether the operator method is applicable to a particular problem may depend on the successful search of such a function in each case.

Our discussions in this article will be divided into six parts. In the first section we lay down the necessary definitions and notions related to the problems with some examples. In the second section we concentrate on the "matrix-forest" theorem, i.e. the theorem leading to any m-forest enumeration [1]. The main task here is to show how the rather abstract method works by analyzing some actual computations. In the third section we concentrate specifically on the complete graphs, bipartite complete graphs and tripartite complete graphs in forest enumerations. The fourth section deals with the enumeration of subgraphs with a preassigned cyclomatic number, in a planar graph. The fifth section explores some further extensions and alternatives to the operator approaches. In the last section we introduce a useful function (no operators involved) which handles efficiently the enumerations of (i) m-forests, (ii) spanning subgraphs with a preassigned cyclomatic number of a planar graph, and (iii) r-matchings (i.e. a collection of r independent edges in the graph).

1. Kirchhoff matrix and cycle-adjacency matrix. For convenience, we assume that the given graphs are connected, labelled and undirected. The *adjacency matrix* of a given graph is denoted by $\mathbf{A}$ with entries a_{ij}. Labelling of all the vertices in the graph, i.e. $V(G)=\{V_1,\cdots,V_n\}$, establishes a $1-1$ correspondence between $V(G)$ and the set $\mathbf{n}=\{1,\cdots,n\}$ and also fixed a_{ij}.

As a computational example, we shall use exclusively the graph of Fig. 1, to be hereafter referred to as the "kite" graph.

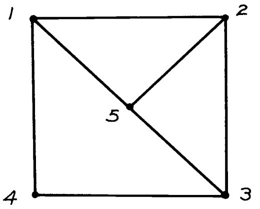


FIG. 1. *"kite" graph.*

Its corresponding adjacency and Kirchhoff matrices are

$$(1)\qquad \mathbf{A}=\begin{pmatrix} 0 & 1 & 0 & 1 & 1\\ 1 & 0 & 1 & 0 & 1\\ 0 & 1 & 0 & 1 & 1\\ 1 & 0 & 1 & 0 & 0\\ 1 & 1 & 1 & 0 & 0 \end{pmatrix},\quad \mathbf{K}=\begin{pmatrix} 3 & -1 & 0 & -1 & -1\\ -1 & 3 & -1 & 0 & -1\\ 0 & -1 & 3 & -1 & -1\\ -1 & 0 & -1 & 2 & 0\\ -1 & -1 & -1 & 0 & 3 \end{pmatrix}.$$

These types of matrices will be the starting point of m-forest enumerations. For the enumeration of the spanning subgraphs with a specific cyclomatic number, of a given *planar* graph G, we introduce the so-called *cycle-adjacency matrix* (this is a terminology we coined [2]), in the following way. First we note that, for a given graph, the notion of cyclomatic number simply means the maximum number of "independent" cycles (i.e. no cycle is *covered* by the union of the other cycles). Let N be the cyclomatic number of G. Consider a set of independent cycles, S, in G over the field $\mathbb{Z}_2$, with

$$S=\{C_1,\cdots,C_r\}$$

where C_i are cycles in S and $r\leqq N$ is called the *degree* of S written $|S|=r$. When S is maximal then r is just the cyclomatic number of G. The *cycle-adjacency matrix* of G *relative* to S, is defined by an $r\times r$ matrix

$$(2)\qquad \mathbf{E}_S=\|E_{ij}\|$$

with $E_{ij}=-\{\text{total number of edges common to } C_i \text{ and } C_j\}$ for $i\neq j$. $E_{ii}=+\{\text{total number of edges belonging to } C_i\}$. It is further required that no edge can belong to more than *two* cycles of S.

As an example, we again take up the "kite" graph (Fig. 1) and indicate a set of independent cycles C_1, C_2, C_3 by dotted lines (Fig. 2).

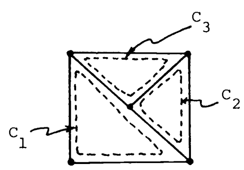


FIG. 2

Let $S = \{C_1, C_2, C_3\}$. Then the cycle-adjacency matrix $\mathbf{E}$ relative to S is

$$\mathbf{E}_S = \begin{pmatrix} 4 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}.$$

If we consider $S' = \{C_1, C_2\}$, then $\mathbf{E}$ relative to S' is

$$\mathbf{E}_{S'} = \begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix}.$$

Or alternatively (see Fig. 3) consider $S'' = \{C_1, C_4\}$. Then

$$\mathbf{E}_{S''} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}.$$

From now on we shall drop the subscript of $\mathbf{E}$ since it is usually clear from the context.

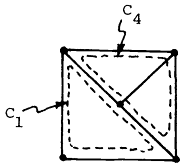


FIG. 3

The following notations will be used for submatrices: for a Kirchhoff matrix $\mathbf{K}$, the ith principal submatrix (i.e. the deletion of ith row and column from K) is written $\mathbf{K}(i)$. Similarly, we denote by $\mathbf{K}(i, j)$ the submatrix obtained from $\mathbf{K}$ by deleting the ith row and column as well as the jth ones. We also introduce the following convention for Kirchhoff matrices:

$$\mathbf{K}(1, \cdots, n) = 1 \tag{3}$$

and

$$\mathbf{K}(i_1, \cdots, i_p) = 0 \quad \text{if } i_j = i_k \text{ for some } j \neq k, \tag{4}$$

where 1 and 0 are just numbers (i.e. 1×1 matrices). Exactly similar notations and conventions can be set up for cycle-adjacency matrices except we have now

$$\mathbf{E}(1, \cdots, N) = 1, \tag{5}$$

and

$$\mathbf{E}(i_1, \cdots, i_p) = 0 \quad \text{if } i_j = i_k \text{ for some } j \neq k. \tag{6}$$

We recall that N is the cyclomatic number of G.

2. Forest enumeration: general exposition and an example. The approach we present here involves, first of all, a formal vector space. Denote $\mathbf{n}^* \equiv \{1, 2, \cdots, n-1\}$. Let $\mathcal{P}(\mathbf{n}^*)$ be the power set of $\mathbf{n}^*$, i.e. the collection of all subsets in $\mathbf{n}^*$. Construct the space W of formal sums generated by $\mathcal{P}(\mathbf{n}^*)$ with coefficients in $\mathbb{R}$. Define a real-valued linear function μ, to be called an *evaluation map*, on the space W by

$$\mu : S \mapsto \det \mathbf{K}(n \cup {}_cS) \tag{7}$$

where $S \subset \mathbf{n}^*$ and ${}_cS$ is the complement of S in $\mathbf{n}^*$. For instance, take the kite graph and label the vertices as in Fig. 1. The Kirchhoff matrix is given by (1). Consider $S\equiv\{1,4\}$. Then under the evaluation map,

$$\mu:\{1,4\}\mapsto\det \mathbf{K}(2,3,5)=\begin{vmatrix}3 & -1\\ -1 & 2\end{vmatrix}=5. \tag{8}$$

For convenience, we shall use the expression that the set S *generates* the number T w.r.t. μ if $T=\mu(S)$. For the example of (8) we say that $\{1,4\}$ generates the number 5. Let $T^{(m)}$ be the total number of spanning m-forest of G. Then, by the celebrated Kirchhoff's matrix-tree theorem [5], we have the fact that $\mathbf{n}^*$ *generates* $T^{(1)}$ since $\mu(\mathbf{n}^*)=T^{(1)}$. It is also trivial to see that $\varnothing$ (the empty set) *generates* 1 since $\mu(\varnothing)=1$.

Our next step is to introduce the following $\mathbb{R}$-linear operator, to be called the "annihilation operator at the ith vertex", on the vector space W:

$$\alpha_i: S\mapsto S-\{i\} \quad \text{if } i\in S, \tag{9}$$

$$S\mapsto 0 \qquad \text{if } i\notin S \tag{10}$$

where $S\in\mathcal{P}(\mathbf{n}^*)$ and the 0 in (10) is the zero formal sum in W. It is important to note that the empty set $\varnothing$ is *not* zero and must be retained in formal sums, e.g., $5\{\varnothing\}+2\{4\}-6\{1,3\}$. It is understood that the above definition of the operators extends $\mathbb{R}$-linearly and also w.r.t. the addition in W. It is obvious by the nature of the definition that the operators α_j and α_i always commute.

The following theorem proved in [1] achieves a generalization of Kirchhoff's matrix-*tree* theorem to the case of any forest (hence we may perhaps call this theorem the matrix-*forest* theorem!).

MATRIX-FOREST THEOREM. *Define the operator*:

$$\alpha\equiv\sum_{i\in\mathbf{n}^*}\alpha_i-\frac{1}{2}\sum_{i,j\in\mathbf{n}^*}a_{ij}\alpha_i\alpha_j \tag{11}$$

and denote

$$\mathbf{T}^{(m)}\equiv\frac{1}{(m-1)!}\alpha^{m-1}\mathbf{n}^*. \tag{12}$$

Then $\mathbf{T}^{(m)}$ *generates* $T^{(m)}$, *the total number of m-forests in the graph.*

Since the theorem was proved in [1] we shall not repeat the details of the mathematical induction. We shall instead use the example of the kite graph to illustrate the application of the theorem. Let the vertices be labelled as in Fig. 1. Then the corresponding adjacency matrix is given by (1). The Kirchhoff matrix is then, by definition,

$$\mathbf{K}=\begin{pmatrix}3 & -1 & 0 & -1 & -1\\ -1 & 3 & -1 & 0 & -1\\ 0 & -1 & 3 & -1 & -1\\ -1 & 0 & -1 & 2 & 0\\ -1 & -1 & -1 & 0 & 3\end{pmatrix}. \tag{13}$$

The set $\mathbf{n}^*$ is in this case simply

$$\mathbf{n}^*=\{1,2,3,4\}. \tag{14}$$

First, we compute the total number of spanning 1-forests, i.e. spanning trees, by the matrix-forest theorem:

$$T^{(1)} = \mu(\mathbf{T}^{(1)}) = \mu(\alpha^0\mathbf{n}^*) = \mu(\mathbf{n}^*) = \det \mathbf{K}(n) = \det \mathbf{K}(5)$$

$$(15) \qquad = \begin{vmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{vmatrix} = 24,$$

i.e. there are 24 spanning trees in the kite graph. They are the graphs shown in Fig. 4.

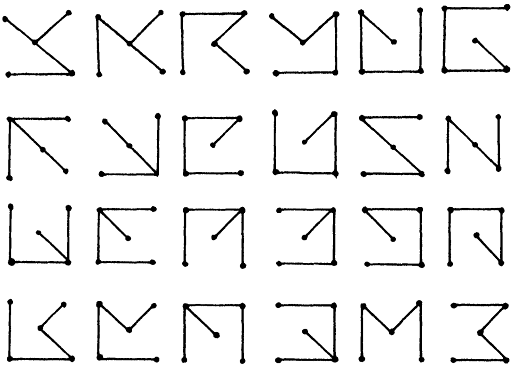


FIG. 4. *Trees (1-forests).*

Next, let us compute the total number of spanning 2-forests:

$$(16) \quad \mathbf{T}^{(2)} = \frac{1}{(2-1)!}\alpha\mathbf{n}^* = \left(\sum_{i=1}^{4} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{4} a_{ij}\alpha_i\alpha_j\right)\mathbf{n}^*$$

$$(17) \qquad = (\alpha_1+\alpha_2+\alpha_3+\alpha_4-\alpha_1\alpha_2-\alpha_1\alpha_4-\alpha_2\alpha_3-\alpha_3\alpha_4)\mathbf{n}^*$$

$$(18) \qquad \begin{aligned} &= \{2,3,4\}+\{1,3,4\}+\{1,2,4\}+\{1,2,3\} \\ &\quad -\{3,4\}-\{2,3\}-\{1,4\}-\{1,2\}. \end{aligned}$$

Then under the evaluation map μ,

$$(19) \qquad \begin{aligned} \mu(\mathbf{T}^{(2)}) &= \mu\{2,3,4\}+\mu\{1,3,4\}+\mu\{1,2,4\}+\mu\{1,2,3\} \\ &\quad -\mu\{3,4\}-\mu\{2,3\}-\mu\{1,4\}-\mu\{1,2\}. \end{aligned}$$

But

$$\mu\{2,3,4\} = \begin{vmatrix} 3 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{vmatrix} = 13, \qquad \mu\{1,3,4\} = \begin{vmatrix} 3 & 0 & -1 \\ 0 & 3 & -1 \\ -1 & -1 & 2 \end{vmatrix} = 12,$$

$$\mu\{1,2,4\} = \begin{vmatrix} 3 & -1 & -1 \\ -1 & 3 & 0 \\ -1 & 0 & 2 \end{vmatrix} = 13, \qquad \mu\{1,2,3\} = \begin{vmatrix} 3 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 3 \end{vmatrix} = 21,$$

$$\mu\{3,4\} = \begin{vmatrix} 3 & -1 \\ -1 & 2 \end{vmatrix} = 5, \qquad \mu\{2,3\} = \begin{vmatrix} 3 & -1 \\ -1 & 3 \end{vmatrix} = 8,$$

$$\mu\{1,4\} = \begin{vmatrix} 3 & -1 \\ -1 & 2 \end{vmatrix} = 5, \qquad \mu\{1,2\} = \begin{vmatrix} 3 & -1 \\ -1 & 3 \end{vmatrix} = 8.$$

Therefore

$$T^{(2)} = \mu(\mathbf{T}^{(2)}) = 13+12+13+21-5-8-5-8=33, \tag{20}$$

i.e. there are 33 2-forests in the kite graph. They are shown in Fig. 5.

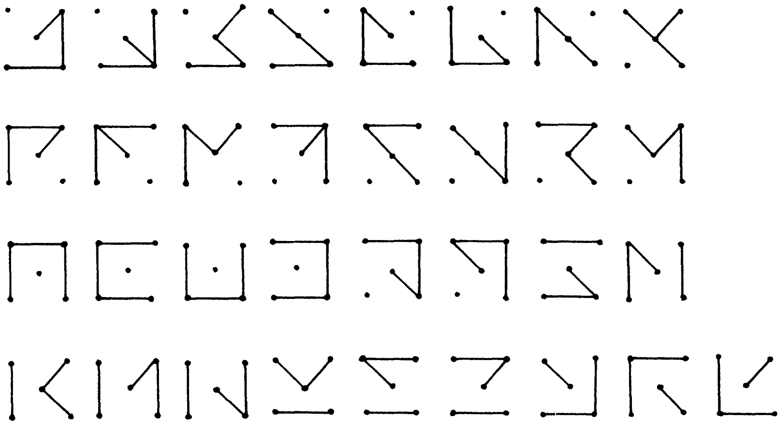


FIG. 5. *2-forests.*

To evaluate the total number of 3-forests, i.e. $T^{(3)}$, we notice that the resulting expression from $T^{(2)}$ can be used to cut down the unnecessary repetition. In other words, the relation, which follows trivially from (12),

$$\mathbf{T}^{(m+1)} = \frac{1}{m}\alpha\mathbf{T}^{(m)} \tag{21}$$

is computationally very useful. By (17) and (18) we have

$$\begin{aligned}\mathbf{T}^{(3)} &= \tfrac{1}{2}(\alpha_1+\alpha_2+\alpha_3+\alpha_4-\alpha_1\alpha_2-\alpha_1\alpha_4-\alpha_2\alpha_3-\alpha_3\alpha_4)\\ &\quad\cdot[\{2,3,4\}+\{1,3,4\}+\{1,2,4\}+\{1,2,3\}-\{3,4\}-\{2,3\}-\{1,4\}-\{1,2\}].\end{aligned} \tag{22}$$

The computations involved in (22) are:

$$\alpha_1[(18)] = \{3,4\}+\{2,4\}+\{2,3\}-\{4\}-\{2\}, \tag{23}$$

$$\alpha_2[(18)] = \{3,4\}+\{1,4\}+\{1,3\}-\{3\}-\{1\}, \tag{24}$$

$$\alpha_3[(18)] = \{2,4\}+\{1,4\}+\{1,2\}-\{4\}-\{2\}, \tag{25}$$

$$\alpha_4[(18)] = \{2,3\}+\{1,3\}+\{1,2\}-\{3\}-\{1\}, \tag{26}$$

$$-\alpha_1\alpha_2[(18)] = -\{4\}-\{3\}+\{\varnothing\}, \tag{27}$$

$$-\alpha_1\alpha_4[(18)] = -\{3\}-\{2\}+\{\varnothing\}, \tag{28}$$

$$-\alpha_2\alpha_3[(18)] = -\{4\}-\{1\}+\{\varnothing\}, \tag{29}$$

$$-\alpha_3\alpha_4[(18)] = -\{2\}-\{1\}+\{\varnothing\}, \tag{30}$$

where the empty sets in (27) to (30) are the result of operations $\alpha_1\alpha_2\{1,2\} = \{\varnothing\}$, etc. It cannot be overemphasized that by definition (9) and (10), we have $\alpha_1\alpha_2\{3,4\} = 0$ not $\varnothing$! Hence

$$\begin{aligned}\mathbf{T}^{(3)} &= \{1,2\}+\{1,3\}+\{1,4\}+\{2,3\}+\{2,4\}+\{3,4\}\\ &\quad-2\{1\}-2\{2\}-2\{3\}-2\{4\}+2\{\varnothing\}.\end{aligned} \tag{31}$$

Therefore, by $T^{(3)}=\mu(\mathbf{T}^{(3)})$, we have

$$\begin{aligned}T^{(3)}=&\begin{vmatrix}3&-1\\-1&3\end{vmatrix}+\begin{vmatrix}3&0\\0&3\end{vmatrix}+\begin{vmatrix}3&-1\\-1&2\end{vmatrix}+\begin{vmatrix}3&-1\\-1&3\end{vmatrix}\\&+\begin{vmatrix}3&0\\0&2\end{vmatrix}+\begin{vmatrix}3&-1\\-1&2\end{vmatrix}-6-6-6-4+2=21\end{aligned}\tag{32}$$

i.e. there are 21 3-forests in the kite graph. They are drawn in Fig. 6.

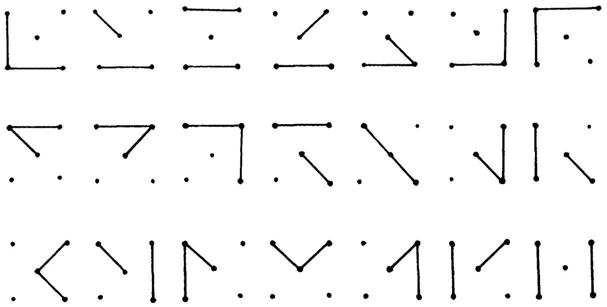


FIG. 6. *3-forests.*

To compute the total number of 4-forests $T^{(4)}$ we now use (21) for $m=3$, i.e.

$$\mathbf{T}^{(4)}=\tfrac{1}{3}\alpha\mathbf{T}^{(3)}.\tag{33}$$

From (22), we can write down by inspection

$$\begin{aligned}\alpha_1\mathbf{T}^{(3)}&=\{2\}+\{3\}+\{4\}-2\{\varnothing\},\\\alpha_2\mathbf{T}^{(3)}&=\{1\}+\{3\}+\{4\}-2\{\varnothing\},\\\alpha_3\mathbf{T}^{(3)}&=\{1\}+\{2\}+\{4\}-2\{\varnothing\},\\\alpha_4\mathbf{T}^{(3)}&=\{1\}+\{2\}+\{3\}-2\{\varnothing\},\end{aligned}\tag{34}$$

and

$$\begin{aligned}-\alpha_1\alpha_2\mathbf{T}^{(3)}&=-\{\varnothing\},&\quad-\alpha_1\alpha_4\mathbf{T}^{(3)}&=-\{\varnothing\},\\-\alpha_2\alpha_3\mathbf{T}^{(3)}&=-\{\varnothing\},&\quad-\alpha_3\alpha_4\mathbf{T}^{(3)}&=-\{\varnothing\}.\end{aligned}\tag{35}$$

Therefore,

$$\mathbf{T}^{(4)}=\{1\}+\{2\}+\{3\}+\{4\}-4\{\varnothing\},\tag{36}$$

$$T^{(4)}=\mu\mathbf{T}^{(4)}=3+3+3+2-4=7,\tag{37}$$

i.e. there are seven 4-forests in the kite graph. They are given in Fig. 7.

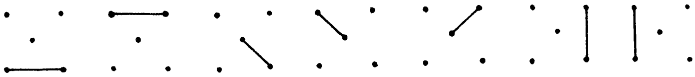


FIG. 7. *The 4-forests.*

It is trivial that there is only one 5-forest since $n=5$; it needs no calculation. But this can also be concluded formally from our computation. Take (36) and use (21) for $m=4$, i.e.

$$\mathbf{T}^{(5)}=\tfrac{1}{4}\alpha\mathbf{T}^{(4)}. \tag{38}$$

We get

$$\mathbf{T}^{(5)}=\tfrac{1}{4}[\varnothing+\varnothing+\varnothing+\varnothing-0]=\{\varnothing\}. \tag{39}$$

Hence

$$T^{(5)}=\mu\mathbf{T}^{(5)}=\mu(\varnothing)=1 \tag{40}$$

which is expected trivially.

3. Forest enumeration: complete graphs, complete bipartite graphs and complete tripartite graphs. We shall take up the complete graphs, complete bipartite graphs and complete tripartite graphs. On the historical account, Cayley [6] first solved the $T^{(1)}$ case of complete graphs in 1889 and, in 1959, Renyi published [7] the solutions of $T^{(m)}$ for any positive integer m. As to the bipartite complete graphs, the solution first appeared in the monograph of John W. Moon [8]. We shall give alternative operator-approach derivations of these classic results but in *simpler* representations. It is important to note that the derivations given here do not require any detailed *combinatorial* argument.

For a complete graph, all the off-diagonal entries in the adjacency matrix are equal to one. This reduces the operator α to

$$\alpha=\sum_{i=1}^{n-1}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n-1}\alpha_i\alpha_j=\left(\sum_{i=1}^{n-1}\alpha_i\right)\left(1-\frac{1}{2}\sum_{j=1}^{n-1}\alpha_j\right). \tag{41}$$

Hence

$$\alpha^{m-1}=\left(\sum_{i=1}^{n-1}\alpha_i\right)^{m-1}\left(1-\frac{1}{2}\sum_{j=1}^{n-1}\alpha_j\right)^{m-1}=\sum_{r=0}^{m-1}\binom{m-1}{r}\left(\frac{-1}{2}\right)^r\left(\sum_{j=1}^{n-1}\alpha_j\right)^{m+r-1}. \tag{42}$$

Then by the matrix-forest theorem (12), we have

$$\mathbf{T}^{(m)}=\sum_{r=0}^{\bar{m}}\frac{1}{r!\,(m-r-1)!}(-1/2)^r\sum_{i_1,\cdots,i_{m+r-1}=1}^{n-1}(\alpha_{i_1}\cdots\alpha_{i_{m+r-1}})\mathbf{n}^* \tag{43}$$

where $\bar{m}$ is the minima of $m-1$ and $n-m$. The prime over the second summation of (43) indicates that $i_1,\cdots,i_{m+r-1}$ are all *distinct* due to the fact that $\alpha_i^2=0$ and $\alpha_i\alpha_j=\alpha_j\alpha_i$. (43) gives immediately

$$\begin{aligned}T^{(m)}&=\mu(\mathbf{T}^{(m)})\\&=\sum\frac{1}{r!\,(m-r-1)!}(-1/2)^r\sum{}'\det\mathbf{K}(i_1,\cdots,i_{m+r-1},n).\end{aligned} \tag{44}$$

It is easy to find that

$$\det\mathbf{K}(i_1,\cdots,i_{m+r-1},n)=n^{n-m-r-1}(m+r). \tag{45}$$

Thus (44) becomes

$$T^{(m)}=(n-1)!\,n^{n-m-1}\sum_{r=0}^{\bar{m}}(-2n)^{-r}(m+r)/\{r!\,(m-r-1)!\,(n-m-r)!\}. \tag{46}$$

Due to the increasing complexity, we list below only the first five cases: $m=1$, 2, 3, 4, 5.

$$\begin{aligned}
T^{(1)} &= n^{n-2},\\
T^{(2)} &= \frac{1}{2}\frac{(n-1)!}{(n-2)!}n^{n-4}(n+6),\\
T^{(3)} &= \frac{1}{2^2(2!)}\frac{(n-1)!}{(n-3)!}n^{n-6}(n^2+13n+60),\\
T^{(4)} &= \frac{1}{2^3(3!)}\frac{(n-1)!}{(n-4)!}n^{n-8}(n^3+21n^2+202n+840),\\
T^{(5)} &= \frac{1}{2^4(4!)}\frac{(n-1)!}{(n-5)!}n^{n-10}(n^4+30n^3+451n^2+3846n+15120).
\end{aligned} \tag{47}$$

A *bipartite* graph G is characterized by the division of $V(G)$ into two subsets such that no two vertices are adjacent in the same subset. Thus for a *complete* bipartite graph, its adjacency matrix is fixed by the graph's 2-partition of vertices, say, $n = p_1 + p_2$. We have

$$a_{ij} = \begin{cases} 1, & \text{if } i \leqq p_1 < j \text{ or } j \leqq p_1 < i, \\ 0, & \text{otherwise.} \end{cases} \tag{48}$$

To compute $T^{(m)}$ we have to evaluate α^{m-1}. From $\alpha_i^2=0$ and $\alpha_i\alpha_j=\alpha_j\alpha_i$ it follows

$$\alpha^{m-1} = \left[\left(\sum_i \alpha_i + \sum_j \alpha_j\right) - \sum_i\sum_j \alpha_i\alpha_j\right]^{m-1}$$

where i runs from 1 to p_1 and j runs from p_1+1 to $n-1$. Thus

$$\begin{aligned}
\alpha^{m-1} &= \sum_{r=0}^{m-1} (-1)^r \binom{m-1}{r}\left(\sum_i \alpha_i + \sum_j \alpha_j\right)^{m-r-1}\left(\sum_i \alpha_i\right)^r\left(\sum_j \alpha_j\right)^r\\
&= \sum_{r=0}^{m-1} \sum_{q_1+q_2=m-r-1} (-1)^r \binom{m-1}{r}\binom{m-r-1}{q_1, q_2}\left(\sum_i \alpha_i\right)^{q_1+r}\left(\sum_j \alpha_j\right)^{q_2+r}.
\end{aligned} \tag{49}$$

Then

$$\begin{aligned}
T^{(m)} &= \frac{1}{(m-1)!}\mu\{\alpha^{m-1}\mathbf{n}^*\}\\
&= \sum_{r=0}^{\overline{m-1}} \sum_{q_1+q_2=m-r-1} (-1)^r \frac{1}{q_1!\,q_2!\,r!}\sum_I\sum_J \det \mathbf{K}(I, J, n)
\end{aligned} \tag{50}$$

where each of the sets $I \equiv (i_1, \cdots, i_{m-1-k})$ and $J \equiv (j_1, \cdots, j_{k+r})$ has distinct elements (i.e. $i_1 \neq i_2 \neq \cdots$, etc.); the ranges of summation over I and J are, respectively, from 1 to p_1 and p_1+1 to $n-1$.

A calculation similar to that of complete graphs yields the following combinatorial formula:

$$\begin{aligned}
T^{(m)} = p_1^{p_2-1}p_2^{p_1-1} \sum_{r=0}^{\overline{m-1}} (-1)^r \frac{(p_1)_r(p_2-1)_r}{r!\,p_1^r p_2^r} &\sum_{q_1+q_2=m-r-1} \binom{p_1-r}{q_1}\binom{p_2-r-1}{q_1}\\
&\times p_1^{-q_2-1}p_2^{-q_1}[p_1p_2-(p_1-q_1-r)(p_2-q_2-r-1)].
\end{aligned} \tag{51}$$

As an example, we find immediately

$$T^{(1)}=p_1^{p_2-1}p_2^{p_1-1}$$

and

$$T^{(2)}=p_1^{p_2-2}p_2^{p_1-2}(p_1^2+p_2^2+p_1+p_2-p_1p_2-2).$$

For a complete tripartite graph, its adjacency matrix A is fixed by the graph's 3-partition of vertices, say, $n=p_1+p_2+p_3$ with

$$a_{ij}=\begin{cases}1 & \text{if } i\leqq p_1<j,\quad j\leqq p_1<i,\quad p_1<i\leqq p_1+p_2<j \text{ or } p_1<j\leqq p_1+p_2<i,\\ 0 & \text{otherwise.}\end{cases}$$

A multinomial expansion similar to that of complete bipartite graphs yields

$$T^{(m)}=\sum_{r=0}^{\overline{m-1}}\sum_{Q=m-r-1}\sum_{R=r}(-1)^r\frac{1}{q_1!\,q_2!\,q_3!\,r_1!\,r_2!\,r_3!}\sum_I\sum_J\sum_L \det\mathbf{K}(I,J,L,n) \tag{52}$$

where

$$Q=q_1+q_2+q_3,\qquad R=r_1+r_2+r_3,$$

$$I=(i_1,\cdots,i_{q_1+r-r_1}),\quad J=(j_1,\cdots,j_{q_2+r-r_2}),\quad L=(l_1,\cdots,l_{q_3+r-r_2})$$

and the range of summation over I, J and L are, respectively, from 1 to p_1, p_1+1 to p_1+p_2 and p_1+p_2+1 to $n-1$. It is readily seen that

$$\sum_I\sum_J\sum_L\det\mathbf{K}(I,J,L,n)=(p_1)_{q_1+r-r_1}(p_2)_{q_2+r-r_2}(p_{3-1})_{q_3+r-r_3}\det\mathbf{R}$$

where

$$\mathbf{R}=\left(\begin{array}{c:c:c}(p_2+p_3)\mathbf{1} & & \mathbf{1}^* \\ \hdashline & (p_1+p_3)\mathbf{1} & \\ \hdashline -\mathbf{1}^* & & (p_1+p_2)\mathbf{1}\end{array}\right)\begin{array}{l}\}b_1\\ \}b_2\\ \}b_3\end{array} \tag{53}$$

with

$$b_1=p_1-q_1-r+r_1,\quad b_2=p_2-q_2-r+r_2,\quad b_3=p_3-q_3-r+r_3-1$$

and $-\mathbf{1}^*$ are matrices with -1 for *every* entry (not just for the diagonals). A detailed calculation leads to

$$\det\mathbf{R}=a_1^{b_1-1}a_2^{b_2-1}a_3^{b_3-1}[a_1a_2a_3-a_1b_2b_3-a_2b_3b_1-a_3b_1b_2-2b_1b_2b_3] \tag{54}$$

where $a_1=p_2+p_3$, $a_2=p_3+p_1$, $a_3=p_1+p_2$.

When (53) and (54) are substituted into (52) we get

$$\begin{aligned}T^{(m)}=\sum_{r=0}^{\overline{m-1}}\sum_{Q-m-r-1}\sum_{R=r}(-1)^r\frac{1}{q_1!\,q_2!\,q_3!\,r_1!\,r_2!\,r_3!}(p_1)_{p_1-b_1}(p_2)_{p_2-b_2}(p_3-1)_{p_3-b_3-1}\\ \times a_1^{b_1-1}a_2^{b_2-1}a_3^{b_3-1}[a_1a_2a_3-a_1b_2b_3-a_2b_3b_1-a_3b_1b_2-2b_1b_2b_3].\end{aligned} \tag{55}$$

For $m=1, 2$ we find

$$T^{(1)}=(p_1+p_2+p_3)(p_2+p_3)^{p_1-1}(p_3+p_1)^{p_2-1}(p_1+p_2)^{p_3-1} \tag{56}$$

and

$$\begin{aligned}
T^{(2)} = {} & (p_1+p_2)^{p_3-3}(p_1+p_3)^{p_2-2}(p_2+p_3)^{p_1-2} \\
& \times\{p_1(p_1+p_2)(p_1+p_3)[(p_1+p_2+p_3)^2+(p_1+p_2+p_3)(p_2-1)-p_2] \\
& +(p_3-1)(p_1+p_2)(p_2+p_3)(p_3+p_1)(p_1+p_2+p_3) \\
& -p_1(p_1+p_3)(p_3-1)[(p_1+p_2+p_3)^2+(p_1+p_2+p_3)(p_1+2p_2-2)-2p_2] \\
& \qquad -p_1p_2(p_1+p_2)(p_1+p_2+p_3-1)^2\} \\
& +(\text{the same terms with } p_1 \text{ and } p_2 \text{ exchanged}).
\end{aligned} \tag{57}$$

4. Enumeration of spanning subgraphs having a fixed cyclomatic number (planar graphs only). Our consideration is restricted to planar graphs. For convenience we shall assume that the given graphs are connected. For a given graph, the cyclomatic number simply means the number of "independent" cycles in the graph. It is therefore reasonable to expect the use of "cycle" may be more effective for this problem. It is, however, not obvious to us how such an approach can be pursued. Rather, we find the notion of "duality," i.e. to each planar graph there is a dual graph, most helpful in formulating our approach: since the m-forest problem is now completely solved (though without planary restriction!) by our operator technique one may *expect* that *duality* may carry the burden of solution, for the *planar* graphs, of the above-mentioned enumeration problems. This turns out to be correct though not in a trivial manner. The key idea is the use of the so-called *cycle-adjacency matrix* (this is not a standard terminology as we mentioned in the definition in § 1 here) in the framework of our operator approach. The rationale is the following: in solving the m-forest problem the Kirchhoff matrix is used though the rows and columns are labelled according to the *vertices.* Duality carries "vertices" to "cycles" (or "faces") for planar graphs. So we may devise a Kirchhoff-like matrix but somehow to be labelled by the *cycles* instead of vertices. This is the way how the term *cycle-adjacency matrix* was coined (clearly, it should perhaps be better called cycle-Kirchhoff matrix but a Kirchhoff matrix is essentially an adjacency matrix).

In what follows, we shall formulate our operator method for enumeration of spanning subgraphs, with a fixed cyclomatic number, in an alternative setting than the one that is used for the forest enumeration. However, the two different formalisms are actually equivalent in the sense that the formalism to be discussed below can be cast into a form that bears direct resemblance to that of § 2 with the correspondence of cycles with vertices and vice versa. Our motivation to present a different setting below is the visualization of graphic meaning in different stages of computation.

Following the notation we set in § 1, we shall first of all construct a real vector space U of formal sums. Let $\mathbf{M}_i$ be real square matrices not necessarily of the same dimension, $i=1,\cdots,r$ for some finite integer r. Then U is the real vector space consisting of formal sums

$$a_1 * \mathbf{M}_1+\cdots+a_r * \mathbf{M}_r$$

where a_i are real numbers and the symbol $*$ now *prohibits* the usual multiplication of the number a_i into the matrix $\mathbf{M}_i$, i.e. a_i are only the "coefficients" of a formal sum. The symbol $*$ defines the following rules of *addition* and *scalar* multiplication to make U a real vector space:

$$\begin{aligned}
&a'(a*\mathbf{M})=(a'a)*\mathbf{M}=(aa')*\mathbf{M}, \\
&(a+a')*\mathbf{M}=a*\mathbf{M}+a'*\mathbf{M}, \\
&a''(a*\mathbf{M}+a'*\mathbf{M}')=(a''a)*\mathbf{M}+(a''a')*\mathbf{M}',
\end{aligned}$$

where a, a', a'' are real numbers and $\mathbf{M}$, $\mathbf{M}'$ are square matrices of generally different dimensions. It cannot be overemphasized that an operation like $a*(a'\mathbf{M})=(aa')*\mathbf{M}$ is *illegal* though either side is well defined. So is $a*(\mathbf{M}+\mathbf{M}')=a*\mathbf{M}+a*\mathbf{M}'$ prohibited here. To simplify notation we shall write hereafter

$$\mathbf{M}^* \equiv 1*\mathbf{M}. \tag{58}$$

Furthermore we shall continue to use the notation $\mathbf{M}(i)$ to denote the matrix obtained from $\mathbf{M}$ by deleting the ith row and column with $\mathbf{M}(i,j)$ to be understood as successive deletions referred to the *original* labelling of entries in $\mathbf{M}$; $\mathbf{M}(i,j)$ yields the *zero* matrix if $i=j$, i.e. if the same row and column are deleted more than once.

The ith *cycle-annihilation* operator γ_i on the vector space U is a linear operator defined by

$$\gamma_i(a*\mathbf{M})=a*\mathbf{M}(i), \tag{59}$$

for $a\in\mathbb{R}$ and a square matrix $\mathbf{M}$. Formally γ_i is similar to α_i of (9) and (10); the former is defined on the space U and the latter on W. Despite the similarity, there is an important difference in their graphic implication: in the case of α_i, under the consideration of $\mathbf{n}^*$, the operator *identifies* the ith vertex with the nth vertex in the original graph (regardless of whether ith and nth vertices are adjacent or not) while γ_i simply "ignores" the ith cycle in the computation without any modification of the original graph (i.e. the ith cycle does not enter into consideration any more after the operation). The evaluation map μ needed here is

$$\mu: U\to\mathbb{R} \tag{60}$$

with

$$\mu(a*\mathbf{M})=a\det\mathbf{M}. \tag{61}$$

The operator γ introduced below is in analogy to the operator α of (11) though they operate on different spaces:

$$\gamma\equiv\sum_{i\in\mathbf{N}}\gamma_i+\frac{1}{2}\sum_{i,j\in\mathbf{N}}E_{ij}\gamma_i\gamma_j \tag{62}$$

where E_{ij} are the entries of the *cycle-adjacency* matrix defined earlier in the first section. $\mathbf{N}\equiv\{1,2,\cdots,N\}$, with N being the cyclomatic number of the given graph G. Let S be a set of N independent cycles of G. Denote by σ_j the number of ways of deleting $(N-j)$ edges, each from a different cycle in S, such that the resulting subgraph remains connected. Obviously $\sigma_0=T^{(1)}$. Similar to the terminology used for forests, we shall say that $\boldsymbol{\sigma}_i$ generates σ_i to mean $\sigma_i=\mu(\boldsymbol{\sigma}_i)$.

The following theorem (for proof see [2]) is essential to the enumeration problem considered.

THEOREM. *Denote*

$$\boldsymbol{\sigma}_m\equiv\frac{1}{m!}\gamma^m(1*\mathbf{E}). \tag{63}$$

Then $\boldsymbol{\sigma}_m$ generates σ_m.

By means of (61) and (62) one can now obtain from (63) an expression for σ_m involving multiple summations and entries of the cycle-adjacency matrix [2]:

$$\sigma_m=\sum_{r=0}^{m}\frac{1}{r!}\sum_{i_1<i_2}\cdots\sum_{i_{2r-1}<i_{2r}}\ \sum_{i_{2r+1}<\cdots<i_{m+r}}E_{i_1i_2}\cdots E_{i_{2r-1}i_{2r}}\det\mathbf{E}(i_1,\cdots,i_{m+r}). \tag{64}$$

Though this expression no longer involves any operator and gives an answer to the question, yet computationally it is more efficient to use formula (131) of § 6 for efficient recurrent reduction.

To show the graphic implications of (61), (62), and (63) we again consider the example of "kite" graph of Fig. 1 with S chosen as in Fig. 2. The corresponding cycle-adjacency matrix is

$$\mathbf{E}=\begin{pmatrix} 4 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}. \tag{65}$$

We have

$$\sigma_0=\det\mathbf{E}=24 \qquad \text{(spanning trees)}. \tag{66}$$

To find the total number σ_1 of spanning subgraphs each containing *one* cycle first write down the general expression for σ_1 from (64):

$$\sigma_1=\sum_{i=1}^{N}\det\mathbf{E}(i)+\sum_{i<j=1}^{N}E_{ij}\det\mathbf{E}(i,j). \tag{67}$$

For the magic-kite graph,

$$\begin{gathered}\mathbf{E}(1)=\begin{pmatrix} 3 & -1 \\ -1 & 3\end{pmatrix},\quad \mathbf{E}(2)=\begin{pmatrix} 4 & -1 \\ -1 & 3\end{pmatrix},\quad \mathbf{E}(3)=\begin{pmatrix} 4 & -1 \\ -1 & 3\end{pmatrix},\\ \mathbf{E}(1,2)=3,\quad \mathbf{E}(1,3)=3,\quad \mathbf{E}(2,3)=4,\\ E_{12}=E_{13}=E_{23}=-1.\end{gathered} \tag{68}$$

Hence

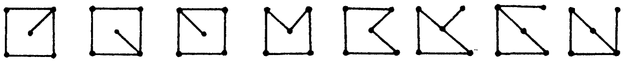
$$\sigma_1=(8+11+11)-(3+3+4) \tag{69}$$

i.e.

$$\sigma_1=20 \qquad \text{(i.e. there are 20 such subgraphs)}. \tag{70}$$

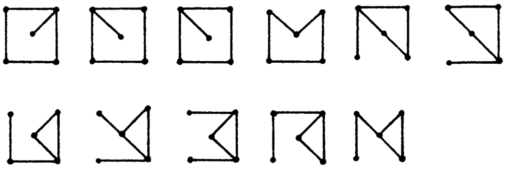
To see what happens graphically we analyze the expression (69) for σ_1. Let us consider the first bracket of (69).Its first term is 8, corresponding to the following subgraphs:

(71)



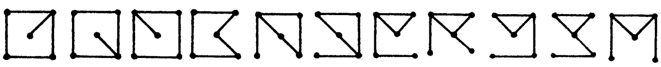
The second term is 11, corresponding to

(72)



The third term is again 11, corresponding to

(73)



So the 30 graphs of the first bracket in (69) are given by (71), (72) and (73). One immediately realizes that there are repetitions among these graphs and they must be removed:

(i) Compare (71) with (72). Remove the following 3 graphs: first, second, fourth from (72). Note that there is still a repetition left, i.e. the third graph in (72) but this is a job for the next step.

(ii) Compare (72) with (73). Remove the following 4 graphs: second, third, fifth, sixth from (73).

(iii) Compare (71) with (73). Remove the following 3 graphs: first, third, fourth. Therefore 10 graphs are removed from (71), (72), (73) so there is no repetition among them. This is the graphic meaning of the second bracket $-(3+3+4)$ in expression (69).

We next compute σ_2. The general expression for σ_2 is

$$\sigma_2 = \sum_{\substack{i,j \\ (i<j)}} \det \mathbf{E}(i,j) + \sum_{\substack{i,j,k \\ (i<j)}} E_{ij} \det \mathbf{E}(i,j,k) + \frac{1}{2} \sum_{\substack{i,j,k,l \\ (i<j;k<l)}} E_{ij}E_{kl} \det \mathbf{E}(i,j,k,l). \tag{74}$$

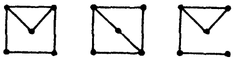
For the example of the kite graph, we have

$$\begin{aligned} \sigma_2 &= \det \mathbf{E}(1,2) + \det \mathbf{E}(2,3) + \det \mathbf{E}(3,1) \\ &\quad + E_{12} \det \mathbf{E}(1,2,3) + E_{23} \det \mathbf{E}(2,3,1) + E_{31} \det \mathbf{E}(3,1,2) \\ &= (3+4+3)-(1+1+1), \end{aligned} \tag{75}$$

i.e. there are 7 spanning subgraphs with two loops. The graphic meaning of the last line of (75) is:

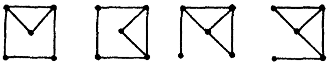
(a) The first term, 3, corresponds to the graphs

(76)



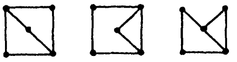
(b) The second term, 4, corresponds to

(77)



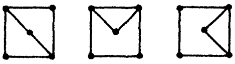
(c) The third term, 3, corresponds to

(78)



To remove the repetition among (76), (77) and (78) we should throw away the following 3 graphs from the graphs:

(79)



corresponding to the last bracket of (75).

σ_3 for the kite graph is just itself. This completes the analysis for this case.

5. Alternative formalisms in the operator approach. In this section, we give a different formalism than the one discussed in § 2 to show that the operator approach considered here is quite flexible and capable of some generalization. In fact, we show that "creation" operators for vertices and edges can be defined as well as the "annihilation" operators we discussed earlier.

Denote

$$\mathbf{n} \equiv \{1, 2, \cdots, n\}. \tag{80}$$

The real vector space X of formal sum is now defined as follows: let $\mathcal{P}(\mathbf{n}\times\mathbf{n})$ be the power set of the Cartesian product $\mathbf{n}\times\mathbf{n}$, then X is the vector space of formal sums generated by $\mathcal{P}(\mathbf{n}\times\mathbf{n})$ with coefficients in $\mathbb{R}$. We shall write a typical element as

$$\sum_{S,S'\subset\mathbf{n}} r_{SS'}\cdot(S; S'), \qquad r_{SS'}\in\mathbb{R}. \tag{81}$$

On X, the ith *vertex-annihilation operator* v_i is a linear operator defined by

$$\begin{aligned} v_i : (S; S') &\mapsto (S; S'-\{i\}) \quad \text{if } i\in S', \\ &\mapsto 0 \quad \text{otherwise.} \end{aligned} \tag{82}$$

Graphically, $\mathbf{n}$ is identified with the set $V(G)$ under a fixed labelling of vertices in G. Similarly, on X, the ijth *edge-annihilation operator* e_{ij} is a linear operator defined by

$$\begin{aligned} e_{ij} : (S; S') &\mapsto (S-\{i, j\}; S') \quad \text{if } \{i, j\}\subset S, \\ &\mapsto 0 \quad \text{otherwise.} \end{aligned} \tag{83}$$

Denote further

$$v \equiv \sum_{i=1}^{n} v_i, \quad e \equiv \sum_{\substack{i,j=1\\(i<j)}}^{n} a_{ij}e_{ij} \quad \text{and} \quad \chi \equiv v-e. \tag{84}$$

Obviously v and e have, respectively, the meaning of total vertex and edge annihilation operators. χ is the "Euler" *annihilation operator*, corresponding to the nonoperator relation for a linear graph:

$$\text{Euler characteristic} = |V(G)| - |E(G)|. \tag{85}$$

For first enumeration problems we define an *evaluation map* ω as a real-valued linear function on the vector space X(corresponding to μ in (7) in a different formalism):

$$\begin{aligned} \omega : (S; S') &\mapsto \det K({}_c\{S\cap S'\}) \quad \text{if } S\cup S' = \mathbf{n}, \\ &\mapsto 0 \quad \text{otherwise.} \end{aligned} \tag{86}$$

We shall again say that F *generates* (w.r.t. ω) the number T if $\omega(F) = T$. Corresponding to (12), we have the following theorem (reference [3]):

THEOREM I. *The formal sum* $(1/m!)\chi^m(\mathbf{n}; \mathbf{n})$ *generates* (w.r.t. ω) *the number of m-forests in G.*

Before we go on with any further exploration, let us again use the kite graph example to illustrate the computations involved. We shall compute first $T^{(1)}$, the number of trees, by this theorem. By putting $m = 1$, we have

$$\chi(\mathbf{n}; \mathbf{n}) = (v-e)(\mathbf{n}; \mathbf{n}) = \left(\sum_{i=1}^{5} v_i - (e_{12}+e_{14}+e_{15}+e_{23}+e_{25}+e_{34}+e_{35})\right)(\mathbf{n}; \mathbf{n}). \tag{87}$$

To simplify notation we shall write

$$(\mathbf{n}; S) \equiv \{S), (S; \mathbf{n}) \equiv (S\}. \tag{88}$$

Then

$$\begin{aligned}\chi(\mathbf{n};\mathbf{n}) &= [\{2345)+\{1345)+\{1245)+\{1235)+\{1234)]\\ &\quad -[(345\}+(235\}+(234\}+(145\}+(134\}+(125\}+(124\}].\end{aligned}\tag{89}$$

Under the evaluation map ω,

$$T^{(1)}=\omega[\chi(\mathbf{n};\mathbf{n})].\tag{90}$$

Since the determinants of submatrices of order $(n-1)$ of a Kirchhoff matrix are equal we see that each term in the first bracket of (89) is equal to:

$$\omega\{2345)=\det \mathbf{K}(1)=24.$$

So the first bracket is equal to 120. Next, we find:

$$\begin{aligned}&\omega(345\}=\det \mathbf{K}(1,2)=13, && \omega(235\}=\det \mathbf{K}(1,4)=16,\\ &\omega(234\}=\det \mathbf{K}(1,5)=13, && \omega(145\}=\det \mathbf{K}(2,3)=13,\\ &\omega(134\}=\det \mathbf{K}(2,5)=12, && \omega(125\}=\det \mathbf{K}(3,4)=16,\\ &\omega(124\}=\det \mathbf{K}(3,5)=13\end{aligned}\tag{91}$$

which gives a value 96 to the second bracket of (89). Thus

$$T^{(1)}=\omega[\chi(\mathbf{n},\mathbf{n})]=120-96=24 \qquad \text{(trees)}.\tag{92}$$

The reader will obviously question the wisdom of such a computational process since by Kirchhoff's method it is simply the single term $\det \mathbf{K}(1)=24$. The point is that we want to pave a way to show the possibility of introducing the so-called "Euler" operators. Such a generalization shows the beautiful further symmetry of the approach [3].

We next compute χ^2:

$$\chi^2(\mathbf{n};\mathbf{n})=\chi(\chi(\mathbf{n};\mathbf{n})).\tag{93}$$

Then typical terms are, for instance,

$$\begin{aligned}&v_1\{2345)=0, \qquad v_1\{1345)=\{345), \text{etc.},\\ &e_{12}\{2345)=(345;2345), \quad \text{etc.},\\ &v_1(345\}=(345;2345), \quad \text{etc.},\\ &e_{12}(234\}=0, \qquad e_{12}(125\}=(5\}, \quad \text{etc.}\end{aligned}\tag{94}$$

The result under the evaluation map ω is, after some straightforward computation,

$$T^{(2)}=33 \qquad \text{(2-forests in the "kite" graph)}.\tag{95}$$

Let us now define the ith *vertex creation* operator $v_i^\dagger$:

$$\begin{aligned}v_i^\dagger:(S;S')&\mapsto(S;S'\cup\{i\}) \quad \text{if } i\notin S',\\ &\mapsto 0 \quad \text{otherwise.}\end{aligned}\tag{96}$$

Similarly we define the ijth *edge creation* operator

$$\begin{aligned}e_{ij}^\dagger:(S;S')&\mapsto(S\cup\{i,j\};S') \quad \text{if } \{i,j\}\subset {}_cS,\\ &\mapsto 0 \quad \text{otherwise.}\end{aligned}\tag{97}$$

For forest enumeration we define the evaluation map as a real-valued function $\omega^\dagger$ on X by

$$\begin{aligned} \omega^\dagger : (S; S') &\mapsto \det \mathbf{K}(S \cup {}_c S') \quad \text{if } S \cap S' = \varnothing, \\ &\mapsto 0 \quad \text{otherwise.} \end{aligned} \tag{98}$$

Then we have the following theorem [3]:

THEOREM II. *The formal sum*

$$\frac{1}{m!}\chi^{\dagger m}(\varnothing; \varnothing) \tag{99}$$

generates ($w.r.t.$ $\omega^\dagger$) *the number of* $(n-m)$*-forests in* G.

A comparison between Theorems I and II shows the amazing symmetry, or rather, "duality," between sets, operators and evaluation maps:

$$\varnothing \text{ versus } n, \quad \chi^\dagger \text{ versus } \chi, \quad \omega^\dagger \text{ versus } \omega \quad \text{and} \quad (n-m) \text{ versus } m. \tag{100}$$

We shall illustrate the typical computations by the kite graph. For $m = 1$, Theorem II gives $T^{(4)}$, i.e. the number of 4-forests in G. The computation would be tedious by Theorem I yet it is very simple by Theorem II.

$$\chi^\dagger(\varnothing; \varnothing) = (v^\dagger - e^\dagger)(\varnothing; \varnothing) = \left(\sum_{i=1}^{N} v_i^\dagger - \sum_{i<j} a_{ij} e_{ij}^\dagger\right)(\varnothing; \varnothing). \tag{101}$$

To simplify notation we write

$$(i, \cdots] \equiv (i, \cdots; \varnothing) \quad \text{and} \quad [i, \cdots) \equiv (\varnothing; i, \cdots). \tag{102}$$

Then

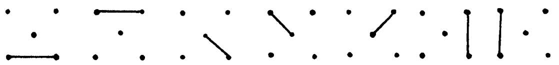
$$\chi^\dagger(\varnothing; \varnothing) = \sum_{i=1}^{5} [i) - (1,2] - (1,4] - (1,5] - (2,3] - (2,5] - (3,4] - (3,5] \tag{103}$$

or

$$T^{(4)} = \omega^\dagger\{\chi^\dagger(\varnothing; \varnothing)\} \tag{104}$$

$$= 3+3+3+2+3-1-1-1-1-1-1-1 = 7 \tag{105}$$

corresponding to the total *seven* 4-forests in the kite graph:

(106) 

We carry out here one more computation, $T^{(3)}$, by Theorem II:

$$(\chi^\dagger)^2(\varnothing; \varnothing) = \chi^\dagger(\text{RHS of (103)}). \tag{107}$$

First,

$$\begin{aligned} v_1^\dagger(\text{RHS of (103)}) = [1,2) + [1,3) + [1,4) + [1,5) - (1; 2, 3) \\ - (1; 2, 5) - (1; 3, 4) - (1; 3, 5) \end{aligned} \tag{108}$$

which leads to under the evaluation map $\omega^\dagger$ the contribution:

$$8+9+5+8-3-3-3-3 = 30-12 = 18. \tag{109}$$

They correspond to the 18 graphs shown on the lower side of Fig. 8.

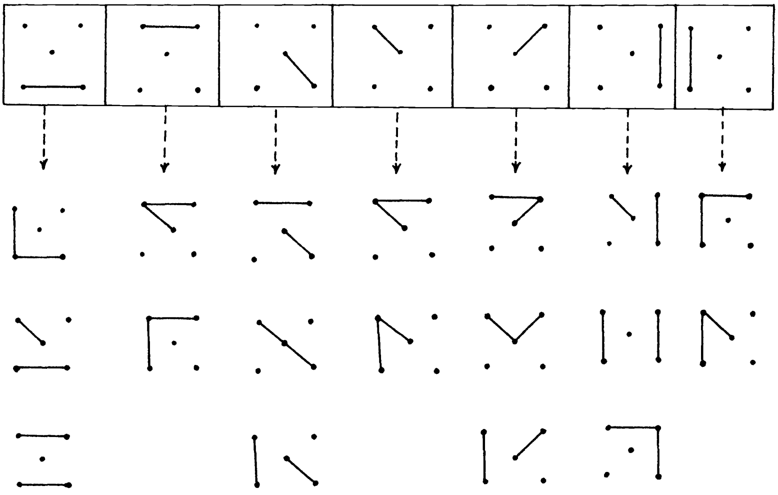


FIG. 8

It is obvious that there are repetitions among the graphs in Fig. 8. Similarly, one finds

$$v_2^\dagger(\text{RHS of }(103)) = [1,2)+[2,3)+[2,4)+[2,5)-(2;1,4) -(2;1,5)-(2;3,4)-(2;3,5) \tag{110}$$

which contributes, under $\omega^\dagger$,

$$8+8+6+8-3-3-3-3=30-12=18. \tag{111}$$

They correspond to the 18 graphs shown in the lower half of Fig. 9.

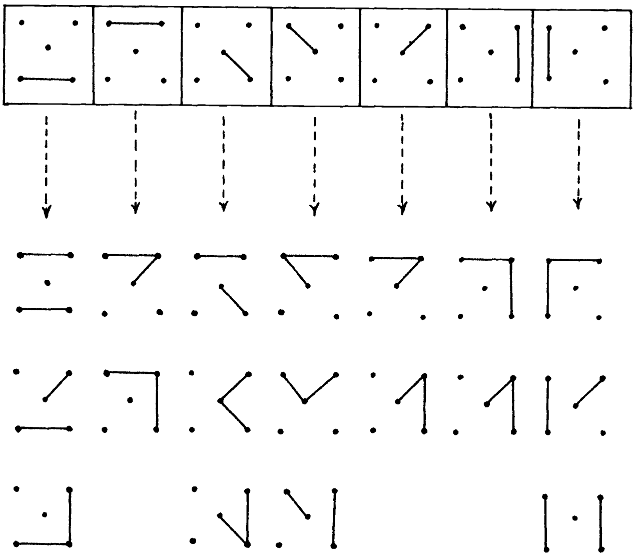


FIG. 9

Next,

$$v_3^\dagger(\text{RHS of }(103)) = [1,3)+[2,3)+[3,4)+[3,5)-(3;1,2) -(3;1,4)-(3;15)-(3;25) \tag{112}$$

which yields under $\omega^\dagger$

(113) $$9+8+5+8-3-3-3-3=30-12=18$$

corresponding to the graphs shown in the lower half of Fig. 10.

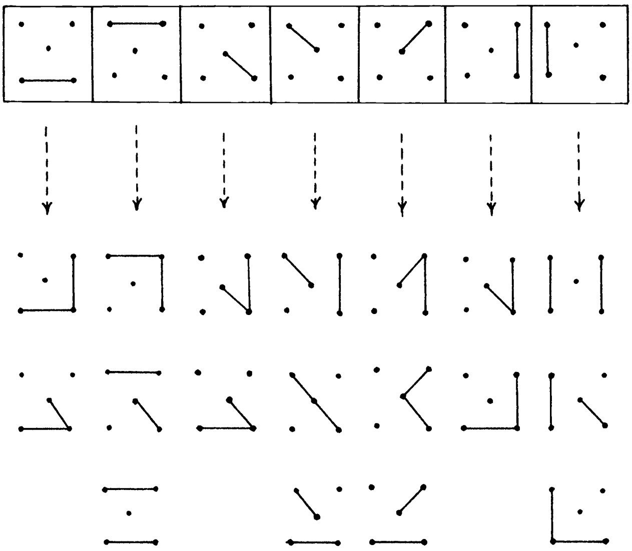


FIG. 10

Next,

$$v_4^\dagger(\text{RHS of }(103)) = [1,4)+[2,4)+[3,4)+[4,5)-(4;1,2)$$
$$-(4;1,5)-(4;2,3)-(4;2,5)-(4;3,5)$$

which yields under $\omega^\dagger$:

(114) $$5+6+5+6-2-2-2-2-2=22-10=12$$

corresponding to the graphs in the lower half of Fig. 11.

For v_5 we have

(115) $$v_5^\dagger(\text{RHS of }(103)) = [1,5)+[2,5)+[3,5)+[4,5)-(5;1,2)$$
$$-(5;1,4)-(5;2,3)-(5;3,4)$$

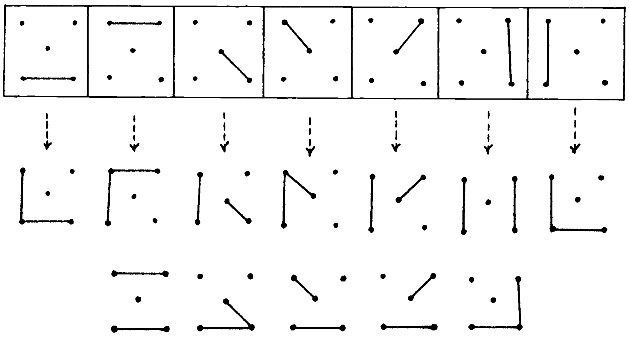


FIG. 11

which yields under $\omega^\dagger$

(116) $$8+8+8+6-3-3-3-3=30-12=18$$

corresponding to the graphs in the lower half of Fig. 12.

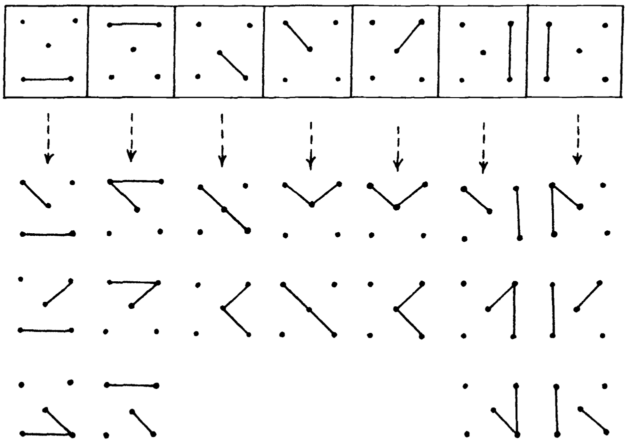


FIG. 12

Next, we evaluate the contribution due to the edge-creation operator, i.e. due to the terms in $e^\dagger\chi^\dagger(\varnothing;\varnothing)$. The edge-creation operators to be considered, for the kite graph, are:

(117) $$e_{12}^\dagger, e_{14}^\dagger, e_{15}^\dagger, e_{23}^\dagger, e_{25}^\dagger, e_{34}^\dagger, e_{35}^\dagger.$$

First, for $e_{12}^\dagger$, we have

(118) $e_{12}^\dagger(\text{RHS of }(103))=(1,2;3)+(1,2;4)+(1,2;5)-(1,2,3,4]-(1,2,3,5]$

which yields under $\omega^\dagger$

(119) $$3+2+3-1-1=6.$$

For $e_{14}^\dagger, e_{15}^\dagger, \cdots, e_{35}^\dagger$, we have the following contributions:

$e_{14}^\dagger:(1,4;2)+(1,4;3)+(1,4;5)-(1,2,3,4]-(1,3,4,5]-(1,2,4,5]$

$\to 3+3+3-1-1-1=6.$

$e_{15}^\dagger:(1,5;2)+(1,5;3)+(1,5;4)-(1,2,3,5]-(1,2,3,4]$

$\to 3+3+2-1-1=6.$

$e_{23}^\dagger:(2,3;1)+(2,3;4)+(2,3;5)-(1,2,3,4]-(1,2,3,5]$

$\to 3+2+3-1-1=6.$

$e_{25}^\dagger:(2,5;1)+(2,5;3)+(2,5;4)-(1,2,4,5]-(2,3,4,5]$

$\to 3+3+2-1-1=6.$

$e_{34}^\dagger:(3,4;1)+(3,4;2)+(3,4;5)-(1,2,3,4]-(1,2,3,5]-(2,3,4,5]$

$\to 3+3+3-1-1-1=6.$

$e_{35}^\dagger:(3,5;1)+(3,5;2)+(3,5;4)-(1,2,3,5]-(1,3,4,5]$

$\to 3+3+2-1-1=6.$

Hence the total contribution due to the edge operator is

$$-6-6-6-6-6-6-6=-42. \tag{120}$$

Summing up the contributions due to both vertex operator and edge operator, i.e. from (109), (111), (113), (114), (116) and (120), we have

$$18+18+18+12+18-42=84-42=42. \tag{121}$$

Hence

$$\omega^{\dagger}:\frac{1}{2!}(\chi^{\dagger})^2(\varnothing;\varnothing)=\frac{1}{2!}(42)=21 \tag{122}$$

which checks the answer of (32) that was obtained in a very simple and direct manner. The point is, as we again emphasize here, that the formalism discussed in this section is used to show the beautiful symmetry of the so-called Euler operator. This also serves to illustrate that a formal simplicity in mathematics does not necessarily imply a computational simplification and, ironically, it could sometimes mean the contrary.

6. A useful function that leads to recurrence relations. In the enumeration of $T^{(m)}$ and σ_m, all the formal operator-free expressions of [1] and [2] involve multiple summations. They are actually too complicated for practical evaluation. One way to resolve this computational difficulty is to find either a function that can be expanded reductively or some recurrent relations that relate the relevant quantities involved. Furthermore, it would be even better if we need only *one* such function to compute *both* $T^{(m)}$ and σ_m. This section deals with this function; it is a real-valued function, on symmetric matrices, that can be reduced recurrently. Its application is such that if a Kirchhoff matrix is substituted into the argument it yields $T^{(m)}$. If, instead, the cycle-adjacency matrix is substituted then one finds σ_m. Furthermore, when the adjacency matrix is used one finds the enumeration of matchings [4].

We now proceed to define this function. Let $\mathbf{X}$ be an $n\times n$ symmetric matrix. Denote by x_{ij} the entries of $\mathbf{X}$ and by $X(i,j,\cdots)$ the principal submatrix obtained from $\mathbf{X}$ after a simultaneous deletion of the ith row and column, the jth row and column, etc. Denote

$$\mathbf{n}\equiv\{1,2,\cdots,n\}, \tag{123}$$

$$\mathbf{X}(1,2,\cdots,n)\equiv 1, \tag{124}$$

and also, for $i_j\in\mathbf{n}$

$$\mathbf{X}(i_1,\cdots,i_m)\equiv 0 \quad \text{if } i_j=i_k \text{ for some } j\neq k. \tag{125}$$

The sought after function is defined by

$$\langle\mathbf{X}\rangle_{q;r}\equiv\frac{1}{q!\,r!}\sum_{i_1,\cdots,i_q}^{n}\sum_{j_1<h_1}^{n}\cdots\sum_{j_r<h_r}x_{j_1h_1}\cdots x_{j_rh_r}\det\mathbf{X}(I_q\cup J_r\cup H_r), \tag{126}$$

if not both q and r are zero, and

$$\langle\mathbf{X}(t_1,\cdots,t_m)\rangle_{0;0}\equiv\det\mathbf{X}(t_1,\cdots,t_m), \tag{127}$$

where

$$I_q\equiv\{i_1,\cdots,i_q\},\quad J_r\equiv\{j_1,\cdots,j_r\}\quad\text{and}\quad H_r\equiv\{h_1,\cdots,h_r\} \tag{128}$$

with the understanding that (125) is in force, i.e. $\langle\mathbf{X}\rangle_{q;r}=0$ if I_q, J_r and H_r are not mutually disjoint sets.

The following *recurrent* properties can be directly verified:
THEOREM.

$$\begin{aligned}\langle \mathbf{X}\rangle_{q;r} &= \frac{1}{q}\sum_{i=1}^{n} \langle \mathbf{X}(i)\rangle_{q-1;r}\\ \langle \mathbf{X}\rangle_{q;r} &= \frac{1}{r}\sum_{j<h}^{n} x_{jh}\langle \mathbf{X}(j, h)\rangle_{q;r-1}.\end{aligned} \tag{129}$$

The matrix-forest theorem, i.e. (12), leads to

$$T^{(m)} = \sum_{r=0}^{\overline{m-1}} \langle \mathbf{K}(n)\rangle_{m-r-1;r} \tag{130}$$

which is an alternative to expression (16) of [1]. It is also easy to see that (64) can be written into:

$$\sigma_m = \sum_{r=0}^{m} \langle \mathbf{E}\rangle_{m-r;r}. \tag{131}$$

We now consider the enumeration of matchings (a matching is a collection of independent edges) in a graph. Denote by $m^{(r)}\{\mathbf{A}\}$ the total number of matchings, each consisting of r edges, for a graph with adjacency matrix $\mathbf{A}$. Theoretically, $m^{(r)}\{\mathbf{A}\}$ may be written into

$$m^{(r)}\{\mathbf{A}\} = \frac{1}{r!}\sum_{j_1<h_1}^{n} \cdots \sum_{j_r<h_r}^{n} a_{j_1h_1}\cdots a_{j_rh_r}. \tag{132}$$

It follows from (131) that

$$m^{(r)}\{\mathbf{A}\} = \langle \mathbf{A}\rangle_{n-2r;r}. \tag{133}$$

In summary, we have just shown the useful role the function

$$\mathbf{X} \mapsto \langle \mathbf{X}\rangle_{q;r}$$

plays in securing recurrence relations and also the "unifying" nature of this function in that a substitution of Kirchhoff, cycle-adjacency and adjacency matrices leads to, respectively, the enumeration of m-forest, spanning subgraphs with given cyclomatic number and r-matchings.

REFERENCES

[1] C. J. LIU AND YUTZE CHOW, *Enumeration of forests in a graph*, Proc. Amer. Math. Soc., 83 (1981), p. 659.

[2] ———, *Enumeration of connected spanning subgraphs of a planar graph*, Acta Mathematica (Budapest), 41 (1983), p. 27.

[3] ———, *An operator approach to some graph enumeration problems*, Discrete Math., 44 (1983), p. 285.

[4] ———, *An operator approach to enumeration of forests and matchings of a graph*, Preprint, 1983.

[5] KIRCHHOFF, *Über die Auflösung der Gleichungen auf welch manbeider Untersuchung der linearen Verteilung galvanischer Ströme gefuhrt wird*, Ann. Phys. Chem., 72 (1847), p. 497.

[6] A. CALEY, *A theorem on trees*, Quart. J. Math., 23 (1889), p. 376.

[7] A. RÉNYI, *Some remarks on the theory of trees*, Publ. Math. Int. Hung. Acad. Sci., 4 (1959), p. 73.

[8] J. W. MOON, *Counting labelled trees*, Seminars at the 12th Canadian Mathematical Congress, Canadian Mathematical Monograph No. 1, 1970.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

010

THE MIDDLE-CUT TRIANGULATIONS OF THE *n*-CUBE*

JOHN F. SALLEE†

Abstract. This paper is concerned with the asymptotic behavior of $\varphi(n)$, the minimum number of simplices required to triangulate the n-cube. Such triangulations are of special interest in connection with algorithms for approximating fixed points of continuous mappings. The standard triangulations of I^n use $n!$ simplices. Let $H(n, m) = \{x \in R^n: \sum x_i = m\}$. Then $H(n, m)$ divides I^n into two polytopes which are then triangulated in a certain fashion. If $K(n, m)$ is the cardinality of this triangulation, then $\lim K(n, n/2)/n! = 0$. Hence $\varphi(n)$ is $o(n!)$. Another measure of the efficiency of a triangulation is the diameter of the dual graph and it is shown that this is $O(n^2)$ for the above triangulation. Finally, a pivoting algorithm for the middle-cut triangulations of the n-cube is also presented.

1. Introduction. In 1967, H. Scarf developed a finite algorithm for approximating a fixed point of a continuous mapping of a simplex into itself [11]. This algorithm was refined and extended by H. Kuhn, B. C. Eaves, and O. Merrill, among others. (See Todd [14], Karamardian [5], and Talman and van der Laan [13], [15] for these and other references.) Several algorithms use a technique of pivoting among simplices which triangulate an n-cube. Current triangulations of the n-cube all have $n!$ simplices and there is an exercise in a well-known text [3] which asserts that this is the minimum. However, in 1970, Mara [7], [8] produced triangulations of the 3, 4 and 5 cubes having 5, 16 and 68 simplices respectively. Recently, Cottle [1], Lee [6] and Sallee [9], [10] have shown that 16 is the minimum for the 4-cube. Lee and Sallee independently found generalizations of Mara's triangulation, showing that the 5-cube can be triangulated with 67 simplices and that if P_n is the cardinality of this triangulation of the cube, then $\lim P_n/n! = (e^2 - 2e - 1)/2 = .4762$.

In this paper a new family of triangulations of the cube is presented. Let $I^n = [0, 1]^n$ and

$$H(n, m) = \{x \in R^n: \sum x_i = m\}.$$

The hyperplane $H(n, m)$ divides I^n into two polytopes if $1 \leqq m \leqq n-1$. Let $K(n, m)$ be the cardinality of this triangulation. It will be shown that $\lim K(2m, m)/(2m)! = 0$. If $\varphi(n)$ is the minimum number of simplices required to triangulate the n-cube, it follows that $\varphi(n)$ is $o(n!)$.

Another measure of the efficiency of a triangulation is the diameter of the dual graph whose vertices are the simplices and two vertices will be joined by an edge if the corresponding simplices have a facet in common. A pivoting algorithm is produced and it is shown that the diameter is $O(n^2)$.

2. The middle-cut triangulation of the *n*-cube. Without specific reference, a number of basic properties of (convex) polytopes will be used. They can all be found in the book of Grünbaum [2]. In particular, an *n-polytope* is an n-dimensional, compact, convex set with a finite number of extreme points (vertices). An *n-simplex* is an n-polytope with $n+1$ vertices. An *affine set* is a translate of a linear set. For $A \subseteq R^n$, aff A (con A) is the intersection of all affine (convex) sets which contain A. A *hyperplane* is an $(n-1)$-dimensional affine set.

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Mathematics, University of Washington, Seattle, Washington, 98195.

An *n-complex* is a finite set $\mathbf{C}$ of n-polytopes such that $P \cap Q$ is a face of both P and Q for all $P, Q \in C$. A *triangulation* of $\mathbf{C}$ is a complex $\mathbf{S}$ of n-simplices such that if $P \in \mathbf{C}$, there is a subset $\mathbf{S}_P$ of $\mathbf{S}$ for which $P = \bigcup \mathbf{S}_P$. If $\mathbf{C} = \{P\}$, $\mathbf{S}$ will be called a triangulation of P. If $P = \bigcup \mathbf{C}$, and $\mathbf{S}$ is a triangulation of $\mathbf{C}$, then $\mathbf{S}$ is a triangulation of P.

A face F of a polytope P is said to be *opposite* a vertex v if $v \notin F$. Let $\mathbf{F}$ be the set of pyramids which have the common vertex v and whose bases are the facets of P opposite v. Then the elements of $\mathbf{F}$ have disjoint interiors and $P = \bigcup \mathbf{F}$. If each of these pyramids is similarly decomposed by choosing a distinguished vertex in each base, the resulting set of pyramids will have relatively disjoint interiors. A little effort will show that if this process is continued, the result will be a set of simplices with disjoint interiors whose union is P. In general, this is not a triangulation of P since there is no guarantee that the intersection of a pair of simplices is a face of both.

For the following, see also Hudson [4]. Let V be an ordering of the vertices of the complex $\mathbf{C}$. For each face F of a polytope $P \in \mathbf{C}$, let v_F be the first vertex of F in the ordering V. Each chain

$$P = F_n \supset F_{n-1} \supset F_{n-2} \supset \cdots \supset F_1 \supset F_0 \neq \varnothing$$

with $v_{F_{i+1}} \notin v_{F_i}$ for $1 \leqq i \leqq n$ has an *associated simplex* $S = \operatorname{con}(\{v_{F_n}, v_{F_{n-1}}, \cdots, v_{F_0}\})$. Let $\mathbf{S}(\mathbf{C}, V)$ be the set of all simplices generated by sequences as above. Then Lemma 1 follows.

LEMMA 1. *$S(C, V)$ is a triangulation of the complex C.*

Proof. See [9]. □

The n-cube I^n has many nice properties, one of which is the following:

LEMMA 2. $\operatorname{vert}(H(n, m) \cap I^n) \subseteq \operatorname{vert} I^n$.

Proof. For any hyperplane H and polytope P, the vertices of $H \cap P$ are the vertices of P or the intersection of H with edges of P. Two vertices of I^n form an edge iff they differ in exactly one coordinate. It is clear that any two vertices of I^n on opposite sides of $H(n, m)$ must differ in at least two coordinates, and the result follows. □

Define $H_-(n, m) = \{x \in R^n : \sum x_i \leqq m\}$, $H_+(n, m) = \{x \in R^n : \sum x_i \geqq m\}$, $A(n, m) = H_-(n, m) \cap I^n$ and $B(n, m) = H_+(n, m) \cap I^n$. The complex of interest in this paper is

$$\mathbf{C}(n, m) = \{A(n, m), B(n, m)\}$$

for $1 \leqq m \leqq n-1$. For completeness, let $\mathbf{C}(n, 0) = \{B(n, 0)\}$ and $\mathbf{C}(n, n) = \{A(n, n)\}$. Note that $A(n, m)$ is the part of I^n lying on the origin side of $H(n, m)$ and $B(n, m)$ is the part lying on the side of $H(n, m)$ away from the origin.

In order to apply Lemma 1 to the complex $C(n, m)$, an ordering of the vertices of I^n is required.

An ordering V of vert I^n will be called an *m-ordering* if it has the following properties:

(i) if $\sum v_i < m$ and $\sum u_i = m$, then $v < u$;

(ii) if $\sum v_i < m$ and $\sum v_i < \sum u_i$, then $v < u$;

(iii) if $m < \sum v_i < \sum u_i$, then $u < v$.

The important property of an m-ordering is that a vertex farther away from $H(n, m)$ preceeds a vertex which is closer to $H(n, m)$ and on the same side of the hyperplane. Lemma 1 guarantees that $\mathbf{S}(\mathbf{C}(n, m), V)$ is a triangulation of $\mathbf{C}(n, m)$ and hence a triangulation of I^n.

An n-cube has 2^n vertices and $2n$ facets, each facet being congruent to I^{n-1}. The facets of I^n are defined by

$$F_{ij} = \{x \in I^n : x_i = j\}, \qquad 1 \leqq i \leqq n, \quad j = 0, 1.$$

The maps f_i defined by

$$f_i(x)=(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)$$

map the F_{ij} onto I^{n-1}. It is clear that the m-ordering V induces an m-ordering on the F_{i0} since $v_i=0$ if $v\in F_{i0}$. A little effort similarly will show that V induces an $(m-1)$-ordering on the F_{i1} for if $x\in F_{i1}$, then $x_i=1$ and $\sum x_i-1=\sum f_{i1}(x)$. Thus the ith coordinate can be ignored when considering the ordering of the vertices of the F_{ij}.

LEMMA 3. *Let V and W be any two m-orderings of* vert I^n; *then* $|\mathbf{S}(\mathbf{C}(n,m),V)|=|\mathbf{S}(\mathbf{C}(n,m),W)|$.

Proof. By induction, assume that the lemma holds for $k<n$. The first vertex in $A(n,m)$ is $v_0=(0,0,\cdots,0)$. The facets opposite v_0 are the $F_{i1}\cap H_-(n,m)$ and the facet $M(n,m)=H(n,m)\cap I^n$. $F_{i1}\cap H_-(n,m)\cong A(n-1,m-1)$ and by induction, the cardinality of the triangulations is the same under either ordering.

Let X be an m-ordering of vert I^{n-1} and Y be an $(m-1)$-ordering. If v is any vertex of I^n, the facets of I^n opposite v are F_{ij}, where $j=1-v_i$ for $1\leqq i\leqq n$. Therefore, if $v\in M(n,m)$, m coordinates of v are 1 and the other $n-m$ coordinates are 0. Hence m of the opposite facets are F_{i0}'s and $n-m$ are F_{i1}'s. Next

$$f_{i0}(H(n,m)\cap F_{i0})=H(n-1,m)\cap I^{n-1}=M(n-1,m)$$

and

$$f_{i1}(H(n,m)\cap F_{i1})=H(n-1,m-1)\cap I^{n-1}=M(n-1,m-1).$$

Therefore

$$\text{(a)}\qquad |\mathbf{S}(M(n,m),V)|=m|\mathbf{S}(M(n-1,m),X)|+(n-m)|\mathbf{S}(M(n-1,m-1),Y)|$$

and the same formula holds for W. It follows by induction that

$$|\mathbf{S}(M(n,m),V)|=|\mathbf{S}(M(n,m),W)|.$$

For either V or W, it follows that

$$\text{(b)}\qquad |\mathbf{S}(A(n,m),V)|=n|\mathbf{S}(A(n-1,m-1),Y)|+|\mathbf{S}(M(n,m),V)|.$$

The vertex $v_1=(1,1,\cdots,1)$ is the first vertex in $B(n,m)$ with any m-ordering. The facets opposite v_1 are the $F_{i0}\cap H_+(n,m)$ and $M(n,m)$. Hence

$$\text{(c)}\qquad |\mathbf{S}(B(n,m),V)|=n|\mathbf{S}(B(n-1,m),X)|+|\mathbf{S}(M(n,m),V)|.$$

This proves the lemma.

From formulas (a), (b) and (c) above, it is possible to derive recursive relations which allow the calculation of $|\mathbf{S}(C(n,m),V)|$. Define $g(n,m)=|\mathbf{S}(M(n,m),V)|$, $h(n,m)=|\mathbf{S}(A(n,m),V)|$ and $K(n,m)=|\mathbf{S}(C(n,m),V)|$.

THEOREM 1.

(i) $g(n,0)=0$ *and* $g(n,1)=1$ *for all* $n\geqq 2$.

(ii) $g(n,m)=g(n,n-m)$.

(iii) $g(n,m)=mg(n-1,m)+(n-m)g(n-1,m-1)$.

(iv) $h(n,m)=g(n,m)+nh(n-1,m-1)$

$$=\cdots$$

$$=\sum_{i=0}^{m-1}\frac{n!}{(n-i)!}\cdot g(n-i,m-i).$$

(v) $K(n,m)=K(n,n-m)=h(n,m)+h(n,n-m)$.

Proof. (i) follows from the fact that $M(n,0)$ is a point and $M(n,1)$ is a simplex. (ii) results from the fact that the cube is symmetric with respect to its centroid. (iii) is just the definition of $g(n,m)$ and formula (a) of Lemma 3. (iv) is formula (b) of Lemma 3. (v) follows from the fact that $B(n,m)\cong A(n,n-m)$. □

LEMMA 4.

$$g(n,m)\leqq\frac{n!}{2^{n-1}}\binom{n-2}{m-1}.$$

Proof. By induction, assume that the result holds for n. By Theorem 1 (iii)

$$\begin{aligned}g(n+1,m)&=mg(n,m)+(n+1-m)g(n,m-1)\\&\leqq\frac{n+1}{2}(g(n,m)+g(n,m-1))\end{aligned}$$

for if $m\leqq n/2$, then $g(n,m)\geqq g(n,m-1)$ and $n+1-m\geqq m$. If $m>n/2$, then $g(n,m-1)\geqq g(n,m)$ and $m\geqq n+1-m$. The result then follows from the addition of binomial coefficients. □

The definition of $g(n,m)$, for $n\geqq 2$, can be extended to include all integers m by setting $g(n,m)=g(n,n-m)=0$ if $m\geqq n$ or $m\leqq 0$. If $m\geqq n$, then $m-1\geqq n-1$ and Theorem 1(iii) holds for the extended function g. Hence the theorem holds for the extended function.

Define $\alpha(n,m,a,i)$ for $0\leqq a\leqq n-2$ and $1\leqq m\leqq n-1$ by

$$g(n,m)=\sum_{i=0}^{a}\alpha(n,m,a,i)\cdot g(n-a,m-i)$$

where the $\alpha(n,m,a,i)$ are defined recursively by Theorem 1(iii). For example $\alpha(n,m,1,0)=m$, $\alpha(n,m,1,1)=n-m$, $\alpha(n,m,2,0)=m^2$, $\alpha(n,m,2,1)=2nm-2m^2-m$ and $\alpha(n,m,2,2)=(n-m)^2$.

LEMMA 5. *The function* α *satisfies the following conditions*:

(i)
$$\alpha(n,m,a+1,i)=\begin{cases}m\cdot\alpha(n,m,a,i) & \textit{if } i=0,\\ (n-a-m+i-1)\cdot\alpha(n,m,a,i-1)\\ \qquad+(m-i)\cdot\alpha(n,m,a,i) & \textit{for } 1\leqq i\leqq a,\\ (n-m)\cdot\alpha(n,m,a,a) & \textit{if } i=a+1;\end{cases}$$

(ii) $\displaystyle\sum_{i=0}^{a}\alpha(n,m,a,i)=\frac{n!}{(n-a)!}$;

(iii) $g(n,m)=\alpha(n,m,n-2,1)$;

Proof. For condition (i), by Theorem 1 (iii)

$$\begin{aligned}g(n,m)&=\sum_{i=0}^{a}\alpha(n,m,a,i)\cdot g(n-a,m-i)\\&=\sum_{i=0}^{a}\alpha(n,m,a,i)[(m-i)\cdot g(n-a-1,m-i)\\&\qquad\qquad+(n-a+m-i)\cdot g(n-a-1,m-i-1)]\end{aligned}$$

$$
\begin{aligned}
&= \alpha(n, m, a, 0) \cdot m \cdot g(n-a-1, m) \\
&\quad + \sum_{i=1}^{a} [(n-a-m+i-1)\alpha(n, m, a, i-1) \\
&\qquad\qquad + (m-i)\alpha(n, m, a, i)]g(n-a-1, m-i) \\
&\quad + (n-m)\alpha(n, m, a, a)g(n-a-1, m-a-1).
\end{aligned}
$$

Condition (i) follows by comparing coefficients.

Condition (ii) is shown to hold by a simple induction. Suppose that $\sum_{i=0}^{a} \alpha(n, m, a+1, i) = n!/(n-a)!$. Then in the above, each $\alpha(n, m, a, i)$ is multiplied by $n-a$ in generating the $\alpha(n, m, a+1, i)$. Hence

$$\sum_{i=0}^{a+1} \alpha(n, m, a+1, i) = (n-a) \cdot \frac{n!}{(n-a)!} = \frac{n!}{(n-(a+1))!}.$$

Since $g(2, i) = 0$ if $i \neq 1$ and $g(2, 1) = 1$,

$$g(n, m) = \sum_{i=0}^{n-2} \alpha(n, m, n-2, i)g(2, i) = \alpha(n, m, 2, 1). \qquad \square$$

LEMMA 6. *Let* $1 \leqq i \leqq m-1$. *If* $m \leqq b \leqq 2m-2$ *and* $1 \leqq m-i \leqq 2m-b-1$, *then* $\alpha(2m, m, b, i) \geqq 0$.

Proof. Since $g(n, m) = 1 \cdot g(2m-0, m-0)$, $\alpha(2m, m, 0, 0) = 1$. Assume that $\alpha(2m, m, b, i) \geqq 0$ for $0 \leqq i \leqq b$ where $b < m$. Then by Lemma 5(i),

$$
\begin{aligned}
\alpha(2m, m, b+1, 0) &= m\alpha(2m, m, b, 0) \geqq 0, \\
\alpha(2m, m, b+1, b+1) &= (2m-m)\alpha(2m, m, b, b) \geqq 0,
\end{aligned}
$$

and

$$
\begin{aligned}
\alpha(2m, m, b+1, i) &= (2m-b-m+i-1)\alpha(2m, m, b, i-1) \\
&\quad + (m-i)\alpha(2m, m, b, i) \\
&\geqq (2m-(m-1)-m+i-1)\alpha(2m, m, b, i-1) \\
&\quad + (m-i)\alpha(2m, m, b, i) \\
&\geqq i\alpha(2m, m, b, i-1) + (m-i)\alpha(2m, m, b, i) \geqq 0
\end{aligned}
$$

for $1 \leqq i \leqq b \leqq m-1$ and each factor in each term above is nonnegative. This establishes the lemma for $b = m$.

The condition that $1 \leqq m-i \leqq 2m-b-1$ is equivalent to the condition that $b-m+1 \leqq i \leqq m-1$. Assume that $m \leqq b < 2m-2$ and that the lemma holds for b. Then

$$
\begin{aligned}
\alpha(2m, m, b+1, i) &= (2m-b-m+i-1)\alpha(2m, m, b, i-1) + (m-i)\alpha(2m, m, b, i) \\
&\geqq (2m-b-m+(b-m+1)-1)\alpha(2m, m, b, i-1) \\
&\quad + (m-(m-1))\alpha(2m, m, b, i) \\
&\geqq 0 \cdot \alpha(2m, m, b, i-1) + 1 \cdot \alpha(2m, m, b, i) \\
&\geqq 0. \qquad \square
\end{aligned}
$$

LEMMA 7. $g(n, m) \leqq (1/m)(2m)!$.

Proof. By Lemma 5 (ii), $\sum_{i=0}^{m} \alpha(2m, m, m, i) = (2m)!/m!$. Since each of the $\alpha(2m, m, m, i)$ is nonnegative, $\alpha(2m, m, m, i) \leqq (2m)!/m!$. Let $b \geqq m$. For i and b

which satisfy Lemma 6,

$$\begin{aligned}\alpha(2m, m, b+1, i) &= (2m-b-m+i-1)\alpha(2m, m, b, i-1) + (m-i)\alpha(2m, m, b, i)\\ &\leqq (2m-b-1)\max\{\alpha(2m, m, b, i-1), \alpha(2m, m, b, i)\}.\end{aligned}$$

In particular, if $b=m$,

$$\alpha(2m, m, m+1, i) \leqq (m-1)\cdot\frac{(2m)!}{m!} = \frac{(m-1)}{m}\cdot\frac{(2m)!}{(m-1)!}.$$

Assume that $\alpha(2m, m, m+j, i) \leqq ((m-j)/m)\cdot(2m)!/(m-j)!$. Then

$$\alpha(2m, m, m+j+1, i) \leqq (m-j-1)\cdot\frac{(m-j)}{m}\cdot\frac{(2m)!}{(m-j)!} \leqq \frac{(m-j-1)}{m}\cdot\frac{(2m)!}{(m-j-1)!}.$$

Hence, by Lemma 5 (iii),

$$g(2m, m) = \alpha(2m, m, 2m-2, 1) \leqq \frac{2}{m}\cdot\frac{(2m)!}{2} = \frac{1}{m}\cdot(2m)!. \qquad \square$$

The notation $a_n \approx b$ is used to mean that $\lim a_n = b$.

THEOREM 2. $\lim_{n\to\infty} K(n, \lfloor n/2\rfloor)/n! = 0$.

Proof. $K(2m, m) = 2h(2m, m)$, and

$$\begin{aligned}K(2m+1, m) &= h(2m+1, m) + h(2m, m)\\ &\leqq 2h(2m+1, m)\\ &\leqq 2g(2m+1, m) + 2(2m+1)\cdot h(2m, m)\end{aligned}$$

by Theorem 1(iv) and (v). By either Lemma 4 or Lemma 7, it is clear that $\lim g(n, m)/n! = 0$. To complete the proof of the theorem, it suffices to show that $\lim_{m\to\infty} h(2m, m)/(2m)! = 0$.

First,

$$\left(\frac{2m-\sqrt{m}\cdot\log m}{2\pi(m-\sqrt{m}\cdot\log m)m}\right)^{1/2} \leqq 1.$$

Next,

$$\left(\frac{m-k\sqrt{m}/2}{m}\right)^{k\sqrt{m}} = \left(1-\frac{k}{2}\cdot\frac{1}{\sqrt{m}}\right)^{\sqrt{m}\cdot k} \approx e^{-(k/2)k} = e^{-k^2/2}.$$

Third,

$$\begin{aligned}&\left(\frac{(m-k\sqrt{m}/2)^2}{(m-k\sqrt{m})m}\right)^{m-k\sqrt{m}} = \left(\frac{m^2-km\sqrt{m}+k^2/4\cdot m}{(m-k\sqrt{m})m}\right)^{m-k\sqrt{m}}\\ &= \left(\frac{(m-k\sqrt{m}+k^2/4}{m-k\sqrt{m}}\right)^{m-k\sqrt{m}} = \left(1+\frac{k^2}{4}\cdot\frac{1}{m-k\sqrt{m}}\right)^{m-k\sqrt{m}} \approx e^{k^2/4}.\end{aligned}$$

By Stirling's formula, $n! \approx \sqrt{2\pi n}(n/e)^n$, hence

$$\frac{1}{2^{2m-k\lfloor\sqrt{m}\rfloor}}\binom{2m-k\lfloor\sqrt{m}\rfloor}{m-k\lfloor\sqrt{m}\rfloor} \approx \left(\frac{2m-k\lfloor\sqrt{m}\rfloor}{2\pi(m-k\lfloor\sqrt{m}\rfloor)m}\right)^{1/2}\frac{((2m-k\lfloor\sqrt{m}\rfloor)/2)^{2m-k\lfloor\sqrt{m}\rfloor}}{(m-k\lfloor\sqrt{m}\rfloor)^{m-k\lfloor\sqrt{m}\rfloor}\cdot m^m}$$

$$\approx\left(\frac{2m-k\sqrt{m}}{2\pi(m-k\sqrt{m})m}\right)^{1/2}$$
$$\cdot\left(\frac{(m-k\sqrt{m}/2)^2}{(m-k\sqrt{m})m}\right)^{m-k\sqrt{m}}\left(\frac{m-k\sqrt{m}/2}{m}\right)^{k\sqrt{m}}$$
$$\leqq 1\cdot e^{k^2/4}\cdot e^{-k^2/2}=e^{-k^2/4}.$$

By Theorem 1(iii) and the fact that $g(n,m)\geqq ng(n-1,m-1)$ for $m\leqq n/2$, it follows that

$$g(n-i,m-i)\geqq\frac{(n-i)!}{(n-i-j)!}\cdot g(n-i-j,m-i-j).$$

Hence

$$\frac{(2m)!}{(2m-i)!}\cdot g(2m-i,m-i)\geqq\frac{(2m)!}{(2m-i-j)!}\cdot g(2m-i-j,m-i-j).$$

By Theorem 3(iv),

$$h(2m,m)=\sum_{i=0}^{m-1}\frac{(2m)!}{(2m-i)!}\cdot g(2m-i,m-i)$$
$$=\sum_{i=0}^{\lfloor\sqrt{m}\cdot\log m\rfloor-1}\frac{(2m)!}{(2m-i)!}\cdot g(2m-i,m-i)$$
$$+\sum_{i=\lfloor\sqrt{m}\cdot\log m\rfloor}^{m-1}\frac{(2m)!}{(2m-i)!}\cdot g(2m-i,m-i)$$
$$\leqq\lfloor\sqrt{m}\cdot\log m\rfloor g(2m,m)$$
$$+m\cdot\frac{(2m)!}{(2m-\lfloor\sqrt{m}\cdot\log m\rfloor)!}\cdot g(2m-\lfloor\sqrt{m}\cdot\log m\rfloor,m-\lfloor\sqrt{m}\cdot\log m\rfloor)$$

by the above

$$\leqq\sqrt{m}\cdot\log m\cdot\frac{1}{m}\cdot(2m)!$$
$$+2m\cdot\frac{(2m)!}{(2m-\lfloor\sqrt{m}\cdot\log m\rfloor)!}\cdot\frac{(2m-\lfloor\sqrt{m}\cdot\log m\rfloor)!}{2^{m-\lfloor\sqrt{m}\cdot\log m\rfloor}}$$
$$\cdot\binom{2m-\lfloor\sqrt{m}\cdot\log m\rfloor}{2m-\lfloor\sqrt{m}\cdot\log m\rfloor}$$

by Lemmas 4 and 7

$$\leqq(2m)!\left(\frac{\sqrt{m}\cdot\log m}{m}+2m\cdot e^{-(\log m)^2/4}\right);$$

by the above

$$\leqq(2m)!\left(\frac{\sqrt{m}\cdot\log m}{m}+\frac{2m}{m^{(\log m)/4}}\right).$$

Hence

$$\frac{h(2m,m)}{(2m)!}\leqq\frac{\sqrt{m}\cdot\log m}{m}+\frac{2m}{m^{(\log m)/4}}.$$

The limit of the right-hand side of this inequality is clearly 0 as $m \to \infty$. This proves the theorem. □

Tables 1 and 2 show the first few values of $g(n, m)$ and give a comparison of $n!$, P_n and $K(n, n/2)$. P_n is the corner-cut triangulation of [6], [9].

TABLE 1.
$g(n, m)$.

m n	2	3	4	5
4	4	·	·	·
5	11	·	·	·
6	26	66	·	·
7	57	302	·	·
8	120	1,191	2,416	·
9	247	4,293	15,619	·
10	502	14,608	88,234	142,684

The other values of $g(n, m)$ are easily found by Theorem 1(i) and (ii).

TABLE 2.
A comparison of $n!$, P_n and $K(n, n/2)$.

n	$n!$	P_n	$K(n, n/2)$
2	2	2	2
3	6	5	5
4	24	16	16
5	120	67	67
6	720	364	324
7	5,040	2,445	1,962
8	40,320	19,296	13,248
9	362,880	173,015	106,181
10	3,628,800	1,720,924	931,300

It should be noted that $\mathbf{S}(\mathbf{C}(n, n), V)$ is known in the literature as the K triangulation.

3. Identifying simplices in S(C(n, m), V). In order to use Lemma 1 and to be able to discuss a middle-cut triangulation, it is necessary to choose a specific m-ordering of vert I^n. For any vertex v of I^n, let COUNT $(v) = \sum v_i$ and VALUE $(v) = \sum v_i \cdot 2^{i-1}$. Note that COUNT (v) is the function used to define an m-ordering. Let V be the m-ordering of vert I^n defined by first comparing COUNT (u) and COUNT (v) and if these values differ, ordering u and v according to properties (i)–(iii) of the definition of an m-ordering; otherwise u and v have the natural ordering given by comparing VALUE (u) and VALUE (v).

Just as the facets of the cuve are defined by setting a particular coordinate to either 0 or 1, any proper face of the cube is defined by choosing a subset L of $\{1, 2, \cdots, n\}$ and for each $i \in L$, setting the ith coordinate to either 0 for all x in the face or to 1 for all x in the face. To be mathematically precise, let $L \subseteq \{1, 2, \cdots, n\}$ and $f: L \to \{0, 1\}$. Then

$$F_{L,f} = \bigcap_{i \in L} F_{i,f(i)}$$

is a face of I^n. It is not difficult to show that F is a proper face of I^n iff there exists an L and a function f as above for which $F = F_{L,f}$.

Since $A(n, m) = I^n \cap H_-(n, m)$, all the proper faces F of $A(n, m)$ are of either *form* 1 where $F = H_-(n, m) \cap F_{L,f}$ or of *form* 2 when $f = H(n, m) \cap F_{L,f}$. A coordinate of a vertex in a face F will be called *fixed* if $i \in L$ and *free* otherwise. The first vertex of $A(n, m)$ in V is $v_0 = (0, 0, \cdots, 0)$. The first vertex v of a face F of form 1 is the vertex v with $v_i = f(i)$ if $i \in L$ and $v_i = 0$ otherwise. The first vertex of a face F of form 2 is found by setting $v_i = f(i)$ if $i \in L$, then setting the first $m - \sum_{i \in L} f(i)$ free coordinates to 1 and the other free coordinates to 0.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	1	0	0
0	0	0	0	1	0	0	1	0	1
1	1	0	0	1	0	0	1	0	1
1	0	1	0	1	0	0	1	0	1
0	0	1	1	1	0	0	1	0	1
0	0	1	0	1	1	0	1	0	1
0	0	1	0	1	0	1	1	0	1
0	0	0	0	1	1	1	1	0	1
0	0	0	0	1	0	1	1	1	1

FIG. 1. *A simplex in* $\mathbf{S}(\mathbf{C}(10, 5), V)$.

As an example of these ideas, Fig. 1 is the matrix of coordinate values of a simplex in $\mathbf{S}(\mathbf{C}(10, 5), V)$. The face opposite $v_{A(10,5)}$ is F_{81}. The vertices $v_3, v_4, \cdots, v_{11}$ all lie in the face $F = F_{L,f} \cap H_-(10, 5)$, where $L = \{5, 8\}$. The vertices $v_7, \cdots, v_{11}$ lie in the face $F = F_{L,f} \cap H(10, 5)$ where $L = \{1, 2, 5, 8, 10\}$ with $f(1) = f(2) = 0$ and $f(5) = f(8) = f(10) = 1$. The first $5 - \text{COUNT}(v) = 2$ free coordinates are 3 and 4 and these are indeed 1.

Using Lemma 1 to construct a simplex in $A(n, m)$, the first vertex is $v_1 = v_0$. The second vertex is in a facet opposite, hence it must have either one nonzero coordinate or exactly the first m coordinates are equal to 1. In the first case, let this coordinate be j, then $L = \{j\}$, and $f(j) = 1$. If the first k vertices have been chosen and $k \leqq m - 1$, then either COUNT $(v_k) = k - 1$ or COUNT $(v_k) = m$. In the first case, $k - 1$ coordinates are fixed.

If COUNT $(v_k) = m$ and $k < n$, then v_k has at most $m - 1$ fixed coordinates which are 1 and at most $n - m - 1$ fixed coordinates which are 0. Also, v_k has at most $k - 2$ fixed coordinates. A face of $A(n, m)$ opposite v_k in $H(n, m)$ has vertices v with $v(i) = 1 - v_k(i)$ for some i.

Thus if $M(S)$ is the *coordinate matrix* of a simplex $S \in \mathbf{S}(A(n, m), V)$, then $s_1(0, 0, \cdots, 0)$ and for $i \geqq 1$, either there exists j such that $s_{k,j} = 1 - s_{i,j}$ for all $k > i$ or $\sum_j s_{k,j} = m$ for all $k \geqq i$ and $\sum_j s_{i,j} < m$.

4. A pivoting procedure for $\mathbf{S}(\mathbf{C}(n, m), V)$. Excellent discussions of the theory and use of pivoting algorithms based on triangulations of R^n to find fixed points of functions $f: R^n \to R^n$ can be found in the books by Todd [14], Talman [13] and van der Laan [15].

Let $\mathbf{S}$ be a triangulation of a polytope P. The *dual graph* G of $\mathbf{S}$ is the graph whose vertices are the simplices of $\mathbf{S}$ and a pair of vertices of G are joined by an edge iff the corresponding simplices have a facet in common. The algorithms in use today find a path in G from a simplex T having certain properties to a simplex S whose vertices have the property that $|f(v) - v)| < \varepsilon$ for $\varepsilon > 0$. Hence all the points in S have this property and are called *approximate fixed points.* Replacing one simplex in the path from T to S by the next simplex in this path is called *pivoting.*

Simplices are readily identified by their vertices. The pivoting algorithms are based on a criterion for replacing a vertex x of a simplex $X \in \mathbf{S}$ with a vertex y of the simplex $Y \in \mathbf{S}$ which shares the facet of X opposite x with X. Thus, given $x \in \operatorname{vert} X$, $X \in \mathbf{S}$, a pivoting procedure must be able to find $y \neq x$ such that $\operatorname{con}((\operatorname{vert} X - \{x\}) \cup \{y\}) \in \mathbf{S}$.

A pivoting procedure must be able to handle pivots for which both simplices belong to the same cube and pivots between simplices which do not belong to the same cube. Pivoting between simplices which do not belong to the same cube will be discussed first.

Let $\{[a_i, b_i]\}$ be an arbitrary set of n unit intervals. Let $v1$ be the vertex of the cube defined by the intervals above whose coordinates are all odd and let $v0$ be the vertex of the cube diagonally opposite $v1$. Each vertex v of this cube has an *associated vertex* t in I^n defined by $t_i = 0$ if $v_i = v0_i$ and $t_i = 1$ if $v_i = v1_i$. It follows that $v_i = v0_i + t_i(v1_i - v0_i)$.

The facets of this cube adjacent to $v0$ and $v1$ are the intersection of the cube with the hyperplanes

$$H(i, v0) = \{x \in R^n \colon x_i = v0_i\}$$

and

$$H(i, v1) = \{x \in R^n \colon x_i = v1_i\}$$

for $1 \leqq i \leqq n$. If the cube is reflected through $H(i, v0)$, then the vertex $v1$ is replaced by the vertex v where

$$v_j = \begin{cases} v1 & \text{if } j \neq i, \\ v1_j - 2(v1_j - v0_j) & \text{if } j = 1. \end{cases}$$

Similarly, if the cube is reflected through the hyperplane $H(i, v1)$, then $v0$ is replaced by the vertex v defined by

$$v_j = \begin{cases} v0_j & \text{if } j \neq i, \\ v0_i + 2(v1_i - v0_i) & \text{if } j = i. \end{cases}$$

If v' is the vertex of the new cube which is the reflection of v, then the vertex of I^n associated with v' is just t, the vertex associated with v.

Let S be an arbitrary simplex in $\mathbf{S}(A(n, m), V)$ and let $M(S)$ be the matrix whose rows are the coordinates of the ordered set of vertices of S. Let $a_i = 0$ if $i = 1$ or $\mathrm{COUNT}\,(s_i) = m$ and $\mathrm{COUNT}\,(s_{i-1}) < m$. Otherwise let $a_i = j$ if the jth coordinate of s_i is fixed and the jth coordinate of s_{i-i} is free. s_n has two free coordinates, hence there are two choices for a_{n+1}. Let a_{n+1} be the rightmost of these coordinates, the one for which $s_{n+1} = 1$. For example, if $M(S)$ is the matrix of Fig. 1, then $a_1 = 0$, $a_2 = 8$, $a_3 = 5$, $a_6 = 2$ $a_7 = 1$, $a_8 = 4$, $a_9 = 7$, $a_{10} = 3$ and $a_{11} = 9$. The two free coordinates of s_{10} are 6 and 9.

Let $L_{ij} = \{a_k \colon k \leqq j\} \cup \{a_i\} - \{0\}$ and define

$$f_{ij}(a_j) = \begin{cases} s_{j,a_j} & \text{if } j \neq i, \\ 1 - s_{i,a_i} & \text{if } j = i \end{cases}$$

for all $a_i L_{ij}$.

If $\mathrm{COUNT}\,(s_j) < m$, let $F'_{ij} = F_{L_{ij}, f_{ij}} \cap H_{-}(n, m)$.

If $\mathrm{COUNT}(s_j) = m$, let $F'_{ij} = F_{L_{ij}, f_{ij}} \cap H(n, m)$.

Finally, let z_{ij} be the first vertex of F'_{ij}.

For example, if $M(S)$ is the matrix of Fig. 1, then $L_{7,4}=\{a_2, a_3, a_4, a_7\}=\{8, 5, 10, 1\}$ and $f_{7,4}(8)=f_{7,4}(5)=f_{7,4}(10)=f_{7,4}(1)=1$ and $z_{7,4}=(1, 0, 0, 0, 1, 0, 0, 1, 0, 1)$. Much more interesting are $z_{4,11}=(1, 1, 1, 0, 1, 0, 0, 1, 0, 0)$ and $z_{9,11}=(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)$, which are the new vertices when the pivoting replaces s_4 and s_9 respectively.

PROCEDURE 1. *Pivoting for a simplex S in $S(A(n, m), V)$ which replaces the kth vertex of S.*

```
Step 1. (replacing $s_1=(0, 0, \cdots, 0)$)
  If $k \neq 1$, go to step 2.
  If COUNT $(s_2)=1$, reflect through $H(a_2, v1)$ as above.
  If COUNT $(s_2)=m$, set $s'_1=(1, 1, \cdots, 1)$.
  Return.
Step 2. (pivoting among vertices not on $M(n, m)$)
  If COUNT $(s_k)=m$, go to step 4.
  If COUNT $(s_{k+1})=m$, go to step 3.
  Do $i=1, n$
      $s'_{k,i}=s_{k+1,i}-s_{k,i}+s_{k-1,i}$
  End do
  Return.
Step 3. (the last vertex not on $M(n, m)$ must be replaced by a vertex on $M(n, m)$)
  Do $j=k+1, n$
    If $s_{j-1}<z_{k,j}<s_j$, then insert $z_{k,j}$; delete $s_k$; return.
  End if.
    If $z_{k,n+1}>s_{n+1}$, then add $z_{k,n+1}$, delete $s_k$.
    Return.
Step 4. (replacing vertices on $M(n, m)$)
    Let $s_i$ be the first vertex on $M(n, m)$.
(If COUNT $(s_k)=m$, then deleting row $k$ from the submatrix $M'(s)$ of $M(S)$ consisting of the rows from $i$ to $n+1$ will either introduce a column of zeros in $M'(S)$, introduce a column of ones into $M'(S)$, introduce a column of ones into $M'(S)$, or the new vertex will be on $M(n, m)$.) Delete row $k$ from $M'(S)$.
    If column $j$ is all ones for some $j$, then set $s'_k=s_{i-1}+e_j$ (the $j$th unit vector); return.
    If column $j$ of $M'(S)$ is all ones, then the pivot is external through the facet
      lying on $H(j, v0)$ and can be handled as described above.
    Do $j=i, n$
      If $s_j<z_{k,j}<s_{k,j+1}$, then insert $z_{k,j}$; delete $s_k$; return.
    End do.
    If $z_{k,n+1}>s_{n+1}$, then add $z_{k,n+1}$ and delete $x_{n+1}$.
  Return.
  End Procedure.
```

That the procedure is correct follows from Lemma 1 and the construction of the triangulation.

5. The diameter of $\mathbf{S}(A(n, m), V)$. The diameter of the K triangulation of I^n is $(n-1)(n-2)/2$. It will be shown in Theorem 3.

THEOREM 3. *The diameter of* $\mathbf{S}(A(n, m), V)$ *is* $O(n^2)$.

Proof. Let S be any simplex in $\mathbf{S}(A(n, m), V)$ and $M(S)$ be its associated matrix. If $1 \leqq \text{COUNT}(s_k)<\text{COUNT}(s_{k+1})=m$, then s_k is replaced by s'_k where COUNT $(s'_k)=m$. Hence S is at most a distance $m-1$ from a simplex in $\mathbf{S}(A(n, m), V)$

whose vertices v_i for $i \geqq 2$ all lie in $M(n, m)$. Let $T(n, m)$ be the matrix as in Fig. 2 whose first row is 0, whose second row contains a block of m ones, and whose next $m-1$ rows are formed by shifting the block of m ones one place to the right and whose last rows are formed by successively shifting the last one in the block as far right as possible.

$$\begin{array}{ccccccc}
0&0&0&0&0&0&0\\
1&1&1&0&0&0&0\\
0&1&1&1&0&0&0\\
0&0&1&1&1&0&0\\
0&0&0&1&1&1&0\\
0&0&0&1&1&0&1\\
0&0&0&1&0&1&1\\
0&0&0&0&1&1&1
\end{array}$$

FIG. 2

Let $M = M(S)$ and M_i be the reduced matrix formed by deleting the rows 2 through $i-1$ and the colums a_3 through a_i, where the a_i are defined as in the last section. M_{n-2} is the matrix

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Suppose $M_i = T(j, k)$ for some j and k. Then M_{i-1} introduces a new row 2 and either a column of zeros or a column of ones as in Fig. 3.

$$\underset{M_i}{\begin{array}{cccccc}
0&0&0&0&0&0\\
1&1&1&0&0&0\\
0&1&1&1&0&0\\
0&0&1&1&1&0\\
0&0&1&1&0&1\\
0&0&1&0&1&1\\
0&0&0&1&1&1\\
 & & & & &
\end{array}}
\qquad
\underset{M_{i-1}}{\begin{array}{ccccccc}
0&0&0&0&0&0&0\\
1&1&1&0&0&0&0\\
1&1&0&1&0&0&0\\
0&1&0&1&1&0&0\\
0&0&0&1&1&1&0\\
0&0&0&1&1&0&1\\
0&0&0&1&0&1&1\\
0&0&0&0&1&1&1
\end{array}}
\quad \text{or} \quad
\underset{M_{i-1}}{\begin{array}{ccccccc}
0&0&0&0&0&0&0\\
1&1&1&1&0&0&0\\
1&1&1&0&0&1&0\\
0&1&1&1&0&1&0\\
0&0&1&1&1&1&0\\
0&0&1&1&0&1&1\\
0&0&1&0&1&1&1\\
0&0&0&1&1&1&1
\end{array}}$$

FIG. 3

In the first case, the block of ones starting in row 3 are successively shifted one place to the right until no further such shifts are possible. In the second case, similar shifts occur. In either case, at most $n-i$ shifts are needed. It follows that any simplex in $\mathbf{S}(A(n, m), V)$ which has a facet in $M(n, m)$ is at most distance $(n-1)(n-2)/2$ from the simplex whose associated matrix is $T(n, m)$. Hence, any simplex is at most $m-1+(n-1)(n-2)/2$ steps from this simplex and the diameter of $\mathbf{S}(A(n, m), V)$ is at most twice the number above. □

6. Concluding remarks. It is clear that $\mathbf{S}(\mathbf{C}(n, m), V)$ is not a minimal triangulation of the n-cube as the facet $M(n, m)$ of $A(n, m)$ and $B(n, m)$ can also be cut near the middle and a suitable ordering of its vertices will result in fewer simplices in the triangulation of $M(n, m)$ and hence in the triangulations of $A(n, m)$ and $B(n, m)$.

It would be nice if a closed form for the estimate of both $K(n, m)$ and the lower bound of $\varphi(n)$, the minimum number of simplices required to triangulate the n-cube, presented in [10] so that these values could be more easily compared. Grünbaum has suggested that lower bounds for $\varphi(n)$ may be found by studying the following two problems. First, what is the minimum cardinality of a set $\mathbf{S}$ for which $\bigcup \mathbf{S} = I^n$? Second, what if the simplices in S are required to have disjoint interiors?

Since the pivoting step for the K triangulation is so simple, it is not clear that the pivoting procedure presented here will be an improvement even though $\mathbf{S}(\mathbf{C}(n, m), V)$ has far fewer simplices.

Finally, the excellent book by Sommerville [12] is highly recommended to the reader with an interest in n-dimensionality geometry.

Acknowledgment. Many thanks are certainly due Victor Klee.

REFERENCES

[1] R. W. COTTLE, *Minimal triangulations of the* 4*-cube*, Discr. Math., 40 (1982), pp. 25–29.

[2] B. GRÜNBAUM, *Convex Polytopes*, Interscience, London, 1967.

[3] J. G. HOCKING AND G. S. YOUNG, *Topology*, Addison-Wesley, Reading, MA, 1961.

[4] J. F. P. HUDSON, *Piecewise linear topology*, University of Chicago Lecture Notes, W. A. Benjamin, New York, 1969.

[5] S. KARAMARDIAN, ed., *Fixed Points* in *Algorithms and Applications*, Academic Press, New York, 1961.

[6] C. W. LEE, *Triangulating the d-cube*, to appear in Discrete Geometry and Convexity, J. E. Goodman et al., eds., New York Academy of Sciences.

[7] P. S. MARA, *Triangulations of a cube*, M.S. thesis, Colorado State Univ., Fort Collins, CO, 1970.

[8] ———, *Triangulations for the cube*, J. Combin. Theory (A), 20 (1976), pp. 170–176.

[9] J. F. SALLEE, *A triangulation of the n-cube*, Discr. Math., 40 (1982), pp. 81–86.

[10] ———, *A note on minimal triangulations of an n-cube*, Discr. Appl. Math., 4 (1982), pp. 211–215.

[11] H. SCARF, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.

[12] D. M. Y. SOMMERVILLE, *An Introduction to the Geometry of n-Dimensions*, Dover, New York, 1958.

[13] A. J. J. TALMAN AND G. VAN DER LAAN, *Variable dimension fixed point algorithms and triangulations*, Mathematical Center, Amsterdam, 1980.

[14] M. J. TODD, *The Computation of Fixed Points and Applications*, Springer-Verlag, Berlin, 1976.

[15] G. VAN DER LAAN AND A. J. J. TALMAN, *Simplicial fixed point algorithms*, Mathematical Center, Amsterdam, 1980.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

011

ON THE ALGORITHMIC COMPLEXITY OF TOTAL DOMINATION*

RENU LASKAR†§, JOHN PFAFF†, S. M. HEDETNIEMI‡ AND S. T. HEDETNIEMI‡

Abstract. A set of vertices D is a dominating set for a graph $G=(V, E)$ if every vertex not in D is adjacent to a vertex in D. A set of vertices is a total dominating set if every vertex in V is adjacent to a vertex in D. Cockayne, Goodman and Hedetniemi presented a linear time algorithm to determine minimum dominating sets for trees. Booth and Johnson established the NP-completeness of the problem for undirected path graphs. This paper presents a linear time algorithm to determine minimum total dominating sets of a tree and shows that for undirected path graphs the problem remains NP-complete.

AMS(MOS) subject classifications. O5C graph theory, 68E discrete mathematics

1. Introduction. We consider undirected graphs $G=(V, E)$ with no loops or multiple edges. A *dominating set* in G is a set D of vertices such that every vertex in $V-D$ is adjacent to at least one vertex in D. The *domination number* of a graph G, denoted $\gamma(G)$, is the minimum number of vertices in a dominating set.

Although the notion of dominating sets of queens on chessboards dates back to the 1800's [1], the modern study of domination can be attributed initially to Ore [15], Berge [2], [3]. For a survey of results on domination see Cockayne and Hedetniemi [5], Cockayne [6] or Laskar and Walikar [13].

For arbitrary graphs the problem of finding a minimum dominating set is NP-complete [10]. For the special case of trees Cockayne, Goodman and Hedetniemi [8] presented a linear time algorithm, which was improved by Natarajan and White [14] for weighted trees.

A graph is an *intersection graph* if there is a correspondence between its vertices and a family of sets (the intersection model) such that two vertices are adjacent in the graph if and only if their two corresponding sets have a nonempty intersection. A graph G is *chordal* if every cycle in G of length >3 has a *chord*, namely an edge joining nonconsecutive vertices on the cycle. It is well known [11] that chordal graphs are exactly the intersection graphs of subtrees of a tree. If the intersection model is further restricted, so that each subtree is a path, a proper subclass of chordal graphs, called *undirected path graphs* results [12]. It was shown by Booth and Johnson [4] that the problem of determining a minimum dominating set remains NP-complete for undirected path graphs. Recently, Farber [9] presented a linear algorithm for strongly chordal graphs, a proper subclass of chordal graphs which includes powers of trees.

In [7] Cockayne, Dawes and Hedetniemi introduced the concept of a total dominating set. A dominating set D is a *total dominating set* if the subgraph $\langle D\rangle$ induced by D has no isolates. The total domination number $\gamma_t(G)$ is the minimum number of vertices in a total dominating set.

In this paper we present a linear algorithm to determine a minimum total dominating set of a tree, and show that for undirected path graphs the problem remains

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631.

‡ Department of Computer Science, Clemson University, Clemson, South Carolina 29631. This research was partly supported by the National Science Foundation under grant MCS-8307832.

§ The research of this author was partly supported by the National Science Foundation under grant ISP-8011451 (EPSCoR).

NP-complete. The problem remains open for directed path graphs, and even for interval graphs.

2. Linear algorithm for finding minimum total dominating set of a tree. Our algorithm actually solves a slightly more general problem, which can be formulated as follows: for a tree T with vertex set V, partitioned into four sets F, B, R_1, and R_2, each consisting of vertices labeled F, B, R_1, and R_2, respectively, we define a *mixed total dominating set* (t-set) of T to be a set TD of vertices of G satisfying:

1. $R_1 \cup R_2 \subset TD$,
2. $x \in B \cup R_1 \rightarrow x$ is adjacent to some vertex in TD.

Clearly, if all vertices of T are labeled B, then a mixed total dominating set of T will be a total dominating set of T, and vice versa. The minimum order of a mixed total dominating set of T will be denoted by $mt(T)$. A linear algorithm for finding a minimum mixed total dominating set (mt-set) of a tree is given.

ALGORITHM MIXED TOTAL DOMINATING. We take as input a tree T with vertices labeled F, B, R_1, or R_2, and produce a set TD which is a minimum mixed total dominating set of T. The algorithm is simply a greedy algorithm which visits an endvertex, makes an appropriate action and deletes this endvertex from the tree, giving a new tree. We denote the endvertex currently being visited by v, and its unique neighbor by u.

Step A. If there are only two vertices left, go to step B. Otherwise let v be an endvertex, adjacent to u.

1. If $v \in F$, delete v.
2. If $v \in B$, and $u \in F \cup R_2$, delete v, and label u with R_2.
3. If $v \in B$, and $u \in B \cup R_1$, delete v, and label u with R_1.
4. If $v \in R_1$, delete v, put v into TD, and label u with R_2.
5. If $v \in R_2$, and $u \in R_1 \cup R_2$, delete v, put v into TD, and label u with R_2.
6. If $v \in R_2$, and $u \in F \cup B$, delete v, put v into TD, and label u with F.

Repeat step A.

Step B. The tree now has only two vertices, u and v.

If u and v both in F, stop.

If v in F, u in B, put v into TD, and stop.

If u and v both in B, put u and v into TD, and stop.

If u or v in R_1, put u and v into TD, and stop.

If u in R_2, v in $F \cup B$, put u into TD, and stop.

If u and v in R_2, put u and v into TD, and stop.

THEOREM 1. *Algorithm Mixed Total Dominating produces a mixed total dominating set of T of minimum order.*

Proof. It is sufficient to consider trees having at least three vertices, since the algorithm clearly finds an mt-set of a two-vertex tree correctly.

Case 1. If $v \in F$ then $mt(T) = mt(T-v)$.

a. Let D be an mt-set of T. If $v \in D$ then u must need to be adjacent to some vertex in D. Let w be any vertex adjacent to u, $w \neq v$. Then $D - \{v\} \cup \{w\}$ is a t-set of $T-v$. Hence $mt(T-v) \leqq mt(T)$. If $v \notin D$ then D is also a t-set of $T-v$. Hence $mt(T-v) \leqq mt(T)$.

b. Conversely, let D be an mt-set of $T-v$. Since $v \in F$, D is also a t-set of T. Thus, $mt(T) \leqq mt(T-v)$.

Thus $mt(T) = mt(T-v)$.

Case 2. If $v \in B$ and $u \in F \cup R_2$, and T' is the tree which results by deleting v from T and labeling u with R_2, then $mt(T) = mt(T')$.

a. Let D be an mt-set of T. Since $v \in B$ we know that $u \in D$. Now if $v \notin D$ then D is a t-set of T' and $mt(T') \leqq mt(T)$. Since the only vertex adjacent to v, namely u, need not be adjacent to any vertex in D, we know $v \notin D$.

b. Conversely let D be an mt-set of T'. Then since $u \in R_2$ (in T') it follows that $u \in D$. Hence D is a t-set of T and $mt(T) \leqq mt(T')$.

Case 3. *If* $v \in B$ and $u \in B \cup R_1$, and T' is the tree which results from deleting v from T and labeling u with R_1, then $mt(T') = mt(T)$.

a. Let D be an mt-set of T. It follows that since $v \in B$, then $u \in D$. Now if $v \notin D$ then D is a t-set of T'. Hence, $mt(T') \leqq mt(T)$. If, however, $v \in D$ then by replacing v with another vertex w adjacent to u, $D - \{v\} \cup \{w\}$ will be a t-set of T', and again $mt(T') \leqq mt(T)$.

b. Conversely, let D be an mt-set of T'. Then since $u \in R_1$, it must follow that $u \in D$. Thus D is also a t-set of T, and $mt(T) \leqq mt(T')$.

Case 4. If $v \in R_1$ and T' is the tree which results from deleting v and labeling u with R_2, then $mt(T) = mt(T') + 1$.

a. Let D be an mt-set of T. Then since $v \in R_1$ we know that v and u are in D. Furthermore, $D - \{v\}$ is a t-set of T' since $u \in R_2$ in T'. Hence $mt(T') \leqq mt(T) - 1$.

b. Let D be an mt-set of T'. Since $u \in R_2$, we know $u \in D$. But then $D \cup \{v\}$ is a t-set of T and $mt(T) \leqq mt(T') + 1$.

Case 5. If $v \in R_2$ and $u \in R_1 \cup R_2$ and T' is the tree which results from deleting v and labeling u with R_2, then $mt(T') + 1 = mt(T)$.

a. Let D be an mt-set of T. Since $u \in R_1 \cup R_2$ and $v \in R_2$ we know that $u \in D$, and furthermore since $u \in R_2$ in T' it follows that $D - \{v\}$ is a t-set of T'. Hence $mt(T') \leqq mt(T) - 1$.

b. Let D be an mt-set of T'. Then $u \in D$ and $D \cup \{v\}$ is a t-set of T. Hence $mt(T) \leqq mt(T') + 1$.

Case 6. If $v \in R_2$ and $u \in F \cup B$, and T' is the tree which results from deleting v and labeling u with F, then $mt(T') = mt(T) - 1$.

a. Let D be an mt-set of T. Then, since $u \in F$ in T' it follows that $D - \{v\}$ is a t-set of T'. Hence, $mt(T') \leqq mt(T) - 1$.

b. Let D be an mt-set of T'. Since $v \in R_2$ in T it follows that $D \cup \{v\}$ is a t-set of T. Hence $mt(T) \leqq mt(T') + 1$.

3. NP-completeness for undirected path graphs. The dominating set problem for undirected path graphs was shown NP-complete by Booth and Johnson [4] using a reduction from the 3-dimensional matching problem. We prove here that the determination of minimum total dominating sets for undirected path graphs is also NP-complete using a slight variation of that reduction. As a matter of fact, the reduction is almost the same as that of Booth and Johnson's except in the formation of one clique which turns a minimum dominating set into a minimum total dominating set. For the sake of completeness, we will describe the reduction completely, which will be mostly Booth and Johnson's reduction.

THEOREM 2. *The problem of finding the total domination number of an undirected path graph is* NP-*complete.*

Proof. We will show that the 3-dimensional matching problem can be reduced to the total domination number problem of a certain undirected path graph.

Following Booth and Johnson [4], let W, X, and Y be three disjoint sets each of cardinality q and let M be a subset of $W \times X \times Y$ having cardinality p. We use the following notation of Booth and Johnson [4]:

$$W = \{w_j | 1 \leqq j \leqq q\},$$

$$X = \{x_k | 1 \leqq k \leqq q\},$$

$$Y = \{y_l | 1 \leqq l \leqq q\},$$

$$M = \{m_i = (w_j, x_k, y_l) | w_j \in W, x_k \in X, y_l \in Y, 1 \leqq i \leqq p\}.$$

For a 3-dimensional matching problem with triples M we may assume that each element of W, X, and Y occurs in at least two triples since otherwise the single triple must occur in any solution so we could reduce the problem to a smaller one. We construct a tree having $6p+3q+1$ cliques from which we will obtain an undirected path graph. The cliques of the tree are explained below.

For each triple m_i in M there are 6 cliques whose vertices depend only upon the triple itself and not upon the elements within the triple. These six cliques form the subtree corresponding to m_i, which is illustrated in Fig. 1.

$\{A_i, B_i, C_i, D_i\}$ (1)

$\{A_i, B_i, D_i, F_i\}$ (2)

$\{C_i, D_i, G_i\}$ (3)

$\{A_i, B_i, D_i, E_i\}$ (4)

$\{A_i, E_i, H_i\}$ (5)

$\{B_i, E_i, I_i\}$ (6) for $1 \leqq i \leqq p$.

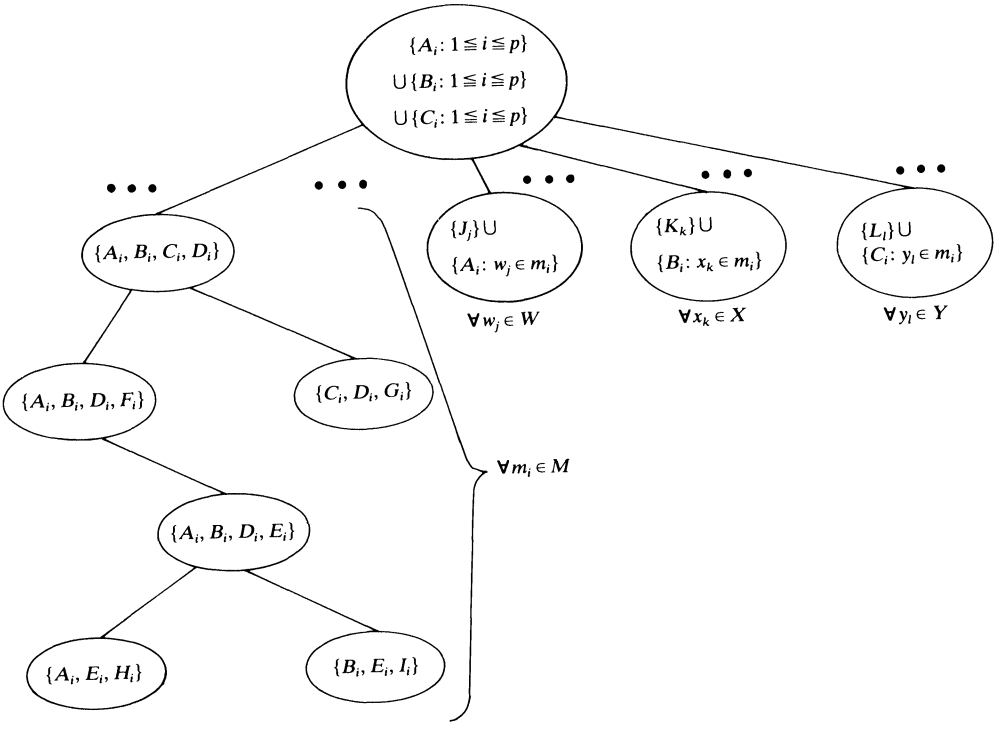


FIG. 1

It may be pointed out here that the above cliques are the same as in Booth and Johnson's paper except the (4) given above, and this little change makes the desired reduction possible.

Next, there is a clique for each element of W, X and Y which depends upon the triples of M to which the element belongs.

$$\{J_j\} \cup \{A_i | w_j \in m_i\} \quad \text{for } 1 \leqq j \leqq q,$$

$$\{K_k\} \cup \{B_i | x_k \in m_i\} \quad \text{for } 1 \leqq k \leqq q,$$

$$\{L_l\} \cup \{C_i | y_l \in m_i\} \quad \text{for } 1 \leqq l \leqq q.$$

And finally there is one large clique, the root of the tree, which contains vertices for each of the triples

$$\{A_i, B_i, C_i | 1 \leqq i \leqq p\}.$$

We see that the sets are cliques by verifying that no set is properly contained within another. We check that each element is contained only in a family of cliques which form an undirected path within the tree; it is then easy to see that there is only one way in which the cliques can be connected into a tree so that these conditions hold. This is the arrangement shown in Fig. 1. We thus know that the graph G whose cliques were built from the 3-dimensional matching problem is an undirected path graph and the clique tree is unique.

We next claim that the undirected path graph G has a total dominating set of size $2p+q$ if and only if the 3-dimensional matching problem has a solution.

Verifying one direction of the claim is easy. If the 3-dimensional matching problem has a solution M' we simply choose for each m_i in M' all of the vertices A_i, B_i and C_i corresponding to that m_i. There are precisely $3q$ of these. For all other m_i not in M' we choose the corresponding D_i and E_i. There are $2p-2q$ of these. Altogether we have chosen $2p+q$ vertices which form a total dominating set of G, since A_i, B_i, C_i are mutually joined, and D_i and E_i are joined.

Conversely, given a total dominating set for G we can assume without loss of generality that for each i either all three of A_i, B_i and C_i or else both of D_i and E_i have been included. This follows from the observation that the only way to totally dominate the subtree corresponding to m_i with two vertices is to choose D_i and E_i, and that any larger total dominating set might just as well consist of A_i, B_i and C_i, since none of the other possible vertices dominate any vertex outside of the subtree.

The proof is completed by noting that if there are t of the m_i for which A_i, B_i and C_i are chosen in a dominating set of size $2p+q$, these account for $3t$ vertices and the remaining E_i and D_i account for $2p-2t$ vertices. It must be the case that $t=q$. Picking the q triples m_i for which A_i, B_i and C_i were chosen yields a solution to the 3-dimensional matching problem. □

REFERENCES

[1] W. W. ROUSE BALL, *Mathematical Recreations and Problems of Past and Present Times*, Macmillan, London, 1892.

[2] C. BERGE, *The Theory of Graphs and Its Applications*, Dunod, Paris, 1958, Methuen, London 1962.

[3] ———, *Graphs and Hypergraphs*, North-Holland, London, 1975.

[4] K. S. BOOTH AND J. H. JOHNSON, *Dominating sets in chordal graphs*, SIAM J. Comput., 11 (1982), pp. 191–199.

[5] E. J. COCKAYNE AND S. T. HEDETNIEMI, *Towards a theory of domination in graphs*, Networks, 7 (1977), pp. 247–261.

[6] E. J. COCKAYNE, *Domination in undirected graphs—a survey, theory and applications of graphs* in America's Bicentennial Year, Y. Alavi and D. R. Lick, eds., Springer-Verlag, Berlin, 1978, pp. 141–147.

[7] E. J. COCKAYNE, R. DAWES AND S. T. HEDETNIEMI, *Total domination in graphs*, Networks, 10 (1980), pp. 211–215.

[8] E. J. COCKAYNE, S. GOODMAN AND S. HEDETNIEMI, *A linear algorithm for the domination number of a tree*, Inform. Processing Lett., 4 (1975), pp. 41–44.

[9] M. FARBER, *Domination, independent domination, and duality in strongly chordal graphs*, Discr. Appl. Math., to appear.

[10] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability—A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[11] F. GAVRIL, *The intersection graphs of subtrees in trees are exactly the chordal graphs*, J. Combin. Theory B, 16 (1974), pp. 47–56.

[12] ———, *A recognition algorithm for the intersection graphs in trees*, Discrete Math., 23 (1978), pp. 221–227.

[13] R. LASKAR AND H. B. WALIKAR, *On domination related concepts in graph theory*, Lecture Notes in Mathematics 885, Springer-Verlag, Berlin, 1980, pp. 308–320.

[14] K. S. NATARAJAN AND L. J. WHITE, *Optimum domination in weighted trees*, Inform. Processing Lett., 7 (1978), pp. 261–265.

[15] O. ORE, *Theory of Graphs*, AMS Colloquium Publications 38, American Mathematical Society, Providence, RI, 1962.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

012

UNLABELLED PARTITION SYSTEMS: OPTIMIZATION AND COMPLEXITY*

P. M. CAMERINI† AND F. MAFFIOLI†

Abstract. In this paper we consider unlabelled partition systems, i.e. independence systems $(S, \mathscr{B})$, where the ground set S—of m elements— is partitioned into n blocks and for each base $B \in \mathscr{B}$ the number of blocks containing i elements of B is exactly c_i— a given nonnegative integer—for each $i = 0, 1, \cdots, m$. For any weighting $w: S \to \mathbb{Z}$, we show that the problem asking for a most weighted base is solvable in polynomial time. When $c_{k-1} + c_k = n$ for some k, $0 < k \leqq m$, we have a matroid, called unlabelled partition matroid. We also introduce a matroid operation, called star, which preserves linear representability. Finally, we investigate the computational complexity of optimum intersection and parity problems for structures of this kind. These arise naturally in many degree-constrained subgraph problems, when only the number of vertices with prescribed degree is assigned, disregarding the vertices identities.

Key words. independence systems, matroids, matching problems, complexity classification

AMS(MOS) subject classifications. 05A05, 05B35, 68C25, 90C09

1. Introduction. A well-solved class of important combinatorial optimization problems is constituted by the degree-constrained subgraph (DCS) problems, both in bipartite and nonbipartite graphs, weighted or not. Not only are the problems solvable in polynomial time when the prescribed degree of each vertex is one (matchings), but also in the more general case of arbitrary degrees, i.e., when for each vertex v a nonnegative integer a_v is given, and the required subgraph has to contain a_v edges incident to v.

Surprisingly little attention has been given to the "unlabelled" DCS problems, namely to the cases where only the number of vertices with prescribed degree is assigned, and the identity of these vertices disregarded. Such problems turn out to arise in some practical applications, such as the optimum assignment of traffic bursts in SS-TDMA systems [4], [5].

In order to investigate in general problems of this kind, in § 2 of this paper we define and study *unlabelled partition systems*, i.e. independence systems whose ground set S of m elements is partitioned into n blocks and whose bases B are such that the number of blocks containing i elements of B is exactly c_i—a given nonnegative integer—for each $i = 0, 1, \cdots, m$. For any weighting $w: S \to \mathbb{Z}$, we show that the problem asking for a most weighted base is solvable in polynomial time. When $c_{k-1} + c_k = n$ for some k, $0 < k \leqq m$, the system is a matroid, which we call *unlabelled partition matroid.*

In § 3 we introduce and discuss a matroid operation, called *star*, yielding a hierarchical matroidal structure and preserving linear representability. In § 4 we investigate thoroughly the computational complexity of both intersection and parity problems involving unlabelled partition systems. In § 5 we motivate this study by applying it to unlabelled degree-contrained subgraph problems and by proposing a reliability problem which can be efficiently solved using the star operation.

* Received by the editors July 6, 1983. This research was partially supported by a research contract of the Italian Ministry of Education. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Dipartimento di Elettronica & Centro di Studio per le Telecomunicazioni Spaziali del CNR-Politecnico di Milano, 20133 Milano, Italy.

2. Unlabelled partition systems. Let $S=\{s_1, s_2, \cdots, s_m\}$ be a *ground set* of *elements*, properly partitioned into *n partition blocks* $S_1, S_2, \cdots, S_n$. For each $X \subseteq S$, $i=0, 1, \cdots, m$, let

$$\gamma_i(X)=|\{S_j\colon |S_j \cap X|=i\}|,$$

i.e., $\gamma_i(X)$ denotes the number of partition blocks having i elements in common with X. Let $c_0, c_1, \cdots, c_m$ be $m+1$ nonnegative integers, called *charges*, such that

$$\sum_{i=0}^{m} c_i = n \tag{2.1}$$

and

$$\sum_{h=i}^{m} c_h \leqq \sum_{h=i}^{m} \gamma_h(S) \tag{2.2}$$

for each $i=1, \cdots, m$.

Because of (2.1), (2.2) the set $\mathscr{B}$ of subsets X of S such that

$$\gamma_i(X)=c_i \tag{2.3}$$

for each $i=0, 1, \cdots, m$, is a nonempty set. Moreover, since all members of $\mathscr{B}$ have the same cardinality,

$$r=\sum_{i=1}^{m} i \cdot c_i, \tag{2.4}$$

the following property is satisfied.

(P.1) $X \in \mathscr{B}$ and $X \subsetneq X'$ implies $X' \notin \mathscr{B}$.

The pair $U=(S, \mathscr{B})$ is therefore an *independence system* [6], [9] which we call an *unlabelled partition system* (UPS). The members of $\mathscr{B}$ are called the *bases* of U. Any subset of a base is an *independent set* of U.

We say that a UPS is given in *concise form* or *concisely*, when its bases, rather than being listed explicitly, are specified as above, i.e. by giving the integer m, the n partition blocks, and $m+1$ nonnegative integers c_i, $i=0, 1, \cdots, m$, satisfying (2.1) and (2.2). Notice that the size of the input data needed to specify an UPS in concise form is bounded above by a polynomial in m—the cardinality of the ground set.

We are concerned with the following decision problem:

WEIGHTED UPS BASE (WUB)
Instance. A UPS $U=(S, \mathscr{B})$, given in concise form;
 a *weight* $w(s) \in \mathbb{Z}$ for each $s \in S$;
 a *threshold* $W \in \mathbb{Z}$.
Question. Does there exist a base $X \in \mathscr{B}$ (i.e. a subset X of S satisfying (2.3)) whose *total weight* $w(X)=\sum_{s \in X} w(s)$ is not smaller than W?

THEOREM 2.1. WUB *is solvable in polynomial time.*

Proof. Consider the following algorithm, whose input is any instance of WUB.

```
begin
1. let $G=(M, N, E)$ be a bipartite graph, where
      $M=\{u_0, u_1, \cdots, u_m\}$, $N=\{v_1, \cdots, v_n\}$ and
      $E=\{\{u_i, v_j\}\colon u_i \in M,\ v_j \in N,\ i \leqq |S_j|\}$;
2. for each edge $e=\{u_i, v_j\}$ of $G$ do
      let $z(e)$ be the sum of the weights of (any) $i$ most weighted elements in $S_j$
      ($z(e)=0$ if $i=0$);
```

```
3. find in $G$ a set $F^*$ of edges such that
    (i) exactly $c_i$ edges of $F^*$ are incident to $u_i$, $i=0,1,\cdots,m$,
    (ii) exactly one edge of $F^*$ is incident to $v_j$, $j=1,\cdots,n$,
    (iii) $Z=\sum_{e\in F^*} z(e)$ is maximized, subject to constraints (i) and (ii) above;
4. if $Z\geqq W$ then return "yes" else return "no"
end
```

We show that this algorithm solves WUB in polynomial time.

First, it is easy to see that if X^* is any UPS base, then

$$F^*=\{\{u_i, v_j\}: i=|S_j\cap X^*|\}$$

satisfies conditions (i) and (ii) of step 3. Moreover, if X^* is an UPS of maximum total weight, then

$$w(X^*)=\sum_{e\in F^*} z(e).$$

(In order for X^* to have maximum total weight, the elements of $S_j\cap X^*$ must be $|S_j\cap X^*|$ most weighted elements of S_j.)

Second, if F^* is any set of edges of G satisfying the three conditions of step 3, then we can construct a corresponding set X^*, by taking for each edge $e=\{u_i, v_j\}\in F^*$, i most weighted elements of S_j, and therefore X^* is an UPS base, of total weight

$$w(X^*)=\sum_{e\in F^*} z(e)=Z.$$

Because of these two facts, the value of Z found in (iii) of step 3 is the maximum total weight of a UPS base, and step 4 provides a correct answer. To see that the algorithm runs in polynomial time, it is enough to note that standard matching techniques [8] can be utilized for executing step 3. □

It is well known [13] that an independence system $(S, \mathscr{B})$ is a *matroid* (defined in terms of its bases) if and only if—in addition to (P.1)—the following property is satisfied.

(P.2) $X\in\mathscr{B}$, $X'\in\mathscr{B}$ and $x\in X$ implies $X-\{x\}\cup\{x'\}\in\mathscr{B}$ for some $x'\in X'$.

The theorem below will be proved in more general form in the next section and is stated here to point out that a UPS is a matroid, called *unlabelled partition matroid* (UPM), when only two consecutive charges, say c_{k-1} and c_k, are different from zero.

THEOREM 2.2. *A UPS such that* $c_{k-1}+c_k=n$ *for some* $k\in\{1,\cdots,m\}$ *is a matroid.*

An UPM for which $c_{k-1}+c_k=n$ is also indicated as a k-UPM, and can be defined in terms of its independent sets, as $M=(S, \mathscr{I})$, where

$$\mathscr{I}=\{I\subseteq S: \gamma_k(I)\leqq c_k, |I\cap S_j|\leqq k, j=1,\cdots,n\}.$$

(Notice that a k-UPM with $c_k=n$ is an ordinary partition matroid [10], whose independent sets have no more than k elements in each partition block.)

As a consequence of Theorem 2, when $c_{k-1}+c_k=n$ for some $k\in\{1,\cdots,m\}$, problem WUB can be solved by the "greedy" algorithm [10], which is much simpler and faster than the method (see proof of Theorem 1) utilized for the general case.

On the other hand, when two nonconsecutive charges are different from zero, the greedy algorithm cannot be used in general, since the corresponding UPS is not necessarily a matroid, as the following example shows. Let $S=\{s_1, s_2, s_3, s_4\}$, $S_1=\{s_1, s_2\}$, $S_2=\{s_3, s_4\}$, $c_0=c_2=1$, $c_1=c_3=c_4=0$: the set of bases $\mathscr{B}=\{S_1, S_2\}$ does not satisfy property (P.2).

3. The star of matroids. In this section we generalize the definition of unlabelled partition matroid, using an operation, called *star*, which can be made on any matroid and any sequence of n matroids, n being the number of elements of the first matroid. This operation yields a new matroid and preserves linear representability, i.e. the star of matroids linearly representable over the same field is linearly representable.

Let N be a matroid, with ground set $T=\{t_1,\cdots,t_n\}$ and base set $\mathscr{C}$. Let $M_1,\cdots,M_n$ be matroids, S_j and $\mathscr{B}_j$ denoting respectively the ground and the base sets of M_j, $j=1,\cdots,n$. All these matroids are arbitrary, but for the fact that we assume their ground sets to be pairwise disjoint, and $|T|=n$. This latter assumption allows the index j to establish a natural one-to-one correspondence between any element t_j of T and the matroid M_j. We say that t_j is the *image* (in T) of any base of M_j.

Let s be the *rank* (i.e. the cardinality of the bases) of N. For each $j=1,\cdots,n$, let r_j be the rank of M_j and

$$\mathscr{A}_j=\{X\subset S_j\colon \exists x\in S_j-X \text{ s.t. } X\cup\{x\}\in\mathscr{B}_j\}$$

be the set of the *sub-bases* of M_j, that is $\mathscr{A}_j$ contains all the independent sets of M_j with r_j-1 elements, or equivalently all the maximal independent sets contained into the hyperplanes [13] of M_j.

Let

$$S=\bigcup_{j=1}^{n} S_j$$

and for each $X\subseteq S$,

$$B(X)=\{t_j\colon X\cap S_j\in\mathscr{B}_j\}.$$

Let also

$$\mathscr{B}=\{X\subseteq S\colon B(X)\in\mathscr{C}, X\cap S_j\in\mathscr{A}_j\cup\mathscr{B}_j, j=1,\cdots,n\};$$

i.e. $\mathscr{B}$ contains every subset of S such that (i) its intersection with each S_j is either a sub-base or a base of M_j, and (ii) the elements of T which are images of these latter bases, form a base of N.

We call $M=(S,\mathscr{B})$ the *star* of N and $M_1,\cdots,M_n$ and write $M=N*(M_1,\cdots,M_n)$ to denote this operation.

Observe that if N is a *free* matroid (i.e. a matroid having the ground set as its unique base), then $N*(M_1,\cdots,M_n)$ is the direct sum [13] of $M_1,\cdots,M_n$.

THEOREM 3.1. *The star of matroids is a matroid.*

Proof. We show that the pair $M=(S,\mathscr{B})$ defined above satisfies properties (P.1) and (P.2) of § 2.

Since all members of $\mathscr{B}$ have the same cardinality

$$r=s+\sum_{j=1}^{n}(r_j-1),$$

(P.1) is obviously satisfied.

In order to show that (P.2) is also satisfied, let X, X' be any two members of $\mathscr{B}$ and let x be any element of X. Let h be the index such that $x\in S_h$. There are two cases.

Case A. $|X\cap S_h|\leqq|X'\cap S_h|$. Hence the independent set of M_h

$$I=(X\cap S_h)-\{x\},$$

has fewer elements than the independent set $I'=X'\cap S_h$.

Because of the properties of the independent sets of a matroid [13], it follows that there exists an element $x' \in I' - I \subseteq X'$ such that $I \cup \{x'\}$ is independent in M_h. Moreover, $I \cup \{x'\}$ is either a base or a sub-base, depending on whether $X \cap S_h$ is a base or a sub-base. Therefore

$$X - \{x\} \cup \{x'\} \in \mathscr{B}.$$

Case B. $|X \cap S_h| > |X' \cap S_h|$. Hence $X \cap S_h$ is a base, $X' \cap S_h$ is a sub-base of M_h, and $t_h \in B(X) - B(X')$. Since both $B(X)$ and $B(X')$ are bases of N, and property (P.2) holds for $\mathscr{C}$, it follows that $B(X) - \{t_h\} \cup \{t_{h'}\} \in \mathscr{C}$ for some $t_{h'} \in B(X') - B(X)$. As a consequence, $X' \cap S_{h'}$ is a base and $X \cap S_{h'}$ is a sub-base of $M_{h'}$. Because of the properties of the independent sets, there exists an element x' of $(X' \cap S_{h'}) - (X \cap S_{h'}) \subseteq X'$ such that $(X \cap S_{h'}) \cup \{x'\}$ is a base of $M_{h'}$, whose image is $t_{h'}$, so that

$$X - \{x\} \cup \{x'\} \in \mathscr{B}. \qquad \square$$

Theorem 3.1 can also be proved as in [3] utilizing the following elementary operations, studied in [2].

(A) *Free extension.* For any matroid L on a ground set R, the free extension of L by an element x is the matroid L' on the set $R \cup \{x\}$, whose bases are all the bases of L, together with all sets obtained by adding x to any sub-base of L.

(B) *Two-sum.* For any two matroids L_1, L_2 on ground sets respectively R_1, R_2, such that $R_1 \cap R_2 = \{y\}$, the two-sum $T_y(L_1, L_2)$ is a matroid on the ground set $R_1 \cup R_2 - \{y\}$, whose bases are all sets of the form $B_1 \cup A_2$ or $A_1 \cup B_2$, where for each $i = 1, 2$, B_i is any base of L_i not containing y, and A_i is any sub-base of L_i such that $A_i \cup \{y\}$ is a base of L_i.

It turns out that

$$N * (M_1, \cdots, M_n) = T_{t_n}(T_{t_{n-1}}(\cdots T_{t_1}(N, M_1') \cdots M_{n-1}'), M_n'),$$

where for each $j = 1, \cdots, n$, M_j' denotes the free extension of M_j by t_j.

The proof of Theorem 2.2 follows easily from Theorem 3.1, observing that any k-UPM $M = (S, \mathscr{B})$ is given by

$$U[T, c_k] * (U[S_1, k], \cdots, U[S_n, k]),$$

where we denote by $U[R, \rho]$ the *uniform matroid* [13] on ground set R, whose bases are all subsets of R with ρ elements.

Since it is known [2] that both operations (A) and (B) preserve linear representability, it follows that the star of matroids, all linearly representable over the same field F, is linearly representable over a suitable extension of F. We refer the reader to [3] for a general way of constructing such a linear representation. In the particular case of a k-UPM, a possible representation over the field of reals is given by the following $r \times m$ matrix:

$$\begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_n \\ \mathbf{T}_1 & \mathbf{T}_2 & \cdots & \mathbf{T}_n \end{bmatrix}$$

where for $j = 1, \cdots, n$

$$\mathbf{S}_j = \begin{bmatrix} 1 & 2 & \cdots & |S_j| \\ 1 & 4 & \cdots & |S_j|^2 \\ \vdots & \vdots & \cdots & \vdots \\ 1^{k-1} & 2^{k-1} & \cdots & |S_j|^{k-1} \end{bmatrix}$$

is a $(k-1)\times|S_j|$ matrix, and

$$\mathbf{T}_j=\begin{bmatrix} j^0 & \cdots & j^0 \\ j^1 & \cdots & j^1 \\ \cdots & \cdots & \cdots \\ j^{c_k-1} & \cdots & j^{c_k-1} \end{bmatrix}$$

is a $c_k\times|S_j|$ matrix.

When, as it happens above, the representation over the field of reals involves only not too large integers, the problem of finding (if any) a base respecting 2-parity conditions may be solved efficiently by the random polynomial algorithm of [11]. However, for the case of a k-UPM, an efficient polynomial deterministic algorithm will be proposed in the next section (see Corollary 4.1).

Finally, let us observe that the star operation can obviously be performed iteratively, thus yielding a multi-level matroidal structure. We shall see in § 5 a possible application of this fact.

4. Intersection and parity problems. In this section we investigate the computational complexity of problems involving more than one UPS, or imposing parity constraints on the elements of the ground set. Specifically, we consider (p-) *intersection problems,* asking for a base—of total weight not smaller than a given threshold W—common to p UPS's given in concise form, all having the same rank and sharing the same ground set of m weighted elements. These problems are denoted by a sequence of p "fields" separated by vertical bars: each field specifies in increasing order, for the corresponding UPS, the indices of those charges which are allowed to be greater than zero. For instance

$$\langle 2,4|3\rangle$$

identifies the problem of intersecting two UPS's: the first having all charges equal to zero, but (possibly) c_2 and c_4; the second with a unique nonzero charge, d_3. This means that for any base X common to both UPS's, c_2 partition blocks of the first UPS share 2 elements with X, and the remaining (c_4) partition blocks have 4 elements in common with X, whereas all (d_3) partition blocks of the second UPS, share 3 elements with X.

We shall also consider (q-) *parity problems,* where a single UPS $U=(S,\mathcal{B})$, weighted on elements, is given in concise form, together with a proper partition of S into *parity blocks* of equal cardinality q (q being a divisor of $|S|$), and we ask for a base $X\in\mathcal{B}$—of total weight not smaller than a given threshold W—satisfying *parity conditions,* i.e. each parity block must share with X either all, or none of its elements. Parity problems of this kind are denoted as

$$\langle h_1,\cdots,h_l\colon q\rangle,$$

where (similarly as for intersection problems) $h_1,\cdots,h_l$ are the indices—in increasing order—of those charges of U, which are allowed to be greater than zero.

Obviously, when $p=q=1$, both intersection and parity problems specialize to problem WUB considered in § 2. A less trivial observation is the following.

Remark 4.1. Any parity problem $\langle h_1,\cdots,h_l\colon q\rangle$ is a special case of $\langle h_1,\cdots,h_l|0,q\rangle$, where the partition blocks of the second UPS have equal cardinality q.

Because of Theorem 2.2, we also note that any problem $\langle h,h+1|k,k+1\rangle$ asks for a base, of weight not smaller than W, common to two (unlabelled partition) matroids, and hence is solvable in polynomial time [10, Ch. 8]. Similar to the problem of intersecting two ordinary partition matroids [10, 12], the optimal intersection of two UPM's can be found by defining a suitable min-cost flow problem, with lower and upper bounds to arc flows (see [14], [5] for an application).

A more general result is the following.

THEOREM 4.1. *Any problem $\langle h, h+1|k, k+1, k+2\rangle$ is solvable in polynomial time.*

Proof. We show that, using polynomial time, it is possible to transform any instance of this problem into an equivalent instance of the weighted degree-constrained subgraph problem [8, Ch. 7]. To this purpose, let us consider the following construction.

Let n_1 and n_2 be respectively the number of partition blocks of the first and the second UPS. For each $i=1, \cdots, m$, let c_i and d_i be respectively the ith charge of the first and the second UPS. Let $V_1=\{v_j\colon j=1, \cdots, n_1\}$ be a set of vertices, in one-to-one correspondence with the partition blocks of the first UPS. For each $l=1, \cdots, n_2$, we associate to the lth partition block of the second UPS a pair of vertices u_l, z_l connected by an edge of zero weight. Let V_2, E_2 denote respectively the set of all these vertices and edges. To each element s_i $(i=1, \cdots, m)$ of the ground set S, assuming that s_i belongs to the jth partition block of the first, and to the lth partition block of the second UPS, we associate the graph of Fig. 1, weighted on edges as shown (i.e. all edges have zero weight, but for $\{v_j, x_i\}$, whose weight is the same as that of s_i). Let E_3 and $V_3=\{x_i, y_i\colon i=1, \cdots, m\}$ denote respectively the set of all edges of these graphs, and the set of all vertices not in $V_1 \cup V_2$. Let $V_4=\{\sigma, \tau\}$ be a set of two other vertices, and let

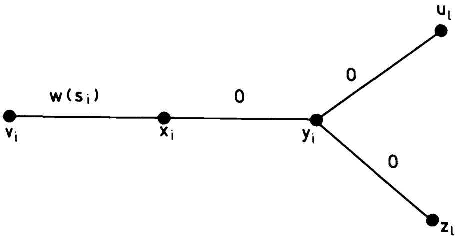


FIG. 1

$$E_1=\{\{\sigma, v_j\}\colon j=1, \cdots, n_1\},$$

$$E_4=\{\{\tau, u_l\}\colon l=1, \cdots, n_2\},$$

be two other sets of edges of zero weight. Let $G=(V, E)$ be the graph such that $V=\bigcup_{\nu=1}^{4} V_\nu$ and $E=\bigcup_{\nu=1}^{4} E_\nu$.

We claim that to any subset F of E having exactly

(i) c_h edges incident to σ,

(ii) d_{k+1} edges incident to τ,

(iii) $h+1$ edges incident to each vertex of V_1,

(iv) $k+1$ edges incident to each vertex z_l $(l=1, \cdots, n_2)$,

(v) one edge incident to each other vertex,

there corresponds one-to-one a base X common to both UPS's, whose total weight is the same as that of F. Once this claim is proved, the theorem follows immediately, since the given construction provides a polynomial transformation [7] from $\langle h, h+1|k, k+1, k+2\rangle$ to the weighted degree-constrained subgraph problem. Figure 2 illustrates an example of this construction, for an instance of $\langle 1, 2|1, 2, 3\rangle$, with $c_1=0$, $c_2=3$, $d_1=d_2=d_3=1$; $n_1=n_2=3$; $m=9$; partition blocks of the first UPS: $\{s_1, s_2, s_3\}$,

$\{s_4, s_5, s_6\}$, $\{s_7, s_8, s_9\}$; partition blocks of the second UPS: $\{s_1, s_4, s_7\}$, $\{s_2, s_4, s_8\}$, $\{s_3, s_6, s_9\}$; $w(s_i) = i$, $i = 1, \cdots, 9$.

Figure 2 represents the corresponding graph G (only nonzero weights are shown): the wavy edges identify the subset F corresponding to $X = \{s_1, s_3, s_5, s_6, s_7, s_9\}$.

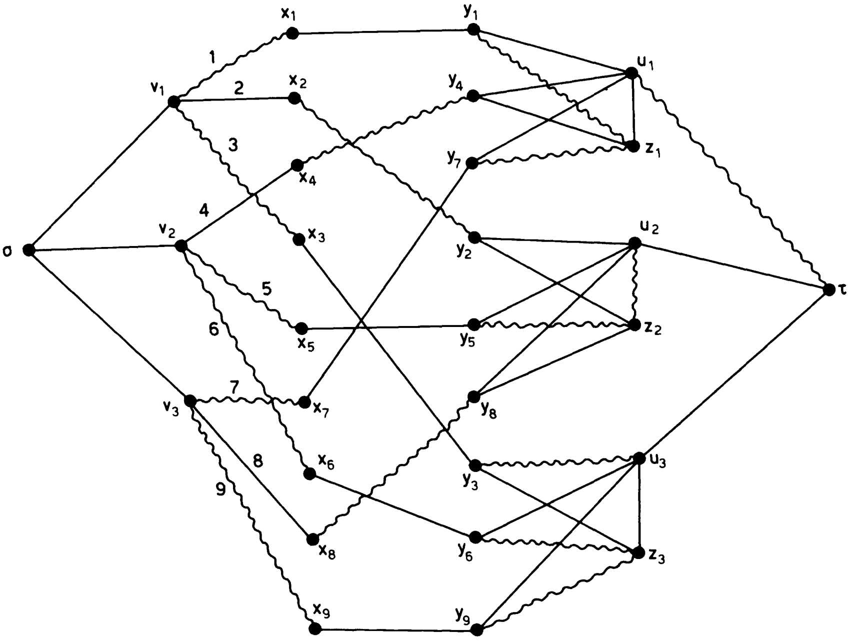


FIG. 2

In order to prove the claim, suppose that we are given a subset F of E satisfying properties (i) through (v) above. Let

$$X = \{s_i \in S\colon \{v_j, x_i\} \in F \text{ for some } j\}.$$

It is trivial to see that F and X have the same total weight. Because of (i) and (iii), there are exactly c_h (respectively $c_{h+1} = n_1 - c_h$) vertices of V_1 incident to exactly h (respectively $h+1$) edges of $F - E_1$. It follows that X is a base of the first UPS. In order to show that X is also a base of the second UPS, observe that for each $l = 1, \cdots, n_2$ the unique edge of F incident to u_l can be either

(a) edge $\{u_l, z_l\}$, or

(b) edge $\{\tau, u_l\}$, or

(c) an edge $\{u_l, y_i\}$ for some i.

In case (a), because of property (iv), F contains k edges of the kind $\{y_i, z_l\}$; for each one of these edges, $\{x_i, y_i\} \notin F$, so that $\{v_j, x_i\} \in F$ for some j; it follows that X contains exactly k elements of the lth partition block of the second UPS.

In case (b), again by property (iv), F contains $k+1$ edges of the kind $\{y_i, z_l\}$; similarly as for case (a), it follows that X contains exactly $k+1$ elements of the lth partition block.

In case (c), F contains $k+1$ edges of the kind $\{y_i, z_l\}$ and one edge of the kind $\{u_l, y_i\}$, so that X contains exactly $k+2$ elements of the lth partition block.

Because of property (ii), case (b) occurs exactly d_{k+1} times, and therefore exactly d_{k+1} partition blocks of the second UPS contain exactly $k+1$ elements of X. Moreover, letting δ_a, δ_c denote respectively the number of times that case (a) and case (c) occur, we have

$$\delta_a + d_{k+1} + \delta_c = n_2,$$
$$k \cdot \delta_a + (k+1)d_{k+1} + (k+2)\delta_c = |X| = r,$$

where r is the rank of both UPS's.

Since the above system of linear equations in the unknowns δ_a, δ_c has a unique solution, and since d_k and d_{k+2} satisfy the system, it follows that exactly d_k (respectively d_{k+2}) partition blocks of the second UPS contain exactly k (respectively $k+2$) elements of X, and therefore X is a base of the second UPS.

Conversely, it is easy to see that given a base X common to both UPS's, one can always construct a corresponding subset F of E satisfying properties (i) through (v), and having total weight equal to that of X. □

COROLLARY 4.1. *Any problem $\langle h, h+1: 2\rangle$ is solvable in polynomial time.*

Proof. This result follows immediately from Remark 4.1 and Theorem 4.1. □

We now turn our attention to NP-complete problems.

THEOREM 4.2. *Problem $\langle 1|0, k\rangle$ is* NP*-complete, for any $k \geqq 3$.*

Proof. Since it is obvious that this problem belongs to NP, we exhibit a polynomial transformation from the following well-known NP-complete problem [7].

EXACT COVER BY 3 SETS
Instance. Set $Q = \{1, \cdots, 3q\}$;
collection C of three-element subsets of Q.
Question. Does C contain an *exact cover* for Q, i.e. a subcollection $C' \subseteq C$ such that every element of Q occurs in exactly one member of C'?

Let

$$S = \{\{x, y\}: y \in C \text{ and either } x \in y \text{ or } x \in Q'\},$$

where $Q' = \{3q+1, \cdots, k \cdot q\}$. For each $j = 1, \cdots, 3q$, let

$$H_j = \{\{j, y\}: j \in y \in C\};$$

for each $j = 3q+1, \cdots, k \cdot q$, let

$$H_j = \{\{j, y\}: y \in C\};$$

for each $l = 1, \cdots, |C|$, let

$$K_l = \{\{x, y_l\}: \text{either } x \in y_l \text{ or } x \in Q'\},$$

where y_l is the lth member of C.

Let U_1 be a UPS having ground set S; partition blocks H_j, $j = 1, \cdots, k \cdot q$; all charges c_i's equal to zero, except $c_1 = k \cdot q$.

Let U_2 be a UPS having the same ground set; partition blocks K_l, $l = 1, \cdots, |C|$; all charges d_i's equal to zero, except $d_0 = |C| - q$, $d_k = q$.

It is easy to see that C contains an exact cover for Q if and only if there exists a base common to both U_1 and U_2. Since the problem asking for *any* base common to U_1, U_2 is a specialization of $\langle 1|0, k\rangle$, the proof follows. □

Theorem 4.2 can be extended with the help of the two following lemmas, which we prove quite concisely.

LEMMA 4.1. *Any problem $\langle h_1, \cdots, h_p|0, k_1, \cdots, k_q\rangle$ transforms polynomially into $\langle h_1, \cdots, h_p|g, k_1+g, \cdots, k_q+g\rangle$ for any $g \in \{1, \cdots, m+k_q\}$, $h_1 > 0$.*

Proof. Let an instance of the first problem be denoted as in the proof of Theorems 4.1, 4.2. We shall use a corresponding primed notation for describing an instance of the second problem. Let

$$n'_1 = n_1 + \left\lceil \frac{n_2}{h_1} \right\rceil \cdot g, \qquad n'_2 = \left\lceil \frac{n_2}{h_1} \right\rceil \cdot h_1;$$

$$S' = \{s_1, \cdots, s_m, s_{m+1}, \cdots, s_{m'}\}$$

where

$$m' = m + n'_2 \cdot g;$$

for each $j = 1, \cdots, n_1$

$$H'_j = H_j;$$

for each $j = n_1 + 1, \cdots, n'_1$

$$H'_j = \{s_{m+(j-n_1-1)h_1+\alpha}\colon \alpha = 1, \cdots, h_1\};$$

for each $l = 1, \cdots, n_2$

$$K'_l = K_l \cup \{s_{m+(l-1)g+\beta}\colon \beta = 1, \cdots, g\};$$

for each $l = n_2 + 1, \cdots, n'_2$

$$K'_l = \{s_{m+(l-1)g+\beta}\colon \beta = 1, \cdots, g\};$$

$$c'_{h_1} = c_{h_1} + n'_1 - n_1, \qquad c'_{h_\alpha} = c_{h_\alpha} (\alpha = 2, \cdots, p);$$

$$d'_g = d_0 + n'_2 - n_2, \qquad d'_{k_\beta + g} = d_{k_\beta} (\beta = 1, \cdots, q);$$

for each $i = 1, \cdots, m$

$$w'(s_i) = w(s_i);$$

for each $i = m + 1, \cdots, m'$

$$w'(s_i) = 0;$$

$$W' = W.$$

□

LEMMA 4.2. *Any problem* $\langle 1|0, k\rangle$ *with* $k \geqq 2$ *transforms polynomially into* $\langle h|0, k\rangle$ *for any* $h \in \{2, \cdots, m\}$.

Proof. Using again the notation of the previous proof to denote corresponding instances of the two problems, we assume that $\delta = (k-1) \cdot h$ divides n_1, for otherwise the following construction yields another equivalent instance of $\langle 1|0, k\rangle$, which satisfies such an assumption.

Add to the original ground set $(\delta - 1) \cdot n_1$ new elements of zero weight; partition these new elements in the two following ways, so as to obtain two sets of new partition blocks, which are added respectively to the two original sets of partition blocks: the first partition consists simply of as many singletons as there are new elements, whereas the second partition is formed by $(\delta - 1) \cdot d_k$ blocks of k elements each (since $n_1 = d_k \cdot k$ is the rank of both UPS's, these two partitions are possible); if we multiply by δ all charges but d_0 it is easy to see that this enlarged instance of $\langle 1|0, k\rangle$ is a yes-instance if and only if the original instance is a yes-instance, and the assumption that δ divides n_1 applies.

Let now

$$n_1' = n_1 + \frac{n_1 \cdot (h-1)}{\delta}, \qquad n_2' = n_2 + \frac{n_1 \cdot (h-1)}{k-1};$$

$$S' = \{s_1, \cdots, s_m, s_{m+1}, \cdots, s_{m'}\},$$

where

$$m' = m + \frac{k \cdot n_1 \cdot (h-1)}{k-1};$$

for each $j = 1, \cdots, n_1$

$$H_j' = H_j \cup \{s_{m+(j-1)\cdot(h-1)+\alpha}: \alpha = 1, \cdots, h-1\};$$

for each $j = n_1 + 1, \cdots, n_1'$

$$H_j' = \{s_{m+(h-1)\cdot n_1 + (j-n_1-1)\cdot h + \alpha}: \alpha = 1, \cdots, h\};$$

for each $l = 1, \cdots, n_2$

$$K_l' = K_l;$$

for each $l = n_2 + 1, \cdots, n_2'$

$$K_l' = \{s_{m+(l-n_2-1)\cdot(k-1)+\beta}: \beta = 1, \cdots, k-1\} \cup \{s_{m+(h-1)\cdot n_1 + l - n_2}\};$$

$$c_h' = n_1';$$

$$d_0' = d_0, \qquad d_k' = d_k + \frac{n_1 \cdot (h-1)}{k-1};$$

for each $i = 1, \cdots, m$

$$w'(s_i) = w(s_i);$$

for each $i = m+1, \cdots, m'$

$$w'(s_i) = 0;$$

$$W' = W. \qquad \square$$

THEOREM 4.3. *Any problem* $\langle h | k_1, k_2 \rangle$ *is* NP-*complete whenever* $k_2 - k_1 \geqq 3$.

Proof. The proof follows immediately from Theorem 4.2 and Lemmas 4.1, 4.3. □

The two following lemmas and Theorem 4.4 are in order to show the NP-completeness of some 2-parity problems.

LEMMA 4.3. *Any problem* $\langle 0, k | 0, k \rangle$ *transforms polynomially into* $\langle 0, k{:}\ 2 \rangle$.

Proof. The first problem is a special case of the second, where the partition blocks can be subdivided into two groups such that no two elements of a parity block belong to partition blocks of the same group. □

LEMMA 4.4. *Any problem* $\langle 0, h{:}\ 2 \rangle$ *transforms polynomially into* $\langle g, h+g{:}\ 2 \rangle$ *for any* $g \in \{1, \cdots, m-h\}$.

Proof. Given any instance of the first problem, we construct a corresponding instance of the second problem in the following way: For each partition block H_j of the first instance, (i) add to the ground set $2 \cdot g$ new elements, partitioned into g new parity blocks $\{a_{j,k}, \bar{a}_{j,k}\}$, $k = 1, \cdots g$; (ii) add $a_{j,k}$ $(k = 1, \cdots, g)$ to H_j; (iii) add a new partition block $\bar{H}_j = \{\bar{a}_{j,k}: k = 1, \cdots, g\}$. Let then the new nonzero charges d_g and d_{h+g} be respectively $c_0 + n$ (n being the number of the original partition blocks) and c_h. □

THEOREM 4.4. *Any problem* $\langle h_1, h_2\colon 2\rangle$ *is* NP*-complete whenever* $h_2-h_1\geqq 3$.

Proof. The proof follows easily from Theorem 4.3 and Lemmas 4.3, 4.4. □

Further results are stated in the following theorem.

THEOREM 4.5. *All p-intersection and q-parity problems with* $p\geqq 3$, $q\geqq 3$ *are* NP*-complete.*

Proof. It is well known that the following problem is NP-complete [7].

3-DIMENSIONAL MATCHING.

Instance. Set $S\subseteq X\times Y\times Z$, where X, Y and Z are disjoint sets having the same number n of elements.

Question. Does S contain a matching, i.e. a subset $X\subseteq S$ such that $|X|=n$ and no two elements of X agree in any coordinate?

It is obvious that this problem is a special case of $\langle 1|1|1\rangle$, where each element of the ground set S has unit weight, and the threshold W equals the number of partition blocks. It is easy to show that $\langle 1|1|1\rangle$ transforms polynomially both to $\langle h\colon 3\rangle$ and to $\langle h|k|l\rangle$ for any h, k, l, thus implying the theorem. □

Remark 4.2. All the NP-completeness proofs given in this section show in fact that the corresponding problems remain in NP-complete, even in their simpler version asking for the existence of any base satisfying the intersection or parity conditions.

The results of Theorems 4.1, 4.3, 4.4, 4.5 and Corollary 4.1 may be summarized by saying that almost all UPS intersection and parity problems have been classified according to their complexity.

The only problems whose complexity remain to be studied are the 2-intersection and the 2-parity problems for which the index range of the nonzero charges is two. Among these problems, the most crucial one is perhaps $\langle 0,2\colon 2\rangle$, for the two following reasons. First, this problem is equivalent to $\langle 0,2|0,2\rangle$, because of Remark 4.1 and Lemma 4.3. Second, should $\langle 0,2\colon 2\rangle$ turn out to be NP-complete, the borderline between "easy" and "hard" problems would become quite sharp, since the following Lemma 4.5, together with Lemmas 4.1, 4.4, would imply the NP-completeness of both $\langle h_1, h_2|k_1, k_2\rangle$ and $\langle h_1, h_2\colon 2\rangle$ whenever $h_2-h_1\geqq 2$ and $k_2-k_1\geqq 2$.

LEMMA 4.5. *Any problem* $\langle 0, h|0, k\rangle$ *with* $k\geqq 2$ *transforms polynomially into* $\langle g, h+g|0, k\rangle$ *for any* $g\in\{1,\cdots,m-h\}$.

Proof. Using the notation of the proof of Lemmas 4.1, 4.2 to denote corresponding instances of the two problems, we assume that $k-1$ divides n_1, for otherwise the following construction yields another equivalent instance of $\langle 0, h|0, k\rangle$ satisfying the assumption: Add to the original ground set $\lceil n_1/(k-1)\rceil - n_1$ new elements of zero weight; partition these new elements into as many singletons; add these singletons to both the original sets of partition blocks; add the number of new elements to both c_0 and d_0.

Let now

$$n_1' = n_1+\frac{n_1}{k-1},\qquad n_2' = n_2+\frac{n\cdot g}{k-1};$$

$$S' = \{s_1,\cdots,s_m,s_{m+1},\cdots,s_{m'}\}$$

where

$$m' = m+n_1'\cdot g;$$

for each $j=1,\cdots,n_1$

$$H_j' = H_j\cup\{\widehat{s^{m+(j-1)\cdot g+\alpha}}\colon \alpha=1,\cdots,g\};$$

for each $j = n_1 + 1, \cdots, n_1'$

$$H_j' = \{s_{m+(j-1)\cdot g+\alpha}\colon \alpha = 1, \cdots, g\};$$

for each $l = 1, \cdots, n_2$

$$K_l' = K_l;$$

for each $l = n_2 + 1, \cdots n_2'$

$$K_l' = \{s_{m+(l-n_2-1)\cdot(k-1)+\beta}\colon \beta = 1, \cdots, k-1\} \cup \{s_{m+n_1\cdot g+l-n_2}\};$$

$$c_g' = c_0 + \frac{n_1}{k-1}, \qquad c_{h+g}' = c_h;$$

$$d_0' = d_0, \quad d_k' = d_k + \frac{n_1 \cdot g}{k-1};$$

for each $i = 1, \cdots, m$

$$w'(s_i) = w(s_i);$$

for each $i = m+1, \cdots, m'$

$$w'(s_i) = 0;$$

$$W' = W. \qquad \square$$

5. Some applications. As we pointed out in the introduction, an important class of combinatorial optimization problems which can be solved in polynomial time is constituted by the degree-constrained subgraph (DCS) problems, both in bipartite and nonbipartite graphs, weighted or not. More formally, for any set X of edges and any vertex v in a given undirected graph $G = (V, E)$, let $\alpha_v(X)$ denote the *degree* of v in the graph (V, X), i.e. the number of edges of X incident to vertex v. (Here and in the sequel, loops are counted twice.) The following search problem has been extensively studied and shown to be solvable in polynomial time [8, Ch. 7].

DEGREE CONSTRAINED SUBGRAPH (DCS).
Given. An undirected graph $G = (V, E)$;
a nonnegative integer a_v for each $v \in V$;
a weight $w(e) \in \mathbb{Z}$ for each $e \in E$.
Find. A subset X of E such that $\alpha_v(X) = a_v$ for each $v \in V$ and the *total weight* $w(X) = \sum_{e \in X} w(e)$ is maximized.

The special case of DCS, where $a_v = 1$ for each $v \in V$ is the well-known perfect matching problem.

Surprisingly little attention has been given so far to the "unlabelled" versions of these problems, namely to problems where only the *number* of vertices with prescribed degree is assigned, without imposing any constraint on the identity of such vertices. Specifically, for any set X of edges in the graph $G = (V, E)$ and for any integer $i = 0, 1, \cdots, |E|$, let $\beta_i(X)$ be the number of vertices of G which are incident to exactly i edges of X. For any set of edges in the bipartite graph $G = (V', V'', E)$ and for any integer $i = 0, 1, \cdots, |E|$, let $\beta_i'(X)$ and $\beta_i''(X)$ be respectively the number of vertices in V' and V'' which are incident to exactly i edges of X. We may then consider the two following search [7] problems.

Unlabelled DCS (UDCS).
Given. An undirected graph $G=(V,E)$;
a nonnegative integer b_i for each $i=0,1,\cdots,2\cdot|E|$;
a weight $w(e)\in\mathbb{Z}$ for each $e\in E$.
Find. A subset X of E such that $\beta_i(X)=b_i$ for each $i=0,1,\cdots,2\cdot|E|$ and $w(X)$ is maximized.

Unlabelled Bipartite DCS (UBDCS).
Given. An undirect bipartite graph $G=(V',V'',E)$;
two nonnegative integers b'_i, b''_i for each $i=0,1,\cdots,|E|$;
a weight $w(e)\in\mathbb{Z}$ for each $e\in E$.
Find. A subset X of E such that $\beta'_i(X)=b'_i$, $\beta_i(X)=b''_i$ for each $i=0,1,\cdots,|E|$ and $w(X)$ is maximized.

It is easy to see that UBDCS and UDCS problems are natural interpretations—in terms of graph optimization—of the 2-intersection and the 2-parity UPS problems, respectively, which have been discussed in the previous section. More precisely, the following one-to-one correspondence between instances of the UBDCS (in its obvious decision form) and instances of the 2-intersection UPS problem preserves the correct answer: $S=E$; for each $j=1,\cdots,|V'|$, the jth partition block of the first UPS contains the edges of G incident to the jth vertex of V'; for each $l=1,\cdots,|V''|$, the lth partition block of the second UPS contains the edges incident to the lth vertex of V''; the charges of the first and the second UPS are $c_i=b'_i$ and $d_i=b''_i$ $(i=0,1,\cdots,|E|)$ respectively.

Similarly, equivalent instances of the UDCS and the 2-parity UPS problem correspond to each other as follows: $S=\{s_1,\bar{s}_1,\cdots,s_m,\bar{s}_m\}$ with $m=|E|$; for each $i=1,\cdots,|E|$, the two elements s_i, $\bar{s}_i$ of the ground set S correspond to the ith edge $e_i=(x_i,\bar{x}_i)$ of G and constitute the ith parity block of the UPS; for each $j=1,\cdots,|V|$, the jth partition block contains all elements s_i or $\bar{s}_i$ such that x_i or $\bar{x}_i$, respectively, coincides with the jth vertex of G; the charges are $c_i=b_i$, $i=0,1,\cdots,2\cdot|E|$; for each $i=1,\cdots,|E|$ the weight of the ith edge e_i is the sum of the weights of the corresponding elements s_i, $\bar{s}_i$.

The complexity results of § 4 can then be directly utilized for studying these unlabelled DCS problems. In particular we note that the open problem $\langle 0,2\colon 2\rangle$ (in its "unweighted" version asking for any base satisfying parity conditions) is equivalent to the problem:

Exact Partial Covering by Circuits (EPCC).
Instance. A graph $G=(V,E)$;
a positive integer $c\leqq|V|$
Question. Does G contain a set of vertex-disjoint circuits covering *exactly* c vertices?

It is easy to see that the nonbipartite matching techniques solve in polynomial time the similar problem asking for the existence of vertex-disjoint circuits covering *at least* c vertices.

We conclude this section by suggesting a possible application—in the field of system reliability—of the UPM's and the star operation studied in §§ 2 and 3.

Being $k\geqq 2$ any fixed integer, for any nonnegative integer l, a (*level*) *l-component* is defined recursively as follows, in terms of its *elements*:
a 0-*component* is an *element*,
an *l-component* is a set of at least k $(l-1)$-components.

Any l-component can be in one of the following three (mutually exclusive) *states*, listed from *better* to *worse*:

a *working*, or *w-state*;

a *critical*, or *c-state*;

a *broken*, or *b-state.*

For any $l > 0$, the state of any l-component T is determined by the states of its $(l-1)$-components according to the following rules:

T is in w-state iff none of its components is in b-state *and* at least k of them are in w-state;

T is in c-state iff none of its components is in b-state *and* exactly $k-1$ of them are in w-state;

T is in b-state otherwise, i.e. iff at least one of its components is in b-state, *or* less then $k-1$ of them are in w-state.

From these rules it follows that the states of all elements determine uniquely the state of any l-component T. In particular, if all elements are in the working state, so is T; if one or more elements are in the broken state, so is T. Note also that T is a *coherent system* in the sense of [1], i.e. improving (worsening) the state of any element either leaves unchanged, or improves (worsens) the state of T. When T is in the critical state, worsening the state of any element causes T to switch to the broken state.

If "costs" are associated to w-states of elements, we might be interested in considering *minimal working sets* for T, i.e. sets of elements such that assigning to them the w- state, and letting the remaining elements be in c-state, T is in w-state, but whenever any element in w-state switches to c-state, so does T. We also consider *critical sets*, i.e. sets of elements such that assigning to them the w-state, and letting the remaining ones be in c-state, T is in c-state. In addition to costs, "weights" may be associated to elements: for instance the *weight* $w(s_i)$ of element s_i may be $\log p_i$, p_i being the probability that s_i remains in *w-state*, once such state has been initially assigned to s_i. Assuming statistical independence among the element states, we may then wish to find a minimal working set X which maximizes for T the probability of remaining in the working state, i.e. maximizes the *total weight* $w(X) = \sum_{s_i \in X} w(s_i)$. This problem can be solved in polynomial time, by the greedy algorithm, as shown by the following matroidal, recursive interpretation.

For any 0-component $T = s$, the (unique) minimal working set and the (unique) critical set is $\{s\}$ and $\varnothing$ respectively, which are the (unique) base and the (unique) sub-base of free-matroid on the ground set $S = \{s\}$; this matroid is given the name of *0-level* k-UPM.

Let $T = \{t_1, \cdots, t_n\}$ be an l-component whose jth $(l-1)$-component t_j has a set of elements S_j, $j = 1, \cdots, n$, all these sets being pairwise disjoint. Then $S = \cup_{j=1}^{n} S_j$ is the set of elements of T. Assuming by inductive hypothesis that for each $j = 1, \cdots, n$, the minimal working sets and the critical sets for t_j are respectively the bases and the sub-bases of an $(l-1)$-level k-UPM M_j on the ground set S_j, it follows that the minimal working sets and the critical sets of T are respectively the bases and the sub-bases of $M = U[T, k] * (M_1, \cdots, M_n)$, $U[T, k]$ being the uniform matroid of rank k on the ground set T: M is called an *l-level k-UPM.*

Remark 5.1. A 1-level k-UPM is a uniform matroid of rank k; a 2-level k-UPM is an (ordinary) k-UPM, as defined in § 2.

6. Conclusions. The results of this paper may be subdivided into three parts. First, we have introduced and discussed, from a set-theoretic point of view, unlabelled partition systems, unlabelled partition matroids and the star operation. Second, we have investigated the computational complexity of combinatorial optimization problems involving these structures. Third, we have suggested a few applications of these ideas to problems in reliability theory and to some interesting variants of the well-known matching problems: the unlabelled degree constrained subgraph problems. In trying to draw the borderline between easy and hard problems of this kind, we have identified a problem on graphs which can be formulated in a very simple way, but whose complexity remains open to date: the exact partial covering by circuits.

Acknowledgments. The authors are grateful to S. Bellini and T. Brylawski for fruitful discussions.

REFERENCES

[1] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing Probability Models*, Holt, Rinehart and Winston, New York, 1975.

[2] T. Brylawski, *Constructions*, in Combinatorial Geometries, H. Crapo, G.-C. Rota and N. White, eds, Ch. 9, to appear.

[3] ———, *Iterated parallel union of matroids*, to appear.

[4] P. M. Camerini, G. Tartara and F. Maffioli, *Some scheduling algorithms for* SS-TDMA *systems*, Fifth International Conference on Digital Satellite Communications (IEEE Catalog No. 81CH1646-9), 1981, Genova, Italy.

[5] G. de Giuseppe and A. Scotti, *Metodi euristici per l'assegnamento ottimo tempo-capacità in sistemi* SS-TDMA, Thesis, Politecnico di Milano, Italy, 1983.

[6] J. Edmonds, *Matroids and the greedy algorithms*, Math. Programming, 1 (1971), pp. 127–136.

[7] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[8] M. Gondran and M. Minoux, *Graphes et algorithmes*, Eyrolles, Paris, 1979.

[9] R. Korte and D. Hausmann, *An analysis of the greedy heuristic for independence systems*, Ann. Discr. Math., 2 (1978), pp. 65–74.

[10] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.

[11] L. Lovász, *On determinants, matchings, and random algorithms*, Jozsef Attila Univ., Bolyai Institute, H-6720 Szeged, Hungary, Res. Rep., 1979.

[12] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[13] D. J. A. Welsh, *Matroid Theory*, Academic Press, London, 1976.

[14] N. Zadeh, *A simple alternative to the out-of-kilter algorithm*, Dept. Operations Research, Stanford Univ., Stanford, CA, 1979, Tech. Rep. no. 35.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

013

ON REDUCING THE SPACE REQUIREMENTS OF A STRAIGHT-LINE ALGORITHM*

DAVID A. CARLSON†

Abstract. The simultaneous time and space requirements of a straight-line algorithm can be determined by playing a well-known "pebble game" on a directed acyclic graph whose vertices and edges represent operations and argument assignments of the algorithm, respectively. In the game, pebble placements are made on vertices only when all predecessors have pebbles, time is the number of such placements made, and space is the number of pebbles used to reach all outputs of the graph. When space is restricted, extra time may be required to repebble some vertices, causing a tradeoff between time and space. Previous research has identified computations exhibiting different tradeoff characteristics, such as an extreme time-space tradeoff, in which a reduction in space causes time to rise from polynomial to superpolynomial in the size of the graph, and a favorable time-space tradeoff, in which a significant decrease in space can be achieved at the expense of a small (e.g. constant factor) increase in time. In this paper, we show that a computation is not limited to having only one tradeoff characteristic, but may exhibit both an extreme and a favorable time-space tradeoff. For three families of graphs, we derive upper and lower bounds on pebbling time for certain values of space that provide evidence of each family possessing both types of tradeoffs. We also provide upper and lower bounds on the maximum amount of time that can result from pebbling a graph when S pebbles are used.

Key words. pebble game, time-space tradeoff, straight-line algorithm, storage allocation, complexity theory

1. Introduction.

The pebble game, played on directed acyclic graphs, is a natural model for determining the simultaneous time and space requirements of a straight-line algorithm. Nodes of the graph correspond to operations in the algorithm, and edges denote dependencies between the operations. Temporary registers used by the algorithm are modeled by pebbles, which are moved from input nodes to output nodes according to three rules:

1. a pebble may be placed on an input node at any time,
2. a pebble may be removed from any node at any time, and
3. noninput nodes may be pebbled only when all their predecessors have pebbles.

Accordingly, the space requirement of the algorithm is reflected in the maximum number of pebbles on the graph at any point during the pebbling process, and time is measured by the number of pebble placements made on the graph (applications of rules 1 and 3). When the amount of available space is decreased, it may be necessary to "recompute" certain nodes of the graph due to the inability to hold pebbles on all intermediate computations, which implies the existence of a tradeoff between time and space.

The study of a straight-line algorithm's time-space tradeoff behavior has been the subject of a large amount of recent research. Since straight-line algorithms form the heart of many algebraic computational problems, it is natural to ask how much time is required to solve such a problem using a certain amount of space. Typically, if an n input problem (graph) is being solved using space S (pebbles), $\Omega(n)$ placements

[*] Received by the editors July 6, 1983. This work was supported in part by the National Science Foundation under grant NSFECS-04894. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

[†] Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, Massachusetts 01003.

must be made for each set of $O(S)$ outputs pebbled, due to the richness of dependencies between inputs and outputs in the graph associated with such a computation. Thus, a pebble game analysis implies $T=\Omega(n^2/S)$ for problems and algorithms such as the Fast Fourier transform [11], integer multiplication [12], and problems whose underlying representations possess concentration properties [14].

The pebble game has also been used to study the tradeoff between time and stack size in the evaluation of a linear recursive function. Paterson and Hewitt [7], Chandra [2], and Swamy and Savage [13] have obtained results for this problem. The most comprehensive are those of Swamy and Savage [13], where optimal pebbling strategies are derived for the graph associated with a linear recursive program. These pebbling strategies also indicate that for such graphs, space can be reduced significantly (e.g. from linear in graph size to the square root of graph size) at the expense of only a constant factor increase in time. In this paper, we refer to such behavior as a *favorable tradeoff.*

In general, any graph having size (number of nodes) N can be pebbled in linear time using N pebbles. Hopcroft, Paul, and Valiant [5] showed that any graph can be pebbled with $O(N/\log N)$ pebbles, so it is natural to ask what the consequences are when the space available to pebble a graph is reduced. Recent research has shown that certain computations possess an *extreme tradeoff,* for which a reduction in space results in superpolynomial pebbling time. Paul and Tarjan [8] exhibited the first extreme tradeoff by constructing a family of graphs for which a constant factor reduction in space results in time rising from polynomial to superpolynomial in graph size. Lengauer and Tarjan [6] extended the analysis of these graph families to show that superpolynomial time occurs when $S=\theta(N/\log N)$, which is the maximum possible reduction in space that can be achieved for a graph [9]. Van Emde Boas and van Leeuwen [3] have shown that superpolynomial time can result from the savings of a single pebble, and Carlson and Savage [1] proved that minimum space growing as any slowly increasing function of graph size is a necessary and sufficient condition for a graph family to possess an extreme tradeoff. Recently, Tompa [15] has shown that an algorithm for the natural problem of computing the transitive closure of a matrix has an extreme tradeoff by considering space-restricted implementations of the algorithm.

In this paper, we consider the general question of what are the consequences of reducing the space available when pebbling the outputs of a directed acyclic graph (equivalently, performing the computations of a straight-line algorithm). We show the existence of graph families that possess both an extreme and a favorable tradeoff; thus these two somewhat complementary behaviors can coexist in the same straight-line algorithm. This implies that it is difficult to know whether or not to attempt to decrease the space available to a specific straight-line algorithm: on one hand a significant decrease may be possible with the associated penalty in time being very small, while on the other hand the time penalty may be so large that such a decrease in space should not be attempted.

We also consider the maximum amount of time that can be associated with the pebbling of an arbitrary N node graph using S pebbles. We prove an upper bound on the pebbling time of the form $T\leqq(eN/S)^S$, and construct graph families, which for certain values of space S, require pebbling time close to this upper bound. This is analogous to previous research that has concentrated on the space requirements of arbitrary N node directed acyclic graphs—the results being that any such graph can be pebbled using space $O(N/\log N)$, and there exist graphs whose minimum space requirement is $\Omega(N/\log N)$ pebbles [5], [9].

2. Combining extreme and favorable tradeoffs. In this section, we present three different graph families that possess both an extreme and a favorable tradeoff. The building blocks for these graph families are the ladder graphs studied by Swamy and Savage [13], which can be defined as follows:

DEFINITION 1. The *ladder graph* l_N on N elements has vertex set $V=\{v_1, v_2, \cdots, v_n, w_1, w_2, \cdots, w_n\}$ and edge set

$$E=\{(v_i, v_{i+1}), (w_i, w_{i+1})|1\leqq i\leqq N-1\}U\{(v_i, w_{N+1-i})|1\leqq i\leqq N\}.$$

Vertex v_1 is the ladder graph's single input, and w_N is its single output.

The pebbling time of a ladder graph can be explicitly characterized by the following lemma:

LEMMA 1 (Swamy and Savage [13]). *The time required to pebble a ladder graph l_n on n elements when S pebbles are available satisfies*

$$T\cong\begin{cases} Sn^{1+1/S}(S/(1+S)), & S\ll \log n, \\ c_1 n \log n, & S=c_2 n \log n, \\ n \log n/\log S, & S\gg \log n. \end{cases}$$

From the above lemma, it is easily seen that ladder graphs (which form the basis for the evaluation of a linear recursive program, see Swamy and Savage [13]) possess a favorable tradeoff, since space can be reduced from linear in n to $n^\varepsilon, 0<\varepsilon<1$ a real constant, at the expense of a constant factor increase in time. Another graph family that can be seen to possess a favorable tradeoff is defined in Pippenger [10]. Pippenger's graphs require time $T=\theta(N \log(N/S))$ when S pebbles are available; thus S can be reduced from N^ε to $O(1)$, while time increases from $c_1N \log N$ to $c_2N \log N(c_1<c_2)$.

We also rely on the notion of a stack of superconcentrators, because it has been shown that such graphs require superpolynomial pebbling time when restricted amounts of space are used.

DEFINITION 2. An n-superconcentrator is a directed acyclic graph with n inputs and n outputs such that for any subset I of k inputs and 0 of k outputs, there exist vertex-disjoint paths joining I and 0.

DEFINITION 3. A stack of k n-superconcentrators $C(n, k)$ is formed by taking k n-superconcentrators C_i, $1\leqq i\leqq k$, and joining the outputs of C_i to the inputs of C_{i+1}, $1\leqq i\leqq k-1$.

Tompa [14] showed that an n-superconcentrator requires $T=\Omega(n^2/S)$ moves to pebble all outputs using S pebbles, and Valiant [16] constructed n-superconcentrators with size $O(n)$. The above lower bound can be iterated to show that a stack of superconcentrators requires $T=\Omega((n/S)^k)$ moves using S pebbles, and Lengauer and Tarjan [6] have improved this to $T=\Omega((N/S)^k)$. This iterative type of argument also forms the heart of Tompa's result [15] for a transitive closure algorithm based on successive matrix squaring operations.

Our first graph family that combines an extreme and a favorable tradeoff is formed by embedding a stack of superconcentrators into the nodes of a ladder graph.

DEFINITION 4. The graph $G(n, k)$ consists of $2k$ stacks of superconcentrators $C(n, \log n)$, where the graphs $C(n, \log n)$ are connected to one another as if each were a single node of a ladder graph. More formally, the outputs of $C_i(n, \log n)$ are connected to the inputs of $C_{i+1}(n \log n)$ for $1\leqq i\leqq k-1$, $k+1\leqq i\leqq 2k-1$ and the outputs of $C_i(n, \log n)$ are connected to the inputs of $C_{2k+1-i}(n, \log n)$ for $1\leqq i\leqq k$. Figure 1 shows the general composition of such a graph.

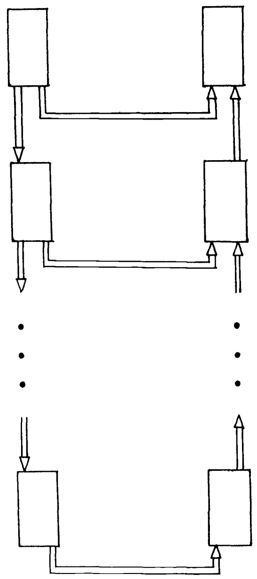


FIG. 1. $G(n, k)$.

Graphs that are similar to stacks of superconcentrators form the basis of algorithms for natural problems [15], thus it is not unrealistic that graphs resembling the ones defined above may be encountered in practice when the solution of a problem leads to a linear recursive program, which can be further decomposed into the above format. The following theorem shows that such a problem decomposition may possess both extreme and favorable tradeoff characteristics.

THEOREM 1. *For* $k = n^{c-1}$ *(c a positive integral constant), the graph family* $\{G(n, k)\}$ *combines an extreme tradeoff with a favorable tradeoff.*

Proof. $G(n, k)$ has size $N = \theta(nk \log n) = \theta(n^c \log n)$. We claim that superpolynomial time is required to pebble the graph when $S = O((N/\log N)^{(1/c)(1-\varepsilon)})$ (note that the graph can be pebbled using $S = O((N/\log N)^{(1/c)(1-\varepsilon)})$ pebbles since $S > S_{\min}$). This follows since by Lengauer and Tarjan [6], any subgraph $C(n, \log n)$ of $G(n, k)$ requires time $T = \Omega((n/S)^{\log n})$ when only S pebbles are available. Since $n = \theta((N/\log N)^{1/c})$, when $S = O((N/\log N)^{(1/c)(1-\varepsilon)})$ (ε a real constant between 0 and 1), we have $T = \Omega((N/\log N)^{(\varepsilon/c^2)\log N})$, which is superpolynomial in N. Thus $G(n, k)$ possesses an extreme tradeoff.

To see that $G(n, k)$ also has a favorable tradeoff, we first note that $G(n, k)$ can be pebbled in linear time with $S = \theta(nk)$ pebbles by pebbling subgraphs $C(n, \log n)$ in linear time ($\theta(n)$ pebbles are required for this) and by holding pebbles on their outputs as we advance through $G(n, k)$. To reduce the amount of space used significantly, we use groups of $\theta(n)$ pebbles each, and pebble the outputs of the subgraphs $C(n, \log n)$ according to the technique of Swamy and Savage [13]. This is done by treating a subgraph $C(n, \log n)$ in $G(n, k)$ as a single node in the ladder graphs analyzed in [13] which can be pebbled in $\theta(n \log n)$ moves using a group of $\theta(n)$ pebbles. Thus, $T = O(((k \log k)/\log t) \cdot (cn \log n))$ can be achieved when space $S = \theta(tn)$ and t grows much faster than $\log k$. It is easy to see that when $S = \theta(k^\gamma n)$ (γ a real constant between 0 and 1), $T = O(nk \log n)$. Thus, a significant decrease in space can be achieved at the expense of only a constant factor increase in time, i.e. $G(n, k)$ has a favorable tradeoff. Q.E.D.

The next graph family that we consider is constructed by connecting the outputs of ladder graphs to the inputs of a stack of superconcentrators. Again, since stacks of superconcentrators can be used to form algorithms for natural problems, the graphs we define here may be encountered in practice when the inputs of a stack of superconcentrators are the results of computations that are performed by a linear recursive program.

DEFINITION 5. The graph $H(n, k)$ consists of a single stack of superconcentrators $C(n, \log n)$ and n ladder graphs l_k where the output of the ith ladder graph is connected to the ith input of $C(n, \log n)$, $1 \leqq i \leqq n$. Figure 2 shows the construction of $H(n, k)$.

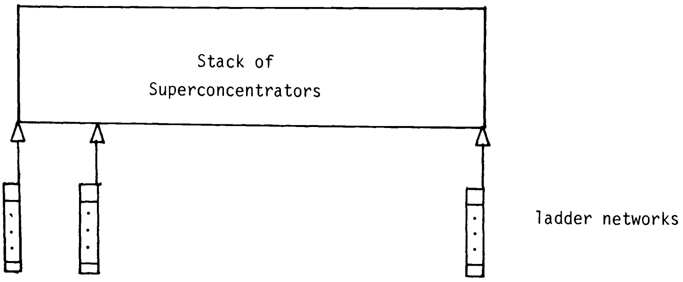


FIG. 2. $H(n, k)$.

THEOREM 2. *For $k = n^{c-1}$ ($c > 2$ and integral constant), the graph family $\{H(n, k)\}$ combines an extreme tradeoff with a favorable tradeoff.*

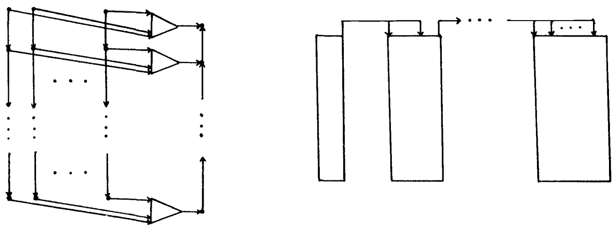
Proof. We use the same techniques as in Theorem 1, relying on the fact that the subgraph $C(n, \log n)$ requires superpolynomial pebbling time when space is suitably restricted, while the space required to pebble the linear recursive networks l_k connected to the inputs of $C(n, \log n)$ can be reduced significantly at the expense of a small increase in time. Specifically, when $S = O(n^{1-\varepsilon}) = O(N^{(1-\varepsilon)/c})$ ($0 < \varepsilon < 1$ a real constant), $T = \Omega\ ((n/S)^{\log n}) = \Omega(N^{(\varepsilon/c^2)\log N})$, which is superpolynomial in N. $H(n, k)$ can be pebbled in linear time using $k + n - 1$ pebbles, and it is easily seen using the techniques of Savage and Swamy [13] that space can be reduced to $S = \theta(N^{\gamma(c-1)/c})$ ($0 < \gamma < 1$ is a real constant and $k^\gamma > n$, i.e. $\gamma > 1/(c-1)$) while time $T = O(nk + n \log n) = O(N)$ increases by only a constant factor. Q.E.D.

It is interesting to note that for the graph family $\{H(n, k)\}$, the results of Swamy and Savage [13] would seem to imply that space can be reduced from $\theta(k+n)$ to $\theta(k^\gamma + n)$ for any $\gamma > 0$. However, the subgraph $C(n, \log n)$ causes time to jump to superpolynomial somewhere in the middle of this progressive reduction of space.

Next, we show that a favorable tradeoff can also be combined with an extreme tradeoff that occurs over a very small change in space in the sense of van Emde Boas and van Leeuwen [3]. The graphs we construct are based again on ladder graphs, but the ladder graphs are augmented so that they have a larger minimum space requirement.

DEFINITION 6. The graph $L(n, k)$ consists of $k-1$ subgraphs $L_{n,2}, \cdots, L_{n,k}$ each with a single output, where the output of $L_{n,i}$ is connected to all inputs of $L_{n,i+1}$ $2 \leqq i \leqq n-1$. The graph $L_{n,k}$ is composed of k spines, each having n nodes $\{s_{k,1}, \cdots, s_{k,n}\}$ and edges $\{(s_{k,i}, s_{k,i+1}) | 1 \leqq i \leqq n-1\}$. For $1 \leqq j \leqq k-1$, nodes $s_{j,i}$ are connected to inputs of a pyramid graph P_{k-1} (see Carlson and Savage [1] for a definition), the output of which is connected to node $s_{k,n+1-i}$ in spine k. Figure 3 outlines the construction of $L(n, k)$.

THEOREM 3. *For $k = O(n^c)$ ($c > 1$ an integral constant), the graph family $\{L(n, k)\}$ combines an extreme tradeoff with a favorable tradeoff.*



a) $L_{n,i}$ b) $L(n, k)$

FIG. 3

Proof. We first show the existence of an extreme tradeoff, relying on ideas from van Emde Boas and van Leeuwen [3] and Carlson and Savage [1]. $L(n, k)$ has been constructed so that it has a minimum space requirement k (k pebbles are necessary and sufficient to pebble its output). To pebble a node on the upward spine of the last subgraph $L_{n,k}$, first the output of subgraphs $L_{n,1}$ to $L_{n,k-1}$ must be pebbled and then k pebbles must be placed on $L_{n,k}$. It is possible to do this using only k pebbles, but in this case, there is a point in time before an advance is made on the spine of $L_{n,k}$ when $L_{n,1}$ to $L_{n,k-1}$ are devoid of pebbles (since the space requirement of reaching any node on the upward spine of $L_{n,k}$ is $k-1$). Thus, $L_{n,1}$ to $L_{n,k-1}$ must be pebbled again in their entirety using $k-1$ pebbles (one pebble is devoted to the upward spine of $L_{n,k}$). If $T(k, S)$ is the time required to pebble $L_{n,1}$ to $L_{n,k}$ using S pebbles, then we have

$$T(k, k) \geqq n \cdot T(k-1, k-1) > n^{k-1}.$$

The size of $L(n, k)$ is $N = O(nk^3)$, thus $T = \Omega((N/S^3)^S)$ with minimum space. This is superpolynomial in N as long as k is not $O(1)$. For example, if $k = n$, then $N = \theta(n^4)$ and $k = \theta(n^{1/4})$ so that $T = \Omega(N^{(1/4)N^{1/4}})$, which is obviously superpolynomial in N. Note that $L(n, k)$ can be pebbled in time $T = O(N^2)$ with $k+1$ pebbles by using the extra pebble to hold the value of the output of L_i as L_{i+1} is being pebbled, and that k can grow as any slowly increasing function of N with T remaining superpolynomial when minimum space is used (this is a simpler proof of the result contained in Carlson and Savage [1]).

$L(n, k)$ can be pebbled in linear time when $n(k-1)+k+1$ pebbles are available, since each subgraph $L_{n,i}$ can be pebbled in linear time by placing pebbles on all $n(i-1)$ nodes of downward spines and by using $i-1$ pebbles to pebble pyramids as a lone pebble advances on the upward spine. To pebble $L(n, k)$ with reduced space, the results of Swamy and Savage [13] are again used. Assuming that we have $\theta(n^\gamma(k-1)+k+1)$ pebbles available ($0<\gamma<1$ a real constant), it is easily seen that $L_{n,i}$ can be pebbled in time $O(ni+ni^2)$ (note that ni^2 term is due to the necessary pebbling of pyramids in L_i once). Thus, $L(n, k)$ can be pebbled in time $O(N)$. If $k = n^c$, c some small positive integral constant, then $N = \theta(n^{3c+1})$ and the above discussion implies that space can be reduced from $\theta(N^{(c+1)/(3c+1)})$ to $\theta(N^{(c+\gamma)/(3c+1)})$ while time remains $O(N)$. The smaller the constant c is, the more significant the reduction in space. If $k = \theta(\log n)$, then space goes from $\theta(N/(\log N)^2)$ to $\theta(N^\gamma/(\log N)^{3\gamma-1})$, which is an even more significant reduction. Q.E.D.

3. Upper and lower bounds on pebbling time. A large amount of research concerning time-space tradeoffs has focused on the question of what is the maximum amount of space reduction that can be made when pebbling an arbitrary directed acyclic graph, and what is the penalty in time associated with such a reduction in space.

It has been determined that space can be reduced to $\theta(N/\log N)$ (N denoting graph size) [5], [9], and that superpolynomial time may be associated with such a reduction [6]. Another similar question has concerned the cutoff point for superpolynomial time: What is the value of space above which any graph can be pebbled in polynomial time. Lengauer and Tarjan [6] also answered this question by showing that any graph can be pebbled in polynomial time with space $S \geqq c_1 N/\log\log N$, while there exist graphs that require superpolynomial time when $S \leqq c_2 N/\log\log N$ ($c_2 < c_1$).

Here, we consider an analogous question regarding the pebbling time of an arbitrary directed acyclic graph, which simply stated is: what is the maximum pebbling time of a general graph when S pebbles are available. We provide an upper bound on the maximum pebbling time of a directed acyclic graph, and construct graph families that come close to meeting this upper bound when pebbled with certain values of space. The upper bound is stated in the following theorem.

THEOREM 4. *When S pebbles are available, an arbitrary graph G with N nodes can be pebbled in* $T \leqq N(eN/S)^S$ *moves.*

Proof. It is easily seen that the total number of moves made on the graph to reach any output is less than or equal to the total number of configurations of $\leqq S$ pebbles on G, which is $\sum_{0 \leqq j \leqq S} \binom{N}{j}$. Thus, $T \leqq N \cdot \sum_{0 \leqq j \leqq S} \binom{N}{j}$, so we desire an upper bound on the sum of binomial coefficients. From the Chernoff bound (see [4]), it can be deduced that when $S < \varepsilon N$,

$$\sum_{0 \leqq j \leqq S} \binom{N}{j} \varepsilon^j (1-\varepsilon)^j = \text{Prob}\,[X < S]$$

$$\leqq \exp\{NH(S/N) + (S/N)\log_e \varepsilon + (1 - S/N)\log_e (1-\varepsilon)\}$$

where X is a random variable having a binomial distribution with parameter ε and H is the entry function $H(x) = -x\log_e x - (1-x)\log_e(1-x)$. Choosing $\varepsilon = 1/2$ yields

$$\sum_{0 \leqq j \leqq S} \binom{N}{j} < \exp\{NH(S/N)\} \quad \text{for } S < N/2.$$

From the definition of the entropy function,

$$\begin{aligned}\exp\{NH(S/N)\} &= \exp\{N[(S/N)\log_e(N/S) + (1-S/N)\log_e(1-S/N)^{-1}]\}\\ &= \exp\left\{S\log_e(N/S) + (N-S)\log_e\left(\frac{N}{N-S}\right)\right\}\\ &= (N/S)^S \cdot \left(\frac{N}{N-S}\right)^{N-S},\end{aligned}$$

Now,

$$\left(\frac{N}{N-S}\right)^{N-S} = \left(\frac{N-S+S}{N-S}\right)^{N-S} = \left(1 + \frac{S}{N-S}\right)^{N-S}.$$

Since the exponential function is defined as $e^x = \lim_{n->\infty}(1+x/n)^n$, and the defined sequence is increasing, it is obvious that $(1+S/(N-S))^{N-S} \leqq e^S$. Thus

$$\sum_{0 \leqq j \leqq S} \binom{N}{j} < (eN/S)^S. \qquad \text{Q.E.D.}$$

We now concentrate on constructing graph families that come close to meeting the bounds of Theorem 4.

THEOREM 5. *There exist graph families that, for certain values of space S, require* $T = \Omega((N/S)^{S/f(N)})$ *moves to pebble a member of the family having size N. Here,* $f(N) = \omega(1)$ *is any slowly increasing function in N.*

Proof. In Lengauer and Tarjan [6], graphs that are stacks of superconcentrators were analyzed to determine their time-space tradeoff behavior. We have seen earlier that a single superconcentrator with n inputs and outputs requires $T=\Omega(n^2/S)$ moves to pebble all outputs using S pebbles [14]. Iterating this argument shows that for a stack of k superconcentrators (outputs of the $(i-1)$st superconcentrator are connected to inputs of the ith), $T=\Omega((n/S)^k)$ moves are required when S pebbles are available. Lengauer and Tarjan [6] improved the lower bound on time to $T=\Omega((nk/S)^k)$, and since superconcentrators of size $O(n)$ exist [16], $T=\Omega((N/S)^k)$. As noted by Lengauer and Tarjan, this implies that graph families exist for which superpolynomial time is required when space $\theta(N/\log N)$ is used. Such graph families are constructed by choosing $k=\log n$.

Here, we analyze the time required to pebble stacks of superconcentrators with different values of k, with the intent of realizing the largest amount of time required for certain values of space. It is easily seen that the minimum space requirement of a stack of k superconcentrators, each having n inputs and outputs, is $k\log n$. For such a stack, $T=\Omega((N/S)^k)$, but since space S is at least $k\log n$, $k\leqq S/\log n$. Thus, we cannot match the upper bound of Theorem 4 exactly, but can come close by pebbling the stack with its minimum space requirement. Pebbling the stack with more available space implies that the match with the bound of Theorem 4 is not as close as with minimum space.

We now consider how close we can come to the bound $T\leqq N(eN/S)^S$. First, consider Lengauer and Tarjan's stack of n-superconcentrators with $k=\log n$ and $T=\Omega((N/S)^k)$. When such a graph is pebbled with minimum space $(\log n)^2$, $T=\Omega((N/S)^{S/\log N})$, since $k=\theta(\log N)$ and $S=\theta((\log N)^2)$. When space $S=N/\log N$ is used, $T=\Omega((N/S)^{\log N})=\Omega((N/S)^{S/(N/(\log N)^2)})$ which is superpolynomial in N, but is not close at all to $T\leqq N(eN/S)^S$. Now, consider a stack of $k=2^n$ n-superconcentrators. The minimum space requirement of such a graph is $2^n\log n$, its size is $N=n2^n$, which implies that $\log\log N=\theta(\log n)$ and $k=S_{\min}/\log\log N$, so that when the graph is pebbled with minimum space, $T=\Omega((N/S)^{S/\log\log N})$ moves are required. Clearly, this argument can be extended to show the existence of a graph family requiring $T=\Omega((N/S)^{S/f(N)})$ moves when minimum space is used. Here, $f(N)=\log\log\cdots\log N$, and since any slowly increasing function in N grows at least as fast as some such $f(N)$, $T=\Omega((N/S)^{S/\omega(1)})$. Q.E.D.

4. Conclusions. The pebble game is a useful tool for studying the simultaneous time and space requirements of a straight-line algorithm, and has been used by previous authors to investigate the various types of tradeoffs associated with different straight-line computations. In this paper, we have shown that a favorable tradeoff (large reduction in space — constant factor increase in time) and an extreme tradeoff (reduction in space — time increase to superpolynomial) can be combined in the same straight-line algorithm. Thus, it is unclear whether a decrease in space that results in a small time penalty should be followed by further decreases in space. We have also studied the maximum pebbling time associated with a pebble game analysis of a straight-line algorithm, and have provided close (but not matching) upper and lower bounds on pebbling time when a certain amount of space is used.

REFERENCES

[1] D. A. Carlson and J. E. Savage, *Extreme time-space tradeoffs for graphs with small space requirements*, Inform. Proc. Lett., 14 (1982), pp. 223–227.

[2] A. K. Chandra, *Efficient compilation of linear recursive programs*, Proc. 14th Symposium on Switching and Automata Theory, October 1973, pp. 16–25.

[3] P. van Emde Boas and J. van Leeuwen, *Move rules and tradeoffs in the pebble game*, in Lecture Notes in Computer Science 67, G. Goos and J. Hartmanis, eds., Springer-Verlag, New York, 1979, pp. 101–112.

[4] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley, New York, 1966.

[5] J. E. Hopcroft, W. J. Paul, and L. G. Valiant, *On time versus space*, J. Assoc. Comput. Mach., 24 (1977), pp. 332–337.

[6] T. Lengauer and R. E. Tarjan, *Upper and lower bounds on time-space tradeoffs*, J. Assoc. Comput. Mach., 29 (1982), pp. 1087–1130.

[7] M. S. Paterson and C. E. Hewitt, *Comparative schematology*, Proc. MAC Conference on Concurrent Systems and Parallel Computation, June, 1970, pp. 119–127.

[8] W. J. Paul and R. E. Tarjan, *Time-space tradeoffs in a pebble game*, Acta Informatica, 10 (1978), pp. 111–115.

[9] W. J. Paul, R. E. Tarjan and J. R. Celoni, *Space bounds for a game on graphs*, Math System Theory, 10 (1977), pp. 239–251.

[10] N. Pippenger, *A time-space tradeoff*, J. Assoc. Comput. Mach., 25 (1978), pp. 509–515.

[11] J. E. Savage and S. Swamy, *Space-time tradeoffs on the* FFT *algorithm*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 563–568.

[12] ———, *Space-time tradeoffs for oblivious sorting and integer multiplication*, Lecture Notes in Computer Science 71, H. A. Mauer ed., Springer-Verlag, New York, 1979.

[13] S. Swamy and J. E. Savage, *Space-time tradeoffs for linear recursion*, Math System Theory, 16 (1983), pp. 9–27.

[14] M. Tompa, *Time-space tradeoffs for computing functions, using connectivity properties of their circuits*, J. Comput. System. Sci., 20 (1980), pp. 118–132.

[15] ———, *Two familiar transitive closure algorithms which admit no polynomial time, sublinear space implementations*, SIAM J. Comput., 11 (1982), pp. 130–137.

[16] L. G. Valiant, *On non-linear lower bounds in computational complexity*, Proc. 7th ACM Symposium on Theory of Computing, May, 1975, pp. 45–53.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 3, September 1984

014

COMBINATORIAL ASPECTS OF THE ORRERY MODEL OF SYNTAX*

MICHAEL O. ALBERTSON† AND DAVID M. BERMAN‡

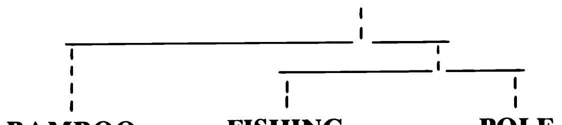
Abstract. The orrery model of syntax offers an approach to language acquisition which requires minimal assumptions. This paper counts the number of orreries on n words, and proposes a weighted generalization which might allow quantitative prediction.

AMS(MOS) subject classifications. Primary 68F99; secondary 05C05

Introduction. Linguists have long acknowledged that the syntax of human language cannot be understood using a model which incorporates only word order and bracketing. Traditional responses to this have been to inject directly into the syntax model aspects of the observed complexity of language. This results in syntactic structures which are difficult to construct and unwieldy to apply. In contrast, Moulton and Robinson [4], [5] propose a theory of language in which syntax need only encode scope and dependency. This paper is a combinatorial investigation of the Moulton–Robinson model of syntax. Our results show that their syntax model is richer than might be expected in its ability to draw distinctions. We also propose a model that might prove useful in establishing a distance between syntactic structures as well as between elements in a given syntactic structure.

1. Linguistic definitions. Two words which are directly related are said to be in each other's scope. The essential features of the scope relationship are that it is symmetric with respect to the two words and that it can be applied recursively i.e. two related words can be formally considered as a unit and this unit may be in the scope of a third word or another compound unit. The notion of dependence is that in any pair of words in each other's scope one is considered the main component and the other the dependent component.

The relations of scope and dependency can be modelled by an orrery, or planetarium. Two words in each other's scope can be linked by a horizontal bracket hanging from a vertical strut. Dependency is indicated by placing the strut closer to the main component than to the dependent component. This supplies a visual suggestion that the main component is heavier. The bracketed pair may then be linked via the strut to another word or compound unit. For example the language string **BAMBOO FISHING POLE** can be represented by the following orrery:



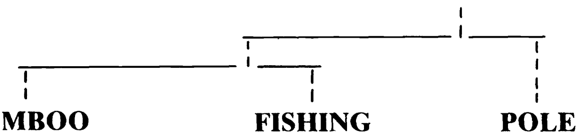
indicating that **FISHING** and **POLE** are in each other's scope and that **POLE** is the main component. **BAMBOO** is then in the scope of **FISHING POLE** and is the dependent component.

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Mathematics, Smith College, Northampton, Massachusetts 01063.

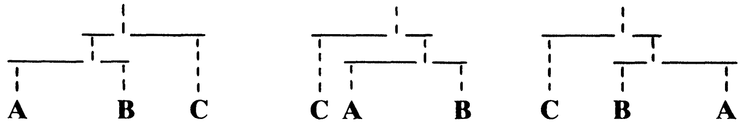
‡ Department of Mathematics, University of New Orleans, New Orleans, Louisiana 70148.

If we meant instead to refer to a pole used in fishing for bamboo, the orrery would be:



Here **BAMBOO** is in the scope of, and dependent on **FISHING**; **BAMBOO FISHING** is then in the scope of and dependent on **POLE**.

It is only for convenience that these orreries are drawn with the words in their natural order. Like the orreries that model the solar system—or like Calder mobiles—they should be thought of as free to pivot about each vertical strut. Thus the following are all representations of the same orrery:



etc.

2. Are there enough orreries? There can be no doubt that the orrery is simpler than the base syntactic structure either for transformation theory or for case theory. For the orrery model of syntax to be a viable alternative it is necessary that there be enough orreries to distinguish fundamentally distinct language strings. Let $f(n)$ denote the number of orreries on n words. It is straightforward to see that $f(1)=1$, $f(2)=2$, and $f(3)=12$. With a little patience one can exhaustively list the 120 orreries on 4 words.

THEOREM. $f(n)=(2n-2)!/(n-1)!$.

At the moment we know of four proofs of the above theorem. One can obtain an inductive argument based on the recurrence [1]

$$f(n)=(4n-6)f(n-1).$$

If $C(n,j)$ denotes the number of ways of selecting j objects from a set of n objects, then standard techniques produce:

$$f(n)=\sum_{j=1}^{n-1} C(n,j)f(j)f(n-j).$$

Given this recurrence, exponential generating functions can be used to obtain the formula for $f(n)$. A direct count for $f(n)/n!$ appears in Even [2]. The argument we will present (due to Jim Henle [3]) is to our knowledge the simplest method which directly counts these Catalan style numbers. It also produces an alternative geometric model of the orrery. We begin by constructing and then counting circular schemes. The second step will be to exhibit a one-to-one onto map, from the orreries to the circular schemes.

Constructing the circular schemes: Given n characters $A, B, C, \cdots$ (which represent the words), we construct a circular scheme by arranging these characters around the circumference of a circle. Next we insert $(n-1)$ ^'s on the circumference. Our only constraint is that no two of these $(2n-1)$ symbols should be at the same point. Finally we equally space the symbols around the circumference. Let $g(n)$ denote the number of distinct circular schemes on n characters. Two circular schemes will be considered to be the same if there is a rotation of the first which is identical to the second. Clearly we may ease the counting burden by forcing the character A to appear at "noon". The remaining $(2n-2)$ symbols may be arranged in $(2n-2)!$ different ways. The ^'s are indistinguishable. Hence $(n-1)!$ of these objects are the same. Thus

$$g(n)=(2n-2)!/(n-1)!.$$

The bijection: We first give an algorithm which will produce a circular scheme given an orrery. Suppose A and B are two words or compound units which are in each other's scope and that B is dependent on A. Replace the two separate words with the unit $\{AB\}$ which will henceforth be considered as one character. Thus we obtain an orrery on $(n-1)$ "words". Place (^**AB**) on the circumference of the circle in clockwise order. We now take a pair of characters in each other's scope in the reduced orrery and repeat the process. If at least one of the two characters is not yet on the circle we add the new symbols so that they are not within any pair of parentheses. (The parentheses serve only this bookkeeping function and are erased once all the characters are on the circle.) If both characters are already on the circle we need only add ^ and a pair of parentheses. We illustrate this step in Fig. 1.

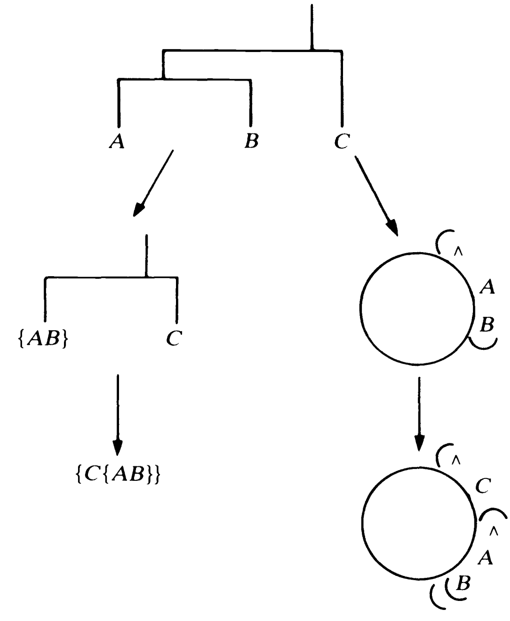


FIG. 1. *From the orrery to the circular scheme.*

Next we produce an orrery from a circular scheme. Since the circular scheme has n characters and only $(n-1)$ ^'s it must be the case that two of the characters are adjacent. Further there must be two such characters which are immediately preceded by a ^. Suppose some segment of the circumference has ^**QR** in clockwise order. Set **Q** and **R** to be in each other's scope with **R** dependent on **Q**. Replace ^**QR** by {**QR**} obtaining a circular scheme with $(n-1)$ characters and $(n-2)$ ^'s. Proceed inductively. We illustrate this in Fig. 2.

3. Combinatorial perspectives. Combinatorially it is more natural to consider scope as a relationship possessed by an entire language string rather than as a relationship between pairs of words or compound units. Thus we can also model the scope relationship for an entire string as a rooted binary tree. The words are identified with the leaves of the tree (i.e. the vertices of degree one). Two words or compound units are in each others scope precisely if the unique path joining them is of length two.

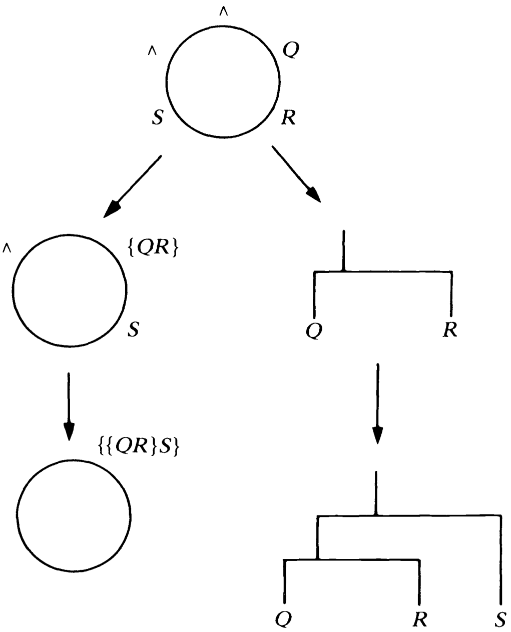
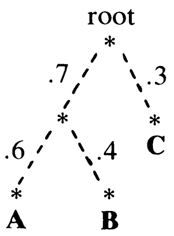


FIG. 2. *From the circular scheme to the orrery.*

There is the natural generalization of allowing the scope of a language string to be modelled by a rooted tree which is not necessarily binary. Moulton and Robinson recognized this possibility, but their model does not appear to require this much generality. At the moment we see no particular mathematical disadvantage in restricting our attention to binary rooted trees.

The combinatorial view of dependency is of a labelling of the edges of the "scope" tree with the designations heavy or light such that the two edges descendant from any vertex are labelled differently. The mathematical generalization here is an assignment to each edge a weight, a number chosen from the interval (0, 1), subject to the condition that the weights assigned to the two edges descendent from a vertex must sum to 1. The weight of a word is then defined to be the product of the weights of the edges in the unique path joining the word with the root. We illustrate below:



Here the weight of A is .42, the weight of B is .28, and the weight of C is .3. Note that the weights of the words sum to 1. This is no accident for we may view the weights of each edge as probabilities and the words as independent outcomes which exhaust the possibilities. It is linguistically reasonable to consider the weights of the words as a measure of their relative importance in the language string. One can recover the weights on the edges of the scope tree from the relative weights of the words. Specifically, given a language string, a scope tree with a word assigned to each leaf, and the relative weights of the words in the string; there is a unique assignment of weights to the edges of the scope tree which will produce the weights of the words. We indicate how to construct this assignment. Begin with two words (or compound units) say **A** and **B** in each other's scope. If the weights of **A** and **B** are a and b respectively, form a smaller tree (and language string) by deleting **A** and **B**, replacing them with the word {**AB**}, and assigning this new unit to the common ancestor of **A** and **B**. The word {**AB**} will be assigned the weight $a+b$. Find the weights on the edges in the smaller tree. In the original tree the weight of the edge terminating at the word **A** will be $a/(a+b)$ and the weight of the edge terminating at the word **B** will be $b/(a+b)$.

The utility of this construction is that while mathematically it is more convenient to examine the weights on the edges, these would seem to be difficult to determine empirically. However, linguists have constructed experiments to measure the relative importance of words in a language string [4].

These weighted binary trees offer the possibility of illuminating two linguistically natural questions. The first of these is the distance between language strings. Given two weighted binary trees T and T', let $\langle T, T'\rangle$ denote the minimal binary tree which contains both T and T'. It is important to note that we mean that there are isomorphisms of T and T' into $\langle T, T'\rangle$. Define a weighted binary tree S by adding to T those edges of $\langle T, T'\rangle$ not already in T and assigning weights of 0 and 1 to the new edges to reflect the dependency in T'. Similarly define S'. Note that S and S' are both weighted versions of the tree $\langle T, T'\rangle$. If $w(e)$ and $w'(e)$ denote the weights of the edge e of depth j in the trees S and S' respectively, define

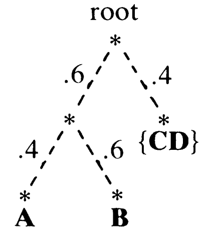
$$z(e) = abs(w(e) - w'(e))/2^j.$$

Attach the weight $z(e)$ to e in $\langle T, T'\rangle$. The distance between two weighted binary trees is defined by

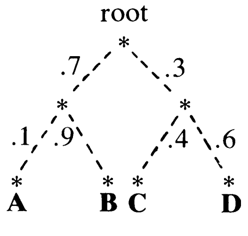
$$d(T, T') = \sum_{e \text{ in} \langle T, T'\rangle} z(e).$$

We illustrate with the example below:

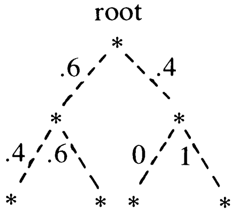
T



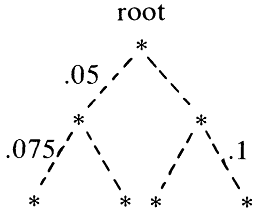
T'



S



$\langle T, T'\rangle$



In this example the weighted tree S' is identical to the tree T' so it is omitted from the above diagram. Note that if e and f are common descendant edges of the same vertex then $z(e)=z(f)$. In the example above $d(T, T')=.45$. In general the distance between two weighted scope trees will be less than the larger of the depths. It is straightforward to check that d is a metric on weighted binary trees. There are of course many different such metrics. We chose this particular one because it emphasizes differences in the weighting which are closer to the root.

Weighted scope trees also might help in the investigation of the distance between words within a given language string. There are at least two potential pitfalls involved in constructing such a definition. First, empirical evidence seems to indicate that the distance between words is an asymmetric quantity [4]. Generally within a given pair of words, say **A** and **B**, if **A** is given more weight than **B** in the weighted scope tree, then **B** will be closer to **A** than **A** is to **B**. Thus the distance function will not be a metric. Second, the question arises as to whether the distance from **A** to **B** depends only on that portion of the weighted scope tree immediately above **A** and **B** or on the entire tree, i.e., distance may not even be a local property.

At this stage, the direction of investigation should be determined by linguistic utility rather than mathematical esthetic.

REFERENCES

[1] R. BRUALDI, *Introductory Combinatorics*, North-Holland, New York, 1977.

[2] S. EVEN, *Graph Algorithms*, Computer Science Press, Rockville, MD, 1979.

[3] J. HENLE, personal communication.

[4] J. MOULTON AND G. ROBINSON, *The Organization of Language*, Cambridge Univ. Press, New York, 1981.

[5] ———, *An empiricist theory of language acquisition*, Behavioral and Brain Science, (invited).

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

MULTIPLICATION OF GENERALIZED POLYNOMIALS, WITH APPLICATIONS TO CLASSICAL ORTHOGONAL POLYNOMIALS*

STEPHEN BARNETT†

Abstract. A simple scheme is presented for computing the product of two polynomials in generalized form, i.e. expressed relative to a given orthogonal polynomial basis. If the polynomials have degrees m and r ($r \leqq m$), then the method requires r multiplications of vectors by a tridiagonal matrix of order $m+r$. No conversions to standard power form are involved. As particular cases, some explicit formulae are easily derived for products of pairs of classical orthogonal polynomials.

AMS(MOS) subject classifications. Primary: 12B05, secondary: 15A99, 42C05

1. Introduction. Take as a basis the orthogonal polynomials $p_i(\lambda)$ defined by the standard relationships

$$p_0(\lambda)=1, \qquad p_1(\lambda)=\alpha_1\lambda+\beta_1, \tag{1.1}$$

$$p_i(\lambda)=(\alpha_i\lambda+\beta_i)p_{i-1}(\lambda)-\gamma_ip_{i-2}(\lambda), \qquad i=2,3,\cdots \tag{1.2}$$

with $\alpha_i>0$, $\gamma_i>0$. Let $a(\lambda)$ be an nth degree generalized polynomial [7] expressed in the form

$$a(\lambda)=p_n(\lambda)+a_1p_{n-1}(\lambda)+\cdots+a_np_0(\lambda). \tag{1.3}$$

Many of the properties of $a(\lambda)$ can be investigated using the comrade matrix [5]:

$$A=\begin{bmatrix} \frac{-\beta_1}{\alpha_1} & \frac{1}{\alpha_1} & 0 & 0 & & \\ \frac{\gamma_2}{\alpha_2} & \frac{-\beta_2}{\alpha_2} & \frac{1}{\alpha_2} & 0 & 0 & \\ 0 & \frac{\gamma_3}{\alpha_3} & \frac{-\beta_3}{\alpha_3} & \frac{1}{\alpha_3} & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \frac{\gamma_{n-1}}{\alpha_{n-1}} & \frac{-\beta_{n-1}}{\alpha_{n-1}} & \frac{1}{\alpha_{n-1}} \\ \frac{-a_n}{\alpha_n} & \cdots & & \frac{-a_3}{\alpha_n} & \frac{-a_2+\gamma_n}{\alpha_n} & \frac{-a_1-\beta_n}{\alpha_n} \end{bmatrix}. \tag{1.4}$$

It was shown in [4] that

$$\det(\lambda I_n-A)=a(\lambda)/(\alpha_1\alpha_2\cdots\alpha_n) \tag{1.5}$$

where I_n denotes the unit matrix of order n. A crucial step in previous work [5] has been to construct the matrix $b(A)$, where $b(\lambda)$ is a generalized polynomial:

$$b(\lambda)=p_m(\lambda)+b_1p_{m-1}(\lambda)+\cdots+b_mp_0(\lambda). \tag{1.6}$$

A very similar construction will be seen to be a key step in this paper. It can be assumed without loss of generality that $m<n$. It was shown in [4] that if the rows of

* Received by the editors February 1, 1983, and in revised form June 16, 1983.

† School of Mathematical Sciences, University of Bradford, Bradford, West Yorkshire, England.

$b(A)$ are denoted by $\rho_1, \rho_2, \cdots, \rho_n$, then

$$\rho_1=[b_m, b_{m-1}, \cdots, b_1, 1, 0, \cdots, 0] \tag{1.7}$$

and subsequent rows satisfy the same recurrence formula as does the basis:

$$\rho_i=\rho_{i-1}(\alpha_{i-1}A+\beta_{i-1}I_n)-\gamma_{i-1}\rho_{i-2}, \qquad i=2, 3, \cdots, n \tag{1.8}$$

with $\rho_0=0$. Previous work has included computation of the greatest common divisor $d(\lambda)$ of $a(\lambda)$ and $b(\lambda)$ [4]. A recent extension [6] permits, in addition, simultaneous determination of $a(\lambda)/d(\lambda)$, and the quotient and remainder on division of $a(\lambda)$ by $b(\lambda)$. Throughout, all computations are performed relative to the given basis, and no conversion of polynomials to standard power form is required. A further extension is given in this paper, enabling the product of $b(\lambda)$, and another generalized polynomial $c(\lambda)$ of degree $r\leqq m$, to be determined. The procedure is presented in Theorem 2 in § 2, being a direct consequence of a result of [6], reproduced below as Theorem 1. A numerical example illustrates the simplicity of the algorithm, which involves the computation of r vectors using a formula like (1.8), but with A replaced by a tridiagonal matrix A_0 obtained by setting $a_i=0$, $\forall i$, in (1.4). The problem of determining the product $p_m(\lambda)p_r(\lambda)$ is considered in § 3 for some special cases where explicit expressions can be obtained. These include formulae for Chebyshev, Hermite and Legendre polynomials.

2. The multiplication procedure. Suppose that in (1.3) the degree of $a(\lambda)$ is equal to $m+r$, where $r\leqq m$, and write the division of $a(\lambda)$ by $b(\lambda)$ in (1.6) in the form

$$a(\lambda)=\mu(m+r, m)[b(\lambda)c(\lambda)+f(\lambda)] \tag{2.1}$$

where

$$c(\lambda)=p_r(\lambda)+c_1p_{r-1}(\lambda)+\cdots+c_rp_0(\lambda), \tag{2.2}$$

$$f(\lambda)=f_0p_{m-1}(\lambda)+f_1p_{m-2}(\lambda)+\cdots+f_{m-1}p_0(\lambda), \tag{2.3}$$

$$\begin{aligned}\mu(m+r, m)&=(\alpha_1\alpha_2\cdots\alpha_{m+r})/(\alpha_1\alpha_2\cdots\alpha_m)(\alpha_1\alpha_2\cdots\alpha_r)\\&=(\alpha_{m+1}\alpha_{m+2}\cdots\alpha_{m+r})/(\alpha_1\alpha_2\cdots\alpha_r).\end{aligned} \tag{2.4}$$

The main result in [5] is conveniently expressed for present purposes as follows:

THEOREM 1. *The rows* $\rho_1, \cdots, \rho_{r+1}$ *of*

$$b(A)=p_m(A)+b_1p_{m-1}(A)+\cdots+b_mp_0(A)$$

where A is the comrade matrix (1.4) *of order $m+r$, satisfy the relationship*

$$\rho_{r+1}+c_1\rho_r+\cdots+c_r\rho_1+[f_{m-1}, f_{m-2}, \cdots, f_1, f_0, 0, \cdots, 0]=0. \tag{2.5}$$

Suppose now that $b(\lambda)$ and $c(\lambda)$ are *given* polynomials in the forms (1.6) and (2.2) respectively, and in (2.1) set $f(\lambda)$ equal to zero. The problem is then to determine $a(\lambda)$, i.e., the product $b(\lambda)c(\lambda)$, and this is given by

THEOREM 2. *Let A_0 denote the tridiagonal matrix of order $m+r$ obtained by setting $a_i=0$, $\forall i$, in* (1.4). *The rows $R_1, R_2, \cdots, R_{r+1}$ of $b(A_0)$ satisfy*

$$R_1=[b_m, b_{m-1}, \cdots, b_1, 1, 0, \cdots, 0], \tag{2.6}$$

$$R_i=R_{i-1}(\alpha_{i-1}A_0+\beta_{i-1}I_{m+r})-\gamma_{i-1}R_{i-2}, \qquad i=2, 3, \cdots, r+1 \tag{2.7}$$

with $R_0 = 0$, *and the product of* (1.6) *and* (2.2) *is*

$$b(\lambda)c(\lambda) = \frac{\alpha_1 \cdots \alpha_r}{\alpha_{m+1} \cdots \alpha_{m+r}} p_{m+r}(\lambda) + x_1 p_{m+r-1}(\lambda) + \cdots + x_{m+r} p_0(\lambda) \tag{2.8}$$

where

$$[x_{m+r}, \cdots, x_2, x_1] = R_{r+1} + c_1 R_r + \cdots + c_r R_1. \tag{2.9}$$

Proof. Recall that in (1.4) we have set $n = m + r$. Since A_0 is still in comrade form, it follows that R_1 is identical to ρ_1 in (1.7), and that the rows of $b(A_0)$ must satisfy a relationship (2.7) having the same form as (1.8).

Furthermore, it follows from (1.8) that the ith row ρ_i of $b(A)$ has its last $r-i$ elements equal to zero, and that ρ_{r+1} is the first row of $b(A)$ to involve the coefficients a_i in A, so that $\rho_i \equiv R_i$, $i = 2, 3, \cdots, r$.

By considering (1.7) and (1.8), it is easy to verify that the last element of ρ_r is

$$\delta = (\alpha_1 \alpha_2 \cdots \alpha_{r-1})/(\alpha_{m+1}\alpha_{m+2} \cdots \alpha_{m+r-1}). \tag{2.10}$$

Hence it follows from (1.8), and (2.7) with $i = r+1$, that

$$\begin{aligned} \rho_{r+1} &= R_{r+1} + \alpha_r R_r (A - A_0) \\ &= R_{r+1} + \alpha_r \delta \left[\frac{-a_{m+r}}{\alpha_{m+r}}, \cdots, \frac{-a_2}{\alpha_{m+r}}, \frac{-a_1}{\alpha_{m+r}} \right]. \end{aligned} \tag{2.11}$$

Since $f(\lambda) \equiv 0$, (2.5) becomes

$$\rho_{r+1} + c_1 \rho_r + \cdots + c_r \rho_1 = 0$$

and substituting from (2.11) and (2.4) produces

$$R_{r+1} + c_1 R_r + \cdots + c_r R_1 = [a_{m+r}, \cdots, a_1]/\mu(m+r, m). \tag{2.12}$$

In (2.1) we now have

$$\begin{aligned} b(\lambda)c(\lambda) &= a(\lambda)/\mu(m+r, m) \\ &= [p_{m+r}(\lambda) + a_1 p_{m+r-1}(\lambda) + \cdots + a_{m+r} p_0(\lambda)]/\mu(m+r, m), \end{aligned}$$

and this reduces to the required expression (2.8) on setting $x_i = a_i/\mu(m+r, m)$, whereby (2.12) becomes (2.9). □

Example. Choose as the basis the Legendre polynomials $P_i(\lambda)$, for which

$$\alpha_i = \frac{2i-1}{i}, \quad \beta_i = 0, \quad \gamma_i = \frac{i-1}{i}, \quad i \geqq 1. \tag{2.13}$$

Suppose that

$$\begin{aligned} b(\lambda) &= P_3(\lambda) + 2P_2(\lambda) + P_1(\lambda) - P_0(\lambda), \\ c(\lambda) &= P_2(\lambda) - 2P_1(\lambda) + 3P_0(\lambda). \end{aligned}$$

Here $m = 3$, $r = 2$, and from (1.4) and (2.13)

$$A_0 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{5} & 0 & \frac{3}{5} & 0 \\ 0 & 0 & \frac{3}{7} & 0 & \frac{4}{7} \\ 0 & 0 & 0 & \frac{4}{9} & 0 \end{bmatrix}.$$

Equations (2.6) and (2.7) give

$$R_1=[-1,1,2,1,0],$$

$$R_2=R_1A_0=[\tfrac{1}{3},-\tfrac{1}{5},\tfrac{23}{21},\tfrac{6}{5},\tfrac{4}{7}],$$

$$R_3=\tfrac{3}{2}R_2A_0-\tfrac{1}{2}R_1=[\tfrac{2}{5},\tfrac{23}{35},-\tfrac{3}{7},\tfrac{13}{15},\tfrac{36}{35}]$$

and from (2.9)

$$\begin{aligned}[x_5,x_4,x_3,x_2,x_1]&=R_3-2R_2+3R_1\\&=[-\tfrac{49}{15},\tfrac{142}{35},\tfrac{71}{21},\tfrac{22}{15},-\tfrac{4}{35}].\end{aligned}$$

The leading coefficient in (2.8) is $\alpha_1\alpha_2/\alpha_4\alpha_5=10/21$, so Theorem 2 gives

$$b(\lambda)c(\lambda)=\tfrac{10}{21}P_5(\lambda)-\tfrac{4}{35}P_4(\lambda)+\tfrac{22}{15}P_3(\lambda)+\tfrac{71}{21}P_2(\lambda)+\tfrac{142}{35}P_1(\lambda)-\tfrac{49}{15}P_0(\lambda).$$

Remark 1. The argument used to prove Theorem 2 remains valid if $\gamma_i=0$, $\forall i$. If in addition we set $\alpha_i=1$, $\beta_i=0$, then formally $p_i(\lambda)\equiv\lambda^i$, and $b(\lambda)$ and $c(\lambda)$ reduce to

$$b(\lambda)=\lambda^m+\sum_{i=1}^m b_i\lambda^{m-i},\qquad c(\lambda)=\lambda^r+\sum_{i=1}^r c_i\lambda^{r-i}.$$

The matrix A_0 has a superdiagonal of 1's with all other elements zero, so by (2.6) and (2.7)

$$R_i=[\underset{\leftarrow(i-1)\rightarrow}{0,\cdots,0},b_m,b_{m-1},b_{m-2},\cdots].$$

It is trivial to check that in this case Theorem 2 is equivalent to collecting powers of λ in the expression

$$\lambda^r b(\lambda)+c_1\lambda^{r-1}b(\lambda)+\cdots+c_r b(\lambda).$$

Remark 2. If a more general polynomial basis is used [5], then the method carries over with A_0 replaced by a lower Hessenberg matrix, obtained from the so-called confederate matrix in the same way that A_0 is obtained from the comrade matrix.

3. Products of two orthogonal polynomials. For certain special cases it is possible to obtain explicit expressions for the product $p_m(\lambda)p_r(\lambda)$, $m\geqq r$. In Theorem 2, set $b(\lambda)=p_m(\lambda)$ and $c(\lambda)=p_r(\lambda)$. The first row R_1 of $p_m(A_0)$ is seen from (2.6) to be just the $(m+1)$th row of I_{m+r}, and will be denoted by e_{m+1}. Also, (2.9) shows that the vector of coefficients in the product $p_m(\lambda)p_r(\lambda)$ is just

$$(3.1)\qquad [x_{m+r},\cdots,x_2,x_1]=R_{r+1}.$$

It is convenient to record at this point that the ith row of A_0 has the form

$$(3.2)\qquad e_iA_0=\frac{1}{\alpha_i}(\gamma_ie_{i-1}-\beta_ie_i+e_{i+1}),\qquad i=2,3,\cdots,m+r-1,$$

$$(3.3)\qquad e_{m+r}A_0=\frac{1}{\alpha_{m+r}}(\gamma_{m+r}e_{m+r-1}-\beta_{m+r}e_{m+r}).$$

Case 1. $\alpha_1=1$, $\beta_1=0$; $\alpha_i=2$, $\beta_i=0$, $\gamma_i=\gamma$, $i\geqq 2$. From (2.7) we have

$$R_2=R_1A_0=e_{m+1}A_0=\tfrac{1}{2}(\gamma e_m+e_{m+2})$$

on applying (3.2), and similarly

$$\begin{aligned}R_3 = 2R_2A_0 - \gamma R_1 &= (\gamma e_m A_0 + e_{m+2}A_0) - \gamma e_{m+1}\\ &= \frac{\gamma}{2}(\gamma e_{m-1} + e_{m+1}) + \tfrac{1}{2}(\gamma e_{m+1} + e_{m+3}) - \gamma e_{m+1}\\ &= \tfrac{1}{2}(\gamma^2 e_{m-1} + e_{m+3}).\end{aligned}$$

It is easy to verify by induction that the $(i+1)$th row of $p_m(A_0)$ is

$$R_{i+1} = \tfrac{1}{2}(\gamma^i e_{m-i+1} + e_{m+i+1}). \tag{3.4}$$

In order to apply (3.1), set $i = r$ in (3.4) and formally ignore the term e_{m+r+1} (this is justified by (3.3)). Hence

$$[x_{m+r}, \cdots, x_1] = \frac{\gamma^r}{2} e_{m-r+1}$$

and Theorem 2 implies

$$p_m(\lambda)p_r(\lambda) = \frac{2^{r-1}}{2^r} p_{m+r}(\lambda) + \frac{\gamma^r}{2} p_{m-r}(\lambda) = \tfrac{1}{2}[p_{m+r}(\lambda) + \gamma^r p_{m-r}(\lambda)]. \tag{3.5}$$

In fact, the $p_i(\lambda)$ are essentially Chebyshev polynomials of the first kind, as the parameter γ can be removed by rescaling. Thus on setting $\gamma = 1$ in (3.5), the well known result [1, p. 782] for Chebyshev polynomials is recovered.

The procedure set out above is followed identically in all the following cases, so only the statements of the results will be given. In each case the expression for R_{i+1} is established by induction. The desired sum for $p_m(\lambda)p_r(\lambda)$ can then be written down using Theorem 2, the coefficient of e_s $(= x_{m+r-s+1})$ being equal to the coefficient of $p_{s-1}(\lambda)$ in the sum.

Case 2. $\alpha_i = 2$, $\beta_i = 0$, $\gamma_i = \gamma$, $i \geqq 1$. The $(i+1)$th row of $p_m(A_0)$ is

$$R_{i+1} = \gamma^i e_{m-i+1} + \gamma^{i-1} e_{m-i+3} + \cdots + \gamma e_{m+i-1} + e_{m+i+1}$$

so by Theorem 2

$$p_m(\lambda)p_r(\lambda) = \sum_{k=0}^{r} \gamma^k p_{m+r-2k}(\lambda). \tag{3.6}$$

Again by rescaling, γ could be removed, so setting $\gamma = 1$ in (3.6) gives a result for Chebyshev polynomials of the second kind [3].

Case 3. $\alpha_i = 1$, $\beta_i = \beta$, $\gamma_i = \gamma$, $i \geqq 1$.

$$\begin{aligned}R_{i+1} &= \sum_{k=0}^{i} \gamma^k e_{m+i-2k+1},\\ p_m(\lambda)p_r(\lambda) &= \sum_{k=0}^{r} \gamma^k p_{m+r-2k}(\lambda).\end{aligned} \tag{3.7}$$

Notice that (3.7) is independent of β; this is because λ could be replaced by $\lambda + \beta$.

Case 4. Hermite polynomials $H_i(\lambda)$, $\alpha_i = 2$, $\beta_i = 0$, $\gamma_i = 2(i-1)$, $i \geqq 1$.

$$R_{i+1} = \sum_{k=0}^{i} 2^k k! \binom{m}{k}\binom{i}{k} e_{m+i-2k+1}$$

where $\binom{i}{k} = i!/k!(i-k)!$,

$$H_m(\lambda)H_r(\lambda)=\sum_{k=0}^{r}2^k k!\binom{m}{k}\binom{r}{k}H_{m+r-2k}(\lambda). \tag{3.8}$$

Case 5. Legendre polynomials $P_i(\lambda)$. These were defined in the numerical example in § 2.

$$R_{i+1}=\sum_{k=0}^{i}\frac{\pi_k\pi_{i-k}i!\binom{m+i-k}{i}(2m+2i-4k+1)e_{m+i-2k+1}}{(2m+2i-2k+1)(2m+2i-2k-1)\cdots(2m-2k+1)}$$

where

$$\pi_r=\alpha_2\alpha_3\cdots\alpha_r,\quad r\geqq 2,\quad \pi_1=1,\quad \pi_0=1,$$

the α_i being given by (2.13).

$$P_m(\lambda)P_r(\lambda)=\sum_{k=0}^{r}\frac{\pi_k\pi_{r-k}r!\binom{m+r-k}{r}(2m+2r-4k+1)P_{m+r-2k}(\lambda)}{(2m+2r-2k+1)(2m+2r-2k-1)\cdots(2m-2k+1)}. \tag{3.9}$$

The formulae (3.8) and (3.9) are known; see [2] and [3, Lecture 5] for discussion and references. As pointed out in [3], one approach to obtaining such expressions is to find the coefficients for small values of m and r, guess the general formula, and prove it by induction. The method of derivation outlined in this paper, whilst also involving an induction step, is simpler and more systematic. The purely algebraic nature of the procedure seems attractively straightforward.

Acknowledgment. The author is grateful to the referees for their helpful suggestions which improved the presentation.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] R. Askey, *Linearization of the product of orthogonal polynomials*, in Problems in Analysis, R. Gunning, ed., Princeton Univ. Press, Princeton, NJ, 1970, pp. 131–138.

[3] ———, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[4] S. Barnett, *A companion matrix analogue for orthogonal polynomials*, Linear Algebra Appl., 12 (1975), pp. 197–208.

[5] ———, *Congenial matrices*, Linear Algebra Appl., 41 (1981), pp. 277–298.

[6] ———, *Division of generalized polynomials using the comrade matrix*, Linear Algebra Appl., to appear.

[7] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

002

HAMILTONIAN CYCLES IN STRIPED GRAPHS: THE TWO-STRIPE PROBLEM*

R. S. GARFINKEL† AND P. S. SUNDARARAGHAVAN‡

Abstract. For a directed graph $G=(N, A)$, the kth stripe of G is $R_k=\{(i, [i+k]_n)\}$ where $[a]_n$ is a (modulo n) and n is the number of nodes. A graph is *striped* if A consists of a set of stripes. The t-stripe problem is to determine whether a graph containing t stripes is hamiltonian. Necessary and sufficient conditions are provided for $t \leqq 2$, although the problem is still open for $t \geqq 3$.

AMS(MOS) subject classifications. 05C38, 05C45

1. Introduction. Consider a directed graph $G=(N, A)$ with nodes indexed by $I=\{0, 1, \cdots, n-1\}$. For $k \in I$, the kth *stripe* of G is defined by $R_k=\{(i, [i+k]_n), i \in I\}$ where $[a]_n$ denotes a (modulo n). G is *striped* if $A=\{R_k, k \in I^*\}$ where $I^* \subset I$.

In this paper we develop conditions for the existence of hamiltonian cycles in striped graphs. Our interest in these graphs was motivated by [1], [3], where traveling salesman problems are solved on graphs having striped distance matrices (i.e. constant distances over the stripes of the graph). Naturally, solution techniques have evolved which entail combining low cost stripes of G.

In [1] it was shown that if $I^*=\{k\}$, *gcd* $(k, n)=1$ is necessary and sufficient for G to be hamiltonian. If card $(I^*)=t>1$, we are interested in the possibility of combining stripes into an h-cycle. Thus we say that an h-cycle is a (nondegenerate) combination of $\{R_k, k \in I^*\}$ if it contains at least one arc from each stripe.

In this work we establish necessary and sufficient conditions for the case $t=2$ which can be checked simply in polynomial time. We also present a necessary condition for the case $t \geqq 3$, although we have not established that the 3-stripe problem is solvable in polynomial time. Sufficient conditions for that problem are given in [3]. We also present a polynomial algorithm for finding all h-cycles when $t=2$.

2. Necessary conditions. A striped graph having card $(I^*)=t$ is denoted by G_t.

LEMMA 1. *G_t is hamiltonian only if $w=gcd$ $(\{k \in I^*\}, n)=1$.*

Proof. Consider the nodes of G_t to be elements of the abelian group H_n with group operation being addition modulo n. Now, suppose $w>1$. Then the subgroup $H_n(0, w)$ of H_n, which contains 0 and is generated by w, is not the full group H_n; therefore it follows that the elements of $H_n(0, w)$ are a component of G_t so that G_t is disconnected in the strong sense.

Attention is now focused on the two-stripe problem. Let $I^*=\{p, q\}$, and denote G_2 by $G(p, q)$. For an arbitrary h-cycle in $G(p, q)$, let the number of arcs of stripes p and q be P and Q respectively. Also let $g=gcd$ $(p-q, n)$.

LEMMA 2. *P is an integral multiple of $d=n/g$.*

Proof. For any k, $[k+Pp+Qq]_n=k$ and from $P+Q=n$ it follows that

$$(p-q)P \equiv 0 \pmod{n}. \tag{1}$$

Thus (1) has g incongruent (mod n) solutions for P in the interval $[0, n)$. It is easily seen that any integer multiple of $d=n/g$ solves (1). Since there are g such solutions in $[0, n)$ the lemma is established.

* Received by the editors April 26, 1983, and in revised form June 27, 1983.

† Management Science Program, University of Tennessee, Knoxville, Tennessee 37916.

‡ Computer Systems and Production Management, University of Toledo, Toledo, Ohio 43606.

From Lemma 2 it follows that stripes p and q can be combined only if $g \geqq 2$. Also if n is prime, all stripes are h-cycles, but no two stripes can be combined into an h-cycle.

Before we state and prove the main result of this paper, namely necessary and sufficient conditions for $G(p, q)$ to be hamiltonian, some useful results about circular permutations are introduced in the next section.

3. Binary circular permutations. A "*binary*" *circular permuation* is a clockwise arrangement of two kinds of elements, namely p's and q's which number P and $n-P$ respectively, around the perimeter of a circle. Let S_c represent a subsequence of length c of a binary circular permutation, corresponding to consecutive (ordered) elements, where c is an integer in $[1, n)$. Define $S_c(p)$ to be the number of elements in S_c equal to p and SS_c to be the sum of the elements in S_c. Also let $\bar{S}_c(p) = \Sigma S_c(p)/n$ and $\overline{SS}_c = \Sigma SS_c/n$, where in each case the sum is taken over all n consecutive sequences of cardinality c. A binary circular permutation is said to be *symmetric with period* e, if $S_e(p)$ is the same regardless of the choice of the set S_e, and if e is the smallest integer with that property.

Clearly not all circular permutations correspond to h-cycles, but if the correspondence does exist, it should be clear to the reader how to derive the h-cycle from the permutation and vice versa.

Remark. A necessary and sufficient condition for a binary circular permutation to correspond to an h-cycle in $G(p, q)$ is that it contains no subsequence S_c, where c is an integer in $[1, n)$, such that

$$SS_c \equiv 0 \pmod n. \tag{2}$$

4. Necessary and sufficient conditions.

THEOREM 1. *$G(p, q)$ has a nondegenerate h-cycle if and only if*

$$gcd(p, q, n) = 1, \tag{3}$$

$g \geqq 2$, and there exists an integer $m \in [1, g-1]$ such that

$$gcd(m^*, n) = g \tag{4}$$

where $m^* = pm + q(g-m)$.

Proof.

Sufficiency. An algorithm is given for construction of a symmetric circular permutation which yields an h-cycle. Suppose there is an m satisfying (4) and further assume that (3) and $g \geqq 2$ hold.

ALGORITHM 1. Choose a straight line permutation S_g with $S_g(p) = m$. Repeat S_g n/g times around a circle to get a symmetric circular permutation, whose period t is a divisor of g. Use this permutation to obtain an h-cycle.

Now it will be shown that there is no sequence S_c in the permutation such that (2) and hence

$$SS_c = km^* + hp + lq \equiv 0 \pmod n \tag{5}$$

holds, where k is the integer part of c/g and $h + l = c - kg$. It follows that

$$h + l < g. \tag{6}$$

Rewrite (5) as

$$km^* \equiv -(h(p-q) + (h+l)q) \pmod n, \tag{7}$$

which has a solution for k if and only if $g = gcd(m^*, n)$ divides $h(p-q)+(h+l)q$. By definition $g|p-q$ and (7) holds only if $g|(h+l)q$. Then (6) implies $gcd(g, q) > 1$ and therefore since $g|n$, $gcd(n, g, q) > 1$. But $g|p-q$, which yields $gcd(n, p-q, q) > 1$, so that $gcd(n, p, q) > 1$ contradicting (3). It follows that (7) has no solution and therefore Algorithm 1 yields an h-cycle.

Necessity. Suppose there does not exist $m \in [1, g-1]$ satisfying (4) and let S_n be a binary circular permutation corresponding to an h-cycle. It follows from Lemma 2 that

(8) $$S_n(p) = mn/g$$

where $m \geqq 1$ and integer. Since $m \in [1, g-1]$ and since $g|m^*$ and $g|n$, it follows from the statement of the theorem that

(9) $$u = gcd(m^*, n) > g.$$

Define

(10) $$s = u/g$$

so that s is an integer in $[2, n/g]$. We will show that there exists $S_{n/s}$ corresponding to a subtour.

From (8), $\bar{S}_{n/s}(p) = mn/gs$ and it follows that there exists $S'_{n/s}$ such that

(11) $$S'_{n/s}(p) = mn/gs.$$

Now

(12) $$SS'_{n/s} = nm^*/gs$$

and from (9) and (10) it follows that $SS'_{n/s} \equiv 0 \pmod n$ and S_n does not yield an h-cycle.

Theorem 1 gives necessary and sufficient conditions for the existence of a nondegenerate h-cycle in $G(p, q)$ as well as an algorithm to find one if it exists. The next theorem establishes that the h-cycles constructed by Algorithm 1 are the only ones possible.

THEOREM 2. *All binary circular permutations corresponding to h-cycles obtained by nondegenerate combinations of stripes p and q in G(p, q) are symmetric with period g or some divisor of g.*

Proof. Suppose there is a circular permutation S_n corresponding to an h-cycle which is not symmetric with period g or any of its divisors. Let $p-q = zg$ where $z \in [1, (n/g)-1]$ and integer. From Theorem 1

(13) $$gcd(zg, n) = g = gcd(m^*, n).$$

Define $\hat{m} = [m^*]_n$ so that

(14) $$gcd(m^*, n) = gcd(\hat{m}, n) = g.$$

Let $Q(n)$ be the abelian group, consisting of integers modulo n, with group operation being addition modulo n. Also let G and G' be the subgroups of $Q(n)$ generated by g and $\hat{m}$ respectively. From (14), $G = G'$ and there must exist $\hat{z} \in [1, (n/g)-1]$ such that

(15) $$zg \equiv \hat{z}\hat{m} \pmod n.$$

We focus attention on $S_{\hat{z}g}$ and note that $S_{\hat{z}g} \subset S_n$ since $\hat{z}g \in [g, n-g]$.

From (13), (15)

$$g = gcd(zg, n) = gcd([zg]_n, n) = gcd([\hat{z}\hat{m}]_n, n),$$

and with (14)

$$g = gcd(\hat{z}\hat{m}, n) = gcd(\hat{z}g, n). \tag{16}$$

From (16) it follows that S_n is not symmetric with period $\hat{z}g$ or any of its divisors. From (8) and (15)

$$\overline{SS}_{\hat{z}g} = m^*\hat{z} \equiv \hat{m}\hat{z} \equiv zg \pmod{n},$$

so that there exists an integer α such that

$$\overline{SS}_{\hat{z}g} = \alpha n + zg. \tag{17}$$

Since S_n is not symmetric with period $\hat{z}g$ or any of its divisors, it follows that there exits $S^*_{\hat{z}g}$ such that

$$S^*_{\hat{z}g}(p) = \bar{S}_{\hat{z}g}(p) - 1$$

so that

$$SS^*_{\hat{z}g} = \overline{SS}_{\hat{z}g} - p + q.$$

Then (17) and the definition of z yield $SS^*_{\hat{z}g} \equiv 0 \pmod{n}$, so that $S^*_{\hat{z}g}$ satisfies (2), contradicting the assumption that S_n yields an h-cycle.

5. Example. Consider the example given by $n = 12$, $p = 7$, and $q = 3$ where (3) is clearly satisfied. Now $g = 4$ and (4) holds for $m = 1, 2$ but not for $m = 3$. Thus all h-cycles in this graph correspond to the binary circular permutations (written linearly) 7333 7333 7333, 7733 7733 7733, and 7373 7373 7373. The corresponding h-cycles are of the form 1-8-11-2-5-12-3-6-9-4-7-10-1, 1-8-3-6-9-4-11-2-5-12-7-10-1, and 1-8-11-6-9-4-7-2-5-12-3-10-1 respectively. Note that for $m = 5$, the subsequence 7773 yields the subtour 1-8-3-10-1.

6. Conclusions. We have established simple necessary and sufficient conditions for a 2-striped graph to be hamiltonian. Essentially these can be checked by calculating on the order $g - 1$ greatest common divisors each of which can be done in $O(\log n)$ time using Euclid's algorithm as discussed in [2]. Since $g < n$, the resulting algorithm is no worse than $O(n \log n)$.

The problem of extending these results for three or more stripes is an intriguing one which the authors are pursuing. The question of the computational complexity of the t-stripe problem is an open one for $t \geqq 3$.

REFERENCES

[1] R. S. Garfinkel, *Minimizing wallpaper waste, Part* 1: *A class of traveling salesman problems*, Oper. Res., 25 (1977), pp. 741–751.

[2] D. E. Knuth, *The Art of Computer Programming*: *Vol.* 2, *Semi-Numerical Algorithms*, Addison-Wesley, Reading, MA, 1969.

[3] P. S. Sundararaghavan, *Hamiltonian cycles and chains in graphs with striped and circulant distance matrices*: *theory and applications*, Ph.D. dissertation, College of Business Administration, University of Tennessee, Knoxville, 1981.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

COMPLEMENTARITY IN ORIENTED MATROIDS*

MICHAEL J. TODD†

Abstract. We extend many of the results and algorithms of linear complementarity theory to the abstract combinatorial setting of oriented matroids.

AMS(MOS) subject classifications. 90C33, 05B35

1. Introduction. The linear complementarity problem is to find w and z satisfying

$$w = Az + b, \tag{1.1a}$$

$$w \geqq 0, \qquad z \geqq 0, \tag{1.1b}$$

$$w^T z = 0, \tag{1.1c}$$

for given $n \times n$ matrix A and n-vector b. Results concerning this problem and discussions of its importance in mathematical programming, game theory and other fields appear in [6], [8], [18], [19], [21]. Generally, the given data A, b and the solution vectors w, z are assumed to be real, but any ordered field suffices to prove the known results and establish the validity of the usual algorithms. In this paper, we will attempt to show that the natural setting of the problem is that of oriented matroids, the study of abstract combinatorial properties of signed linear dependencies introduced by Bland [1], Las Vergnas [14] and Lawrence, whose thesis [17] presents and extends work of Folkman. See also Bland and Las Vergnas [4], Folkman and Lawrence [10]. To justify this assertion, we will extend several important results and algorithms of linear complementarity theory to the context of oriented matroids. This extension is proper, since there are oriented matroids that are not representable over any ordered field [4]. Our proofs must therefore avoid any dependence on monotonicity. To some extent, our task is simpler than that of Bland [2] or Edmonds and Fukuda (see [11]) in extending the theory and algorithms of linear programming to oriented matroids, since the usual complementary pivot algorithms rely very little on monotonicity. However, they do rely heavily on nondegeneracy (or lexicographic resolution of degeneracy) and we need to extend such techniques to oriented matroids.

The two results we have chosen to extend are the characterizations of those matrices A for which there is exactly one solution for each b and for which there is at least one solution for each principal submatrix of A and each commensurate b. Such matrices are called P-matrices and completely-Q matrices respectively and were characterized by Samelson, Thrall and Wesler [22] and Cottle [5].

Samelson, Thrall and Wesler showed that the following are equivalent:

(1.2a) The linear complementarity problem (1.1) has exactly one solution for every b;

(1.2b) the matrix A has positive principal minors; and

(1.2c′) if $e_1, e_2, \cdots, e_n$ are the columns of the identity matrix, and $a_1, a_2, \cdots, a_n$ the columns of A, then

* Received by the editors October 5, 1982, and in revised form May 5, 1983. This research was partially supported by a Fellowship from the Alfred P. Sloan Foundation and by the National Science Foundation under grant ECS-7921279.

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853.

(i) if $u_i \in \{e_i, -a_i\}$ for $i=1, 2, \cdots, n$, then the u_i's are linearly independent, and

(ii) if $u_i \in \{e_i, -a_i\}$ for $i=1, 2, \cdots, n$, then for each j, e_j and $-a_j$ lie on opposite sides of the hyperplane spanned by $u_1, \cdots, u_{j-1}, u_{j+1}, \cdots, u_n$.

This result was rediscovered independently by Ingleton [13] and Murty [21]. Other equivalent conditions are known. For instance, Murty [21] showed that the condition below suffices:

(1.2ã) The linear complementarity problem (1.1) has at most one solution for every b.

It is also obvious that the following is equivalent:

(1.2b′) The matrix $-A^T$ has alternating principal minors, i.e. its principal submatrices of order k have determinants with the same sign as $(-1)^k$.

We shall see that there is a duality relationship between (b) and (b′). There is also a dual version of (1.2c′), in which (ii) is replaced by

(ii) If $u_i \in \{e_i, -a_i\}$ for $i=1, 2, \cdots, n$, and for some j, $\{u_j, u_j'\} = \{e_j, -a_j\}$, then the linear dependence among $u_1, \cdots, u_{j-1}, u_j, u_j', u_{j+1}, \cdots, u_n$ involves u_j and u_j' with the same sign.

We will call the resulting condition (1.2c).

Gale and Nikaido [12], in investigating the class of matrices with positive principal minors, provided another equivalent condition:

(1.2d) There is no nonzero x with $y = Ax$ and $x_i y_i \leqq 0$ for all $i=1, 2, \cdots, n$.

If $y = Ax$, then $y - Ax = 0$ so that (1.2d) is related to dependencies among the e_i's and $-a_i$'s; it is easily seen that (1.2d) implies (1.2c). There is a dual version of (1.2d), also equivalent:

(1.2d′) There is no hyperplane H such that for each $i=1, 2, \cdots, n$, e_i and $-a_i$ lie in one of the two closed half-spaces associated with H.

If any of (1.2a)–(1.2d′) holds we say A is a P-matrix. Note that all the conditions involve sign properties rather than numerical values. Indeed, they can all be expressed as properties of the signed linear dependencies among the columns of the matrices

$$(1.3) \qquad [I, -A] \quad \text{and} \quad [I, -A, -b]$$

or the signs of vectors in the row spaces of these matrices. (Note that if $(x, y) \neq 0$ is in the row space of $[I, -A]$, then there is a hyperplane H with $x_j(y_j)$ zero, positive or negative according as $e_j(-a_j)$ is in H, on the positive side or on the negative side of H.) Thus it is natural to ask whether these conditions remain equivalent in the context of oriented matroids. Theorem A states that they do.

Among the algorithms designed to solve (1.1) are those of Lemke [18] and Van der Heyden [26]. Both are known to yield a solution whenever A is strictly semi-monotone, i.e., whenever $0 \neq x \geqq 0$ implies that for some k $x_k > 0$ and $(Ax)_k > 0$. Indeed, Cottle [5] has recently shown that the following conditions are equivalent:

(1.4a) (1.1) has at least one solution for each b, and so does every principal subproblem (i.e., a linear complementarity problem whose matrix is a principal submatrix of A);

(1.4b) A is strictly semi-monotone.

Eaves [7] had previously shown that the following is equivalent to (1.4b):

(1.4c) For every $b \geqq 0$, (1.1) has precisely one solution.

Theorem B asserts the equivalence of combinatorial versions of (1.4a), (1.4b) and a slightly strengthened form of (1.4c) that asserts that the condition holds for all principal submatrices of A also. We will prove Theorem B constructively, using algorithms that generalize those of Lemke and Van der Heyden. These algorithms will therefore be guaranteed to find a solution if our translations of conditions (1.4b) or (1.4c) hold, and hence if our versions of any of conditions (1.2b)–(1.2d′) hold. Another algorithm that can be applied to the linear complementarity problem is the principal pivoting method of Cottle and Dantzig [6]. The usual argument that guarantees convergence for this method involves monotonicity of the chosen negative basic variable—this reasoning is no longer valid in the context of oriented matroids. However, when A is a P-matrix, the principal pivoting method can be seen to be equivalent to the algorithm of Van der Heyden—see [26, p. 339].

Our main purpose is to show precisely what structure is sufficient to establish the major results of linear complementarity. (In a related vein, Bland [3] has shown that oriented matroids are the most general combinatorial objects for which the Farkas property together with certain natural symmetry and closure properties holds. Thus this structure is in some sense necessary as well as sufficient to prove the main results in linear programming.) We also show that related constructions, such as lexicographic rules, can be extended to oriented matroids, using results of Las Vergnas on extensions of matroids [15].

However, we also provide insight into the linear complementarity problem itself. Since the way in which (1.1) arises from linear or quadratic programming problems incorporates both primal and dual problems, it has been believed that there is little duality applicable to (1.1) itself (apart from the use of Farkas' lemma to establish infeasibility). However the arguments we use rely heavily on the dual oriented matroid and thus stress the importance of duality. Indeed, (1.2b′)–(1.2d′) are conditions on the dual oriented matroid, and all arguments involving hyperplanes containing columns e_i and $-a_j$ and with columns e_k and $-a_l$ on various sides (in particular, discussions of complementary cones) use duality. Our combinatorial versions of conditions (1.2b)–(1.2d′) will exhibit this duality very clearly. Next, we show that Van der Heyden's variable dimension algorithm (when it succeeds, or when a strong nondegeneracy assumption holds) is a special case of Lemke's, where the artificial column is $d = (\delta^n, \delta^{n-1}, \cdots, \delta)^T$ for sufficiently small positive δ. This provides some answer to Cottle's first question [5] concerning the applicability of Van der Heyden's algorithm; it will process only problems which Lemke's algorithm will process for some positive d, and it will process all "strongly nondegenerate" problems which Lemke's algorithm will process for all positive d.

Finally, our development raises the hope that quadratic programming can be studied in the context of oriented matroids, just as linear programming was by Bland [2] and Edmonds and Fukuda [11]. We may also be able to derive a new algorithm for linear programming in oriented matroids, related to Dantzig's self-dual parametric algorithm [7, pp. 245–246] as is Bland's [2] to Dantzig's simplex method. Indeed, linear and quadratic programming duality results for oriented matroids can be proved constructively using the ideas of this paper—see [24].

We conclude this section with an outline of the remainder of the paper.

Section 2 is an introduction to the theory of oriented matroids, containing the concepts and many of the results we shall need later. In § 3 we state our generalizations of the theorems of Samelson–Thrall–Wesler and Cottle. Section 4 establishes an important "unique pivoting" result. We show in § 5 how lexicographic extensions can be used to cure the ills of degeneracy. Section 6 proves our first main result, while § 7 proves the second and also describes two algorithms that can be considered as extensions of those of Van der Heyden and Lemke. The paper concludes with a result relating the former algorithm to a special case of the latter, and a discussion of the orientation of the solution found by our algorithms.

2. Oriented matroids. We follow [4]. Let E be a finite set. A *signed set* in E is a pair $X=(X^+,X^-)$ with $X^+\subseteq E$, $X^-\subseteq E$ and $X^+\cap X^-=\emptyset$. The *opposite* of $X=(X^+,X^-)$ is the signed set $-X=(X^-,X^+)$ and the *set underlying* X is $\underline{X}=X^+\cup X^-$. We say that X contains e if $e\in\underline{X}$ and that e and f appear in X with the same sign (opposite signs) if $e,f\in X^+$ or $e,f\in X^-$ ($e\in X^+$ and $f\in X^-$, or $e\in X^-$ and $f\in X^+$).

The pair $\mathscr{M}=(E,\mathscr{C})$ is an oriented matroid if E is a finite set and $\mathscr{C}$ a collection of signed sets in E, called circuits, satisfying

$(\emptyset,\emptyset)\notin\mathscr{C}$,
(C1) $C\in\mathscr{C}\Rightarrow -C\in\mathscr{C}$,
$C, C'\in\mathscr{C}$ and $\underline{C}\subseteq\underline{C}'\Rightarrow C=C'$ or $C=-C'$;

$C_1, C_2\in\mathscr{C}$ and $e\in(C_1^+\cap C_2^-)\cup(C_1^-\cap C_2^+)$
(C2) imply that there exists $C_3\in\mathscr{C}$ with
$C_3^+\subseteq(C_1^+\cup C_2^+)\backslash\{e\}$, $C_3^-\subseteq\{C_1^-\cup C_2^-\}\backslash\{e\}$.

(C2) is the so-called "signed elimination property." If we let $\underline{\mathscr{C}}=\{\underline{C}: C\in\mathscr{C}\}$ then $\underline{\mathscr{M}}=(E,\underline{\mathscr{C}})$ is an unoriented matroid as introduced by Whitney [27], and (C1) and (C2) become the usual circuit axioms.

Let M be a matrix over an ordered field, with its columns indexed by E. Then M gives rise to an oriented matroid $\mathscr{M}(M)$ as follows. Let $\underline{C}\subseteq E$ index a minimal nonempty linearly dependent set of columns. Thus $Mx=0$ where x is nonzero only on indices in $\underline{C}$. The minimality condition implies that x is unique up to scalar multiplication. Let the signed set C be defined by setting $C^+(C^-)$ to be the index set of the positive (negative) components of x. Clearly $-x$ gives rise to the signed set $-C$. The collection $\mathscr{C}$ of all signed sets arising this way is the set of circuits of $\mathscr{M}(M)$. We can also view $\mathscr{M}$ as the collection of signed supports of elementary vectors (i.e., vectors of minimal nonempty support) in the null space of M. It is then natural to consider the oriented matroid obtained similarly from the orthogonal subspace, the row space of M. This gives rise to another oriented matroid in the representable case, called the dual of $\mathscr{M}(M)$.

Note that there are oriented matroids which are not representable over any ordered field, i.e. which cannot be written as $\mathscr{M}(M)$ for any M. However, corresponding to *every* oriented matroid $\mathscr{M}=(E,\mathscr{C})$ there is a uniquely defined *dual* oriented matroid denoted $\mathscr{M}^*=(E,\mathscr{D})$. The circuits of $\mathscr{M}^*$, also called *cocircuits* of $\mathscr{M}$, are signed sets D, satisfying the *orthogonality property* (2.1) below, which have minimal nonempty underlying sets:

(2.1) for all $C\in\mathscr{C}$, either
$\underline{C}\cap\underline{D}=\emptyset$, or $(C^+\cap D^+)\cup(C^-\cap D^-)\neq\emptyset$ and $(C^+\cap D^-)\cup(C^-\cap D^+)\neq\emptyset$.

Further, $(\mathscr{M}^*)^*=\mathscr{M}$.

In the oriented matroid $\mathscr{M}$, a subset of E is called *dependent* if it contains (the underlying set of) some circuit, and *independent* otherwise. A maximal independent subset of E is called a *base* of $\mathcal{M}$ (*cobase* of $\mathcal{M}^*$). For any $F \subseteq E$, all maximal independent subsets of F have the same cardinality, called the *rank* of F. In particular, all bases have the same size, called the *rank* of $\mathcal{M}$. The bases of $\mathcal{M}^*$ (cobases of $\mathcal{M}$) are just the complements in E of the bases of $\mathcal{M}$. If B is a base of $\mathcal{M}$ and $e \notin B$, then there is exactly one circuit C with $e \in C^+$ and $\underline{C} \subseteq B \cup \{e\}$; we call it the *fundamental circuit* associated with B and e.

If D is a cocircuit of $\mathcal{M}$, then $E \backslash \underline{D}$ is a maximal subset of E of rank one less than the rank of $\mathcal{M}$—it is termed a *hyperplane*. The elements of $\underline{D}$ can be viewed as lying on one side or the other of this hyperplane according as they belong to D^+ or D^-.

All of this notation arises from and is a natural generalization of the representable case. Next we show how to obtain new oriented matroids from old, by reorienting certain elements or taking minors.

Let $\mathcal{M} = (E, \mathscr{C})$ be an oriented matroid and $\mathcal{M}^* = (E, \mathscr{D})$ its dual. Let F be a subset of E, and let $\mathscr{C}(F) = \{(((C^+ \backslash F) \cup (C^- \cap F)), ((C^- \backslash F) \cup (C^+ \cap F)): C \in \mathscr{C}\}$ and $\mathscr{D}(F)$ be defined similarly. Then clearly $\mathcal{M}' = (E, \mathscr{C}(F))$ is an oriented matroid, and $(\mathcal{M}')^* = (E, \mathscr{D}(F))$ its dual. We say $\mathcal{M}'$ is obtained from $\mathcal{M}$ by *reversing signs* on F. In the representable case, this merely corresponds to negating the columns indexed by F.

To describe oriented matroid minors, we need some terminology. For $F \subseteq E$ and C a signed set in E we denote by $C \backslash F$ the signed set $(C^+ \backslash F, C^- \backslash F)$. For $\mathscr{C}$ a collection of signed sets in E we say $C \in \mathscr{C}$ is a minimal nonempty member of $\mathscr{C}$ if $C \neq (\emptyset, \emptyset)$ and for no $C' \in \mathscr{C}$ is $\emptyset \subsetneqq \underline{C}' \subsetneqq \underline{C}$. Now let $F, G \subseteq E$ with $F \cap G = \emptyset$. Let $\mathscr{C} \backslash F/G$ denote the collection of minimal nonempty signed sets in $\{C \backslash G : C \in \mathscr{C}, \underline{C} \cap F = \emptyset\}$. Then $(E \backslash (F \cup G), \mathscr{C} \backslash F/G)$ is an oriented matroid, written $\mathcal{M} \backslash F/G$ and called the *matroid minor of* $\mathcal{M}$ *obtained by deleting F and contracting G*. We have $(\mathcal{M} \backslash F/G)^* = \mathcal{M}^* \backslash G/F$. We write $\mathcal{M} \backslash F$ $(\mathcal{M}/G)$ for $\mathcal{M} \backslash F/\phi$ $(\mathcal{M} \backslash \phi/G)$. If C is a circuit of $\mathcal{M}$ and $G \subseteq \underline{C}$, then it is easy to see that $C \backslash G$ is a circuit of $\mathcal{M}/G$. A useful result [4, Prop. 4.4] is that even if $G \nsubseteq \underline{C}$, there is a circuit $\bar{C}$ of $\mathcal{M}/G$ with $\bar{C}^+ \subseteq C^+ \backslash G$, $\bar{C}^- \subseteq C^- \backslash G$.

If $p \notin E$ and $\mathcal{M} = (E, \mathscr{C})$, $\hat{\mathcal{M}} = (E \cup \{p\}, \hat{\mathscr{C}})$ are oriented matroids with $\hat{\mathcal{M}} \backslash \{p\} = \mathcal{M}$, we say $\hat{\mathcal{M}}$ is a point extension of $\mathcal{M}$. Las Vergnas [15] has studied such point extensions and characterized those with rank equal to that of $\mathcal{M}$. (There is only one trivial point extension with rank one greater than that of $\mathcal{M}$.) Henceforth by a point extension of a matroid we shall mean one with equal rank.

Again, minors have straightforward interpretations in the representable case. Deleting F corresponds to removing from M the columns indexed by F. To contract G, we first project each remaining column orthogonal to the subspace spanned by the columns indexed by G, and then delete the latter columns.

All the material above (except point extensions) can be found in Bland and Las Vergnas [4]. The final concept we need, orientation of ordered bases, was introduced in the context of oriented matroids by Las Vergnas [16]. In the representable case, each ordered basis of a subspace can be given a sign, depending on the sign of the determinant of its representation in terms of a canonical positive ordered basis. Clearly, this is related to our interest in matrices with positive principal minors. Note that if β is an ordering of the basis B, and C is the fundamental circuit associated with B and e, with $f \in B \cap \underline{C}$, then the ordered basis β' obtained from β by replacing f with e has the same sign as β if and only if e and f appear in C with opposite signs. Las Vergnas [16] proved that such an assignment of signs was possible also in the nonrepresentable case.

Call any two orderings of the same set *equivalent* if they can be obtained from each other by an even permutation, *opposite* otherwise. Las Vergnas showed that there are precisely two assignments ε of signs to orderings of bases in an oriented matroid (one being the opposite of the other) with the following properties:

(2.2)
(i) if β and β' are orderings of a base B, then $\varepsilon(\beta)=\varepsilon(\beta')$ iff β and β' are equivalent;
(ii) if β and β' are orderings of two bases B and B', and β and β' agree in all but one position, then $\varepsilon(\beta)=\varepsilon(\beta')$ iff the two elements of $B \triangle B'$ appear in the circuit C with $\underline{C} \subseteq B \cup B'$ with opposite signs.

We call such an ε an *orientation* of the bases of $\mathcal{M}$. If ε^* is an orientation of the bases of $\mathcal{M}^*$, we also call ε^* an orientation of the cobases of $\mathcal{M}$.

3. Main results. In this section we translate conditions (1.2) and (1.4) into the language of oriented matroids and state our main results.

We will assume

(3.1) $S=\{s_1, s_2, \cdots, s_n\}$, $T=\{t_1, t_2, \cdots, t_n\}$, $S \cap T=\emptyset$ and $E=S \cup T$;
$p \notin E$ and $\hat{E}=E \cup \{p\}$;
$\mathcal{M}=(E, \mathcal{C})$ is an oriented matroid with dual $\mathcal{M}^*=(E, \mathcal{D})$ and S is a base of $\mathcal{M}$.

We say a set $F \subseteq \hat{E}$ is *complementary* (*almost-complementary*) if F contains both s_i and t_i for no i (for at most one i). A signed set X is complementary (almost-complementary) iff its underlying set is. We say X is *positive on a set* $F \subseteq \hat{E}$ if $X^- \cap F=\emptyset$ and *positive* if it is positive on $\hat{E}$ itself. Finally, we call X *sign-preserving* (*sign-reversing*) if for each $i=1, 2, \cdots, n$, if s_i and t_i both belong to $\underline{X}$, they appear in X with the same sign (opposite signs).

To view the conditions (3.2a)–(3.2d′) below as extensions of (1.2a)–(1.2d′), think of $\mathcal{M}$ and $\hat{\mathcal{M}}$ as the oriented matroids represented by the matrices $[I, -A]$ and $[I, -A, -b]$, with S, T and p indexing the columns of I, of $-A$, and $-b$ respectively.

(3.2a) Every point extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ to $\hat{E}$ contains precisely one positive complementary circuit.

(3.2ã) Every point extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ to $\hat{E}$ contains at most one positive complementary circuit.

(3.2b) Every complementary subset U of E with cardinality n is a base of $\mathcal{M}$. Furthermore, there is an orientation ε of the bases of $\mathcal{M}$ with $\varepsilon(\upsilon)=(-1)^{|U \cap T|}$, where υ is the natural ordering $(u_1, u_2, \cdots, u_n)$ of the complementary base U with $u_i \in \{s_i, t_i\}$ for each i.

(3.2b′) Every complementary subset U of E with cardinality n is a cobase of $\mathcal{M}$. Furthermore, there is an orientation ε^* of the cobases of $\mathcal{M}$ with $\varepsilon^*(\upsilon)=+1$, where υ is the natural ordering $(u_1, u_2, \cdots, u_n)$ of the complementary cobase U with $u_i \in \{s_i, t_i\}$ for each i.

(3.2c) There is no almost-complementary sign-reversing circuit in $\mathcal{M}$.

(3.2c′) There is no almost-complementary sign-preserving cocircuit in $\mathcal{M}$.

(3.2d) There is no sign-reversing circuit in $\mathcal{M}$.

(3.2d′) There is no sign-preserving cocircuit in $\mathcal{M}$.

THEOREM A. *Given* (3.1), *conditions* (3.2a)–(3.2d′) *are all equivalent.*

Let us make a few comments about the relationship of conditions (3.2) to conditions (1.2). First, the positive complementary circuit in $\hat{\mathscr{M}}$ whose existence is claimed in (3.2a) must include the element p. Otherwise, $\mathscr{M}$ contains a positive complementary circuit, and then the extension of $\mathscr{M}$ to $\hat{\mathscr{M}}$ by $p=0$ would contain two positive complementary circuits, that in $\mathscr{M}$ and $(\{p\}, \emptyset)$. (We will discuss extensions of $\mathscr{M}$ by various p's, including $p=0$, in § 5; however, the meaning should be clear at this stage.)

Second, we have stated all the conditions in terms of circuits (minimal dependencies) rather than members of the "signed span" $K(\mathscr{C})$ of the circuits of $\mathscr{M}$ (dependencies), and similarly with cocircuits. However, the equivalences are straightforward using the conformal decomposition of members of the signed span (see Bland [2, Thm. 3.2]).

Third, we have stated the conditions (3.2b)–(3.2d′) in "dual pairs" to stress the symmetry between conditions on circuits and bases and those on cocircuits and cobases. In the representable case, this symmetry can be expressed by noting that A has positive principal minors iff its transpose has; for generalizing to oriented matroids, we stated this in the rather less transparent form of (1.2b) and (1.2b′)—note that $\mathscr{M}=\mathscr{M}(I, -A)$ iff $\mathscr{M}^*=\mathscr{M}(A^T, I)$. There is also a natural symmetry between s_i and t_i, that is obvious in conditions (3.2b′)–(3.2d′) and easily derived in (3.2b). In the representable case, this symmetry corresponds to the fact that A has positive principal minors iff every principal pivot transform of it has (Tucker [25]).

In order to state conditions generalizing (1.4a)–(1.4c), we need to define various important minors of $\mathscr{M}$.

DEFINITION 3.3. Let I and J be disjoint subsets of $\{1, 2, \cdots, n\}$. Then $\mathscr{M}_I^J$ denotes the oriented matroid minor obtained by deleting all t_i, $i \in I$, and all s_j, $j \in J$, and contracting all s_i, $i \in I$, and t_j, $j \in J$. We write $\mathscr{M}_I$ for $\mathscr{M}_I^{\emptyset}$ and $\mathscr{M}^J$ for $\mathscr{M}_{\emptyset}^J$. We call $\mathscr{M}_I$ a principal submatroid of $\mathscr{M}$.

We can now state conditions extending those of (1.4).

(3.4a) For every $I \subseteq \{1, 2, \cdots, n\}$, every point extension $\hat{\mathscr{M}}_I$ of $\mathscr{M}_I$ contains at least one positive complementary circuit containing the new element of $\hat{\mathscr{M}}_I$.

(3.4b) $\mathscr{M}$ contains no sign-reversing circuit that is positive on T.

(3.4c) For every $I \subseteq \{1, 2, \cdots, n\}$, if $\hat{\mathscr{M}}_I$ is a point extension of $\mathscr{M}_I$ containing a positive complementary circuit C with $\underline{C} \cap T = \emptyset$, then $\hat{\mathscr{M}}_I$ contains no other positive complementary circuit.

THEOREM B. *Given* (3.1), *conditions* (3.4a)–(3.4c) *are equivalent.*

4. Preliminaries. Our aim in this section is to prove a pivoting result, (4.2), that is closely related to Bland's Claim 4.3 [2] and crucial to the algorithms we shall develop. To prove this, we require the following apparent strengthening of the signed elimination axiom (C2):

THEOREM 4.1 ([4, Thm. 2.1]). *Under the condition* (C1), *the elimination property* (C2) *is equivalent to*

(C3) $C_1, C_2 \in \mathscr{C}$, $e \in (C_1^+ \cap C_2^-) \cup (C_1^- \cap C_2^+)$ *and* $f \in (C_2^+ \backslash C_1^-) \cup (C_2^- \backslash C_1^+)$ *imply that there exists* $C_3 \in \mathscr{C}$ *with* $f \in \underline{C}_3$, $C_3^+ \subseteq (C_1^+ \cup C_2^+) \backslash \{e\}$ *and* $C_3^- \subseteq (C_1^- \cup C_2^-) \backslash \{e\}$.

From this we obtain the important theorem.

THEOREM 4.2. *Let C_1 and C_2 be circuits of the oriented matroid $\mathcal{M} = (E, \mathscr{C})$. Let $f \in \underline{C}_2 \backslash \underline{C}_1$, and suppose $(C_1^+ \cap C_2^-) \cup (C_1^- \cap C_2^+) \neq \emptyset$. Then there exists a circuit C_3 such that*

(i) $C_3^+ \subseteq C_1^+ \cup C_2^+$ *and* $C_3^- \subseteq C_1^- \cup C_2^-$;

(ii) $C_3^+ \cap C_1^- = C_3^- \cap C_1^+ = \emptyset$; *and*

(iii) $f \in (C_3^+ \cap C_2^+) \cup (C_3^- \cap C_2^-)$.

Furthermore, if $\underline{C}_2 \subseteq \underline{C}_1 \cup \{f\}$, then such a C_3 is unique and satisfies

(iv) $\underline{C}_1 \backslash \underline{C}_2 \subseteq \underline{C}_3$.

Proof. For the first part, we use induction on $d(C_1, C_2) \equiv |(C_1^+ \cap C_2^-) \cup (C_1^- \cap C_2^+)|$. If this is one, then condition (C3) yields a circuit C_3 that is easily seen to satisfy (i)–(iii). Suppose the result is true whenever $d(C_1, C_2) < k$, and consider the case where $d(C_1, C_2) = k$. Choose any e in $(C_1^+ \cap C_2^-) \cup (C_1^- \cap C_2^+)$ and apply (C3) to get a circuit C_2'. If $C_3 = C_2'$ satisfies (ii) we are done. Otherwise, C_1 and C_2' satisfy the hypotheses and disagree in sign only where C_1 and C_2 do. Since $e \notin \underline{C}_2'$, $d(C_1, C_2')$ is strictly smaller than k. Thus we may apply the inductive hypothesis to obtain C_3 satisfying (i)–(iii) with respect to C_1 and C_2', and hence also with respect to C_1 and C_2. This completes the inductive step.

For the second part, assume first that (iv) fails. Then apply the elimination property to C_2 and $-C_3$ to eliminate f. This yields some $C_4 \in \mathscr{C}$ with $\underline{C}_4 \subsetneq \underline{C}_1$, a contradiction. To prove uniqueness, suppose C_5 and C_6 both satisfy (i)–(iv), with $e_5 \in \underline{C}_5 \backslash \underline{C}_6$ and $e_6 \in \underline{C}_6 \backslash \underline{C}_5$. Obtain C_7 by eliminating f in C_5 and $-C_6$. Then $\underline{C}_7 \backslash \underline{C}_1$, so $C_7 = \pm C_1$. But e_5 appears in C_1, C_5 and thus C_7 with the same sign, whereas e_6 appears in C_1, C_6 and thus $-C_7$ with the same sign. This contradiction establishes uniqueness.

We will usually apply Theorem 4.2 when the final hypothesis, $\underline{C}_2 \subseteq \underline{C}_1 \cup \{f\}$, is also satisfied. In this case we say that C_3 is obtained by applying the unique pivot result Theorem 4.2 to the circuits C_1 and C_2.

Suppose that C_i is the fundamental circuit associated with a base B and e_i, $i = 1, 2$ (with $f = e_2$). In order that $\underline{C}_2 \subseteq \underline{C}_1 \cup \{f\}$ be satisfied, it is sufficient that $\underline{C}_1 = B \cup \{e_1\}$—this assumption is related to a nondegeneracy assumption that is customarily made in complementary pivot theory. In that context it is justified by the use of lexicographic rules. We shall also employ such techniques to handle degeneracy in § 6. The next section treats lexicographic extensions.

5. Lexicographic extensions.

We require lexicographic extensions of a matroid for two reasons: first, they allow us to resolve the problems of degeneracy in algorithms to generate positive complementary circuits; and second, they provide a sufficiently rich class of extensions of a matroid that we may prove not only the sufficiency of conditions (3.2b)–(3.2d) for (3.2a) and (3.2ã), but also their necessity.

The basic result we need follows from Theorem 1.2 and the lemma of § 3 in Las Vergnas [15]. We state it as a theorem.

THEOREM 5.1. *Let $\mathcal{M} = (E, \mathscr{C})$ be an oriented matroid, and let $\{e_1, e_2, \cdots, e_k\} \subseteq E$ be independent and $p \notin E$. Then there is precisely one point extension $\hat{\mathcal{M}} = (\hat{E}, \hat{\mathscr{C}})$ of $\mathcal{M}$ with $\hat{E} = E \cup \{p\}$ and dual $\hat{\mathcal{M}}^* = (\hat{E}, \hat{\mathscr{D}})$ satisfying*

$\hat{D} \in \hat{\mathscr{D}}$ *if* $\hat{D} \in \mathscr{D}$ *and* $\{e_1, e_2, \cdots, e_k\} \cap \underline{\hat{D}} = \emptyset$, *and*

$\hat{D} \in \hat{\mathscr{D}}$ *if* $\hat{D} \backslash \{p\} = D \in \mathscr{D}$ *and* $\{e_1, e_2, \cdots, e_k\} \cap \underline{D} \neq \emptyset$, *with p appearing in $\hat{D}$ with the same sign as the first e_i in D.*

Moreover, each cocircuit $\hat{D}$ of $\hat{\mathcal{M}}$ containing p is of the latter form.

Note that in the representable case, such an extension arises if we set the vector p to be $e_1+\varepsilon e_2+\cdots+\varepsilon^{k-1}e_k$ for some suitably small positive ε. Motivated by this, we make the following definition.

DEFINITION 5.2. Suppose $\hat{\mathcal{M}}$ arises as above from $\mathcal{M}$. Then, if $k>0$, we say $p=\operatorname{lex}(e_1, e_2, \cdots, e_k)$ extends $\mathcal{M}$ to $\hat{\mathcal{M}}$ and call $\hat{\mathcal{M}}$ a lexicographic extension of $\mathcal{M}$. If $k=0$, then p lies in no cocircuit of $\hat{\mathcal{M}}$ and thus is a loop (forms a one-element circuit). We therefore say $p=0$ extends $\mathcal{M}$ to $\hat{\mathcal{M}}$ and call $\hat{\mathcal{M}}$ the zero-extension of $\mathcal{M}$. If $k>0$ and $\tilde{\mathcal{M}}$ is obtained from $\hat{\mathcal{M}}$ by reversing the sign of p, we say $p=-\operatorname{lex}(e_1, e_2, \cdots, e_k)$ extends $\mathcal{M}$ to $\tilde{\mathcal{M}}$. Similarly, if we reverse the sign of e_1 in $\mathcal{M}$ to get $\check{\mathcal{M}}$, then extend $\check{\mathcal{M}}$ to $\hat{\mathcal{M}}$, then reverse the sign of e_1 to get $\dot{\mathcal{M}}$, we say $p=\operatorname{lex}(-e_1, e_2, \cdots, e_k)$ extends $\mathcal{M}$ to $\dot{\mathcal{M}}$, and so on.

Now we establish some important properties of lexicographic extensions. The first shows that they can be used to resolve degeneracy, which corresponds to circuits of cardinality smaller than the rank of $\mathcal{M}$ plus one.

PROPOSITION 5.3. *If $p=\operatorname{lex}(e_1, e_2, \cdots, e_k)$ extends $\mathcal{M}$ to $\hat{\mathcal{M}}$, then every circuit $\hat{C}$ of $\hat{\mathcal{M}}$ containing p contains at least $k+1$ elements.*

Proof. Suppose $\hat{C}\in\hat{\mathscr{C}}$, $p\in\underline{\hat{C}}$, and $|\hat{C}|\leqq k$. Let I be the independent (in $\mathcal{M}$) set $\underline{\hat{C}}\backslash\{p\}\subseteq E$. Then since I has fewer than k elements and $\{e_1, e_2, \cdots, e_k\}$ is independent, there is some $e_j\notin I$ with $I\cup\{e_j\}$ independent. Let B be a base of $\mathcal{M}$ containing $I\cup\{e_j\}$. Let e_i be the first e_k in $B\backslash I$ (note that e_j is some such e_i), and let $D\in\mathscr{D}$ be the fundamental cocircuit associated with the cobase $E\backslash B$ and e_i. Then D gives rise to a cocircuit $\hat{D}$ of $\hat{\mathcal{M}}$ with $p\in\underline{\hat{D}}$. However, $\underline{\hat{C}}$ and $\underline{\hat{D}}$ then intersect only in p contradicting orthogonality.

Next we show that $\hat{\mathcal{M}}$ contains the obvious circuit.

PROPOSITION 5.4. *If $p=-\operatorname{lex}(e_1, e_2, \cdots, e_k)$ extends $\mathcal{M}$ to $\hat{\mathcal{M}}$, then there is a positive circuit $\hat{C}$ of $\hat{\mathcal{M}}$ with $\underline{\hat{C}}=\{p, e_1, e_2, \cdots, e_k\}$.*

Proof. Let $\hat{C}=(\{p, e_1, e_2, \cdots, e_k\}, \phi)$. Then it is easy to see that $\hat{C}$ is orthogonal to every $\hat{D}\in\hat{\mathscr{D}}$ containing p. Suppose $\hat{D}\in\hat{\mathscr{D}}$ does not contain p, is positive on $\{e_1, e_2, \cdots, e_k\}$ and contains e_i as its first e_j. There is a cocircuit D_1 of $\mathcal{M}$ meeting $\{e_1, e_2, \cdots, e_k\}$ just in e_i with $e_i\in D_1^-$. This gives rise to a cocircuit $\hat{D}_1$ of $\hat{\mathcal{M}}$. Now apply (C3) to $\hat{D}$ and $\hat{D}_1$ eliminating e_i, to get $\hat{D}_2$ with $p\in\underline{\hat{D}}_2$. But $p\in\hat{D}_1^+$, so $p\in\hat{D}_2^+$ and $e_j\in\underline{\hat{D}}_2$ implies $e_j\in\hat{D}_2^+$. This contradicts Theorem 5.1. Thus $\hat{C}$ is orthogonal to all cocircuits of $\hat{\mathcal{M}}$. Since it can contain no circuit by Proposition 5.3, it must itself be a circuit of $\hat{\mathcal{M}}$.

Our final proposition is concerned with recognizing circuits involving p in the lexicographic extension $\hat{\mathcal{M}}$ by knowing the circuits and cocircuits of $\mathcal{M}$. As a corollary we show how lexicographic perturbations can be removed. For simplicity we only consider the case where $k=\operatorname{rank}\mathcal{M}$.

PROPOSITION 5.5. *Let $p=\operatorname{lex}(e_1, e_2, \cdots, e_k)$ extend $\mathcal{M}$ to $\hat{\mathcal{M}}$, where $\mathcal{M}$ has rank k. Let $\hat{C}$ be a circuit of $\hat{\mathcal{M}}$ with $p\in\hat{C}^+$ and let B be the base $\underline{\hat{C}}\backslash\{p\}$ of $\mathcal{M}$. Then $e\in B\cap\hat{C}^+$ or $e\in B\cap\hat{C}^-$ according as the fundamental cocircuit D of $\mathcal{M}$ associated with the cobase $E\backslash B$ and e has $e_i\in D^-$ or $e_i\in D^+$, where e_i is the first e_j in $\underline{D}$.*

Proof. Associated with the fundamental cocircuit D is a cocircuit $\hat{D}$ of $\hat{\mathcal{M}}$, with $\underline{\hat{D}}=\underline{D}\cup\{p\}$ since D contains at least one e_j. Then $\hat{D}$ is orthogonal to $\hat{C}$ and meets it only in e and p. Hence $e\in B\cap\hat{C}^+$ iff $p\in\hat{D}^-$, which holds exactly when $e_i\in D^-$, where e_i is the first e_j in $\underline{D}$.

COROLLARY 5.6. *Let $p=\operatorname{lex}(e_1, e_2, \cdots, e_k)$ extend $\mathcal{M}$ to $\hat{\mathcal{M}}$, where $\mathcal{M}$ has rank k. Let $\hat{C}$ be a circuit of $\hat{\mathcal{M}}$ with $p\in\hat{C}^+$ and $e_1\notin\underline{\hat{C}}$. Then there exists exactly one circuit C of $\mathcal{M}$ with $e_1\in C^+$, $C^+\backslash\{e_1\}\subseteq\hat{C}^+$ and $C^-\subseteq\hat{C}^-$.*

Proof. Let B be the base $\underline{\hat{C}}\setminus\{p\}$ of $\mathcal{M}$ and let C be the fundamental circuit of $\mathcal{M}$ associated with B and e_1. Consider any $e \in B$ and let D be the fundamental cocircuit of $\mathcal{M}$ associated with the cobase $E\setminus B$ and e. Then $e \in \hat{C}^+$ implies $e_1 \in D^-$ or $e_1 \notin \underline{D}$, which implies, by orthogonality of C and D, that $e \in C^+$ or $e \notin \underline{C}$. This proves existence. Moreover, any circuit C' satisfying the conditions has $\underline{C}' \subseteq B \cup \{e_1\}$, so must be the unique fundamental circuit C.

6. Proof of Theorem A. Throughout this section we assume that (3.1) holds, and abbreviate conditions (3.2a)–(3.2d') by (a)–(d'). For many of the arguments, the reader may find it helpful to represent circuits pictorially, so that a circuit $\hat{C}$ with $\hat{C}^+ = \{s_1, s_4, p\}$, $\hat{C}^- = \{t_2\}$ could be represented as:

s_1	s_2	s_3	s_4	p
+			+	+
	−			
t_1	t_2	t_3	t_4	

For reasons of space we have given the arguments concisely, avoiding such diagrams.

The proof is divided into several parts. First, we note the trivial implications

(6.1) (a)$\Rightarrow$(ã), (d)$\Rightarrow$(c) and (d')$\Rightarrow$(c').

Next we prove the following lemma.

LEMMA 6.2. (b)$\Leftrightarrow$(b')$\Leftrightarrow$(c)$\Leftrightarrow$(c').

Proof. Since $\mathcal{M}$ has rank n, the nonexistence of a complementary cocircuit (circuit) implies that every complementary subset of E of size n is a cobase (base), hence a base (cobase), and thus that there is no complementary circuit (cocircuit). Further, given that every complementary n-set is a base, the nonexistences of almost-complementary sign-reversing circuits and sign-preserving cocircuits are equivalent by orthogonality of fundamental circuits and cocircuits. Hence (c)$\Leftrightarrow$(c').

Now assuming (c), we have seen that every complementary n-set is a base. Let ε be the orientation of bases of $\mathcal{M}$ such that $\varepsilon(s_1, s_2, \cdots, s_n) = +1$. Then (b) is true when $|U \cap T| = 0$. If (b) is true when $|U \cap T| < k$ and U is a complementary n-set with $|U \cap T| = k$, we choose $t_i \in U$ and let $U' = (U\setminus\{t_i\}) \cup \{s_i\}$. Then by the induction hypothesis, $\varepsilon(v') = (-1)^{k-1}$. Now apply (2.2ii) and (c) to find $\varepsilon(v) = (-1)^k$. This establishes (c)$\Rightarrow$(b). The reverse implication follows similarly. An analogous argument proves (c')$\Leftrightarrow$(b').

LEMMA 6.3. (ã)$\Rightarrow$(c).

Proof. Suppose (ã) holds. If $\mathcal{M}$ contained a positive complementary circuit C, then the 0-extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ would contain two positive complementary circuits, namely C and $(\{p\}, \varnothing)$. Now suppose $\mathcal{M}$ contains a complementary circuit C with C^+, C^- both nonempty. Let $\underline{C} = \{e_1, e_2, \cdots, e_k, e_{k+1}\}$ with $e_{k+1} \in C^+$. Let $p = -\text{lex}\,(e_1, e_2, \cdots, e_k)$ extend $\mathcal{M}$ to $\hat{\mathcal{M}}$. Then by Proposition 5.4 there is a positive complementary circuit $\hat{C}_1$ in $\hat{\mathcal{M}}$ with $\underline{\hat{C}}_1 = \{p, e_1, e_2, \cdots, e_k\}$. Now apply the unique pivot theorem, Theorem 4.2, to $\hat{C}_1$ and C to obtain a circuit $\hat{C}_3$ with $e_{k+1} \in \hat{C}_3^+$. Also, $\hat{C}_3$ is positive and complementary, contradicting (ã). Hence $\mathcal{M}$ has no complementary circuit.

Next assume that C is an almost complementary circuit in $\mathcal{M}$ with $s_i \in C^+$, $t_i \in C^-$ for some $i = 1, 2, \cdots, n$. Then $\underline{C}\setminus\{s_i\}$ is complementary, so it can be extended to a complementary base B, say $\{e_1, e_2, \cdots, e_{n-1}, t_i\}$. Let $p = -\text{lex}\,(e_1, e_2, \cdots, e_{n-1}, t_i)$

extend $\mathcal{M}$ to $\hat{\mathcal{M}}$. Then, again, Proposition 5.4 gives a positive complementary circuit $\hat{C}_1$ with $\underline{\hat{C}}_1 = B \cup \{p\}$. Now use the signed elimination axiom (C2) (eliminating t_i) to get a circuit $\hat{C}_2$ of $\hat{\mathcal{M}}$ with $\underline{\hat{C}}_2 \subseteq (\underline{\hat{C}}_1 \cup \{s_i\}) \backslash \{t_i\}$, $s_i \in \hat{C}_2^+$ and $p \in \hat{C}_2^+$. Suppose $e_j \in \hat{C}_2^-$. By Proposition 5.3, $p \in \hat{C}_2^+$ implies $|\underline{\hat{C}}_2| = n+1$, so that $\underline{\hat{C}}_2 = (\underline{\hat{C}}_1 \cup \{s_i\}) \backslash \{t_i\}$ and $\underline{\hat{C}}_2 \backslash \{p\} = (B \cup \{s_i\}) \backslash \{t_i\}$ is a base B' of $\hat{\mathcal{M}}$. Let $\hat{D}$ be the fundamental cocircuit of $\hat{\mathcal{M}}$ associated with $\hat{E} \backslash B'$ and e_j. Then $e_j \in \hat{D}^+$, and since $B \cap \underline{\hat{D}} \subseteq \{e_j, t_i\}$, $p \in \hat{D}^-$. But $\underline{\hat{C}}_2 \cap \underline{\hat{D}} = \{p, e_j\}$ implies that $\hat{C}_2$ and $\hat{D}$ are not orthogonal. Thus $\hat{C}_2$ is a positive complementary circuit which together with $\hat{C}_1$ contradicts (ã). Hence we have proved (ã)⇒(c).

Recall that if $I, J \subseteq \{1, 2, \cdots, n\}$ with $I \cap J = \varnothing$, then the matroid $\mathcal{M} \backslash \{t_i : i \in I\} \cup \{s_j : j \in J\} / \{s_i : i \in I\} \cup \{t_j : j \in J\}$ is denoted $\mathcal{M}_I^J$.

LEMMA 6.4. *If* (c) *or* (c′) *holds for* $\mathcal{M}$, *it holds for all minors of the form* $\mathcal{M}_I^J$ *also.*

Proof. It is sufficient to prove the conclusion for $\mathcal{M}_i = \mathcal{M}_{\{i\}}^{\varnothing}$ and $\mathcal{M}^j = \mathcal{M}_{\varnothing}^{\{j\}}$. But if $\mathcal{M}_i$ contains a complementary circuit C_i, then $\mathcal{M}$ contains a complementary circuit C with $\underline{C} \subseteq \underline{C}_i \cup \{s_i\}$. Similarly if $\mathcal{M}_i$ contains an almost-complementary sign-reversing circuit C_i, then $\mathcal{M}$ contains a similar circuit C with $\underline{C} \subseteq \underline{C}_i \cup \{s_i\}$. The argument is analogous for $\mathcal{M}^j$.

LEMMA 6.5. (c)⇒(d) *and* (c′)⇒(d′).

Proof. We show (c)⇒(d); (c′)⇒(d′) follows by duality, reversing signs on T. The proof is by induction on n; for $n = 1$, the implication is clear. Thus suppose the result is true for $n < k$ and consider the case $n = k$. We assume (c) is true for $\mathcal{M}$, and hence for all $\mathcal{M}_I^J$. Suppose C is a sign-reversing circuit in $\mathcal{M}$, i.e. if $\{s_i, t_i\} \subseteq \underline{C}$, s_i and t_i appear in C with opposite signs. If for any i, $|\{s_i, t_i\} \cap \underline{C}| = 1$, then we find a sign-reversing circuit C_i in $\mathcal{M}_i$ (if $s_i \in \underline{C}$) or $\mathcal{M}^i$ (if $t_i \in \underline{C}$). The induction hypothesis then implies that (c) fails for $\mathcal{M}_i$ or $\mathcal{M}^i$, a contradiction. Thus $n_i \equiv |\{s_i, t_i\} \cap \underline{C}| = 0$ or 2 for all i. If $n_i = 0$ for all i, $\underline{C}$ is empty, a contradiction. If $n_i = 2$ for just one i, then (c) is violated. Since $|\underline{C}| \leqq n+1$, if $n_i = 2$ for more than one i, then $n_j = 0$ for some j. If $|\underline{C}| < n+1$, then $\operatorname{rank}(\underline{C}) < n$ so there is some $s_i \notin \operatorname{cl}(\underline{C})$ (the closure $\operatorname{cl}(\underline{C})$ of $\underline{C}$ is $\underline{C}$ together with all $e \in E$ such that $\underline{C} \cup \{e\}$ contains a circuit containing e). Thus C is also a sign-reversing circuit in $\mathcal{M}_i$, and again a contradiction arises. Suppose now $|\underline{C}| = n+1$. Let C_1 be the fundamental circuit associated with the base $\underline{C} \backslash \{s_i\}$ and s_j, where $n_i = 2$ and $n_j = 0$. By the arguments above C_1 is not sign-reversing; thus for some $k \neq i$ with $n_k = 2$, s_k and t_k appear in C_1 with the same sign (and in C with opposite signs). Now apply the unique pivot theorem, Theorem 4.2, to C and C_1 to obtain a sign-reversing circuit C_3 with $|\{s_j, t_j\} \cap \underline{C}_3| = 1$ for some j. As in the first part of the proof, this leads to a contradiction.

To prove Theorem A, it only remains to prove the implication (c)⇒(a). This entails two parts—existence and uniqueness. We will prove existence using an inductive proof that is the basis of our algorithms in the next section. Uniqueness will follow from:

LEMMA 6.6. (d)⇒(ã).

Proof. Suppose (d) holds but some point extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ has two positive complementary circuits $\hat{C}_1$ and $\hat{C}_2$. Apply the signed elimination axiom (C2) to $\hat{C}_1$ and $-\hat{C}_2$, eliminating p. The result is a circuit C of $\mathcal{M}$, and it is easy to see that C is sign-reversing.

We prove existence of a positive complementary circuit in $\hat{\mathcal{M}}$ inductively. The result is trivial for $n = 0$, for the empty matroid $\mathcal{M}$ has rank 0, and the only point extension $\hat{\mathcal{M}}$ of it with the same rank is the 0-extension, with the positive complementary circuit $(\{p\}, \varnothing)$. (If the reader dislikes the case $n = 0$ as a basis for the induction, the proof for $n = 1$ is straightforward: indeed it is the inductive step we prove below.)

So we assume that (c)⇒(a) for $n < k$ and consider the case $n = k$. Let $\hat{\mathcal{M}}$ be a point extension of $\mathcal{M}$. If p lies in some circuit of $\hat{\mathcal{M}}$ of size smaller than $n+1$, we proceed as follows. Let s_i lie in the fundamental circuit associated with S and p; without loss of generality, $i=1$. Then $\{p, s_2, s_3, \cdots, s_n\}$ is a base of $\hat{\mathcal{M}}$. Let $q = \operatorname{lex}(p, s_2, s_3, \cdots, s_n)$ extend $\hat{\mathcal{M}}$ to $\mathring{\mathcal{M}}$, and let $\tilde{\mathcal{M}} = \mathring{\mathcal{M}}\setminus(p\}$. Then $\tilde{\mathcal{M}}$ is a point extension of $\mathcal{M}$, and every circuit of $\tilde{\mathcal{M}}$ containing q contains $n+1$ elements. Moreover, if we find a positive complementary circuit $\tilde{C}$ in $\tilde{\mathcal{M}}$, then Corollary 5.6 will yield a positive complementary circuit $\hat{C}$ in $\hat{\mathcal{M}}$. Hence we may assume without loss of generality that $\hat{\mathcal{M}}$ is an extension such that every circuit of $\hat{\mathcal{M}}$ containing p contains $n+1$ elements.

Now consider the principal submatroid $\mathcal{M}_n$ and its extension $\hat{\mathcal{M}}_n = \hat{\mathcal{M}}\setminus\{t_n\}/\{s_n\}$. By Lemma 6.4, if (c) holds for $\mathcal{M}$ it holds for $\mathcal{M}_n$, so by the induction hypothesis there is a positive complementary circuit in $\hat{\mathcal{M}}_n$, and thus there is a complementary circuit $\hat{C}_1$ in $\hat{\mathcal{M}}$ with $\hat{C}_1^- \subseteq \{s_n\}$. If $s_n \in \hat{C}_1^+$, $\hat{C}_1$ is the desired positive complementary circuit. Thus assume $s_n \in \hat{C}_1^-$.

(If we did not wish to give a constructive proof, we could derive (c)$\Rightarrow$(a) simply from here. Indeed, a similar argument involving $\mathcal{M}^n$ either gives the desired positive complementary circuit or gives a complementary circuit $\hat{C}_2$ with $\hat{C}_2^- = \{t_n\}$. In the latter case the signed elimination axiom (C2) applied to $\hat{C}_1$ and $-\hat{C}_2$ to eliminate p gives a sign-reversing circuit in $\mathcal{M}$, thus contradicting (d) and hence by Lemma 6.5 (c). However, we wish to motivate the algorithms of the next section.)

We will attempt to convert $\hat{C}_1$ into a positive complementary circuit by performing a sequence of (complementary) pivots. The circuits we construct are of the following form.

DEFINITION 6.7. A signed subset $\hat{X}$ of $\hat{E}$ is *n-distinguished* if $\hat{X}^- \subseteq \{s_n\}$, $s_n \notin \hat{X}^+$, $p \in \hat{X}^+$ and $\underline{\hat{X}}\setminus\{s_n\}$ is complementary. Two n-distinguished circuits $\hat{C}_1$ and $\hat{C}_2$ of $\hat{\mathcal{M}}$ are *adjacent* if $(\hat{C}_1^+ \cup \hat{C}_2^+, \hat{C}_1^- \cup \hat{C}_2^-)$ is a signed set that is n-distinguished.

The algorithm we use will construct a sequence of n-distinguished circuits of $\hat{\mathcal{M}}$ with each consecutive pair adjacent. The Lemke–Howson argument that precludes cycling in linear complementarity [18], [19] is also valid in oriented matroids:

LEMMA 6.8. *Assume* (c) *holds and consider the graph whose nodes are n-distinguished circuits with adjacency as defined above. In this graph, a node has degree* 1 *if it is complementary and* 2 *otherwise.*

Proof. Let $\hat{C}$ be a complementary n-distinguished circuit. Then $|\underline{\hat{C}}| = n+1$ and so $\hat{C}$ contains precisely one of each pair $\{s_i, t_i\}$ together with p. Suppose $s_n \in \hat{C}^-$. Then consider the fundamental circuit C of $\mathcal{M}$ associated with the base $\underline{\hat{C}}\setminus\{p\}$ and t_n. We have $s_n \in C^+$ by (c). Thus we may apply the unique pivot result Theorem 4.2 to $\hat{C}$ and C to obtain an adjacent n-distinguished circuit $\hat{C}'$. Moreover, $\hat{C}'$ is unique, since the only possible member of $\underline{\hat{C}}'\setminus\underline{\hat{C}}$ is t_n so that $\underline{\hat{C}}' \subseteq \underline{\hat{C}} \cup \{t_n\}$ and Theorem 4.2 applies. If $s_n \notin \hat{C}^-$, so that $t_n \in \hat{C}^+$, the argument is similar using $-C$, where C is the fundamental circuit of $\mathcal{M}$ associated with $\underline{\hat{C}}\setminus\{p\}$ and $s_n (s_n \in (-C)^-, t_n \in (-C)^-$ by (c)). Thus $\hat{C}$ has degree 1.

Now let $\hat{C}$ be an n-distinguished circuit that is *not* complementary. Then, since $|\underline{\hat{C}}| = n+1$ and $\{s_n, t_n\} \subseteq \underline{\hat{C}}$, there is precisely one index i such that $\hat{C}$ contains neither s_i nor t_i. Let B be the base $\underline{\hat{C}}\setminus\{p\}$ and let C_s be the fundamental circuit of $\mathcal{M}$ associated with B and s_i. Then $s_i \in C_s^+$ and C_s contains s_n and t_n with the same sign by (c). Thus the unique pivot result Theorem 4.2 applied to $\hat{C}$ and C_s gives an adjacent n-distinguished circuit $\hat{C}_s$. It is easy to see that $\hat{C}_s$ is the only such circuit containing s_i. A similar argument using t_i instead of s_i throughout gives another adjacent n-distinguished circuit $\hat{C}_t$, the only such containing t_i. But if $\hat{C}'$ is an n-distinguished circuit adjacent to $\hat{C}$, the only choices for the member of $\underline{\hat{C}}\setminus\underline{\hat{C}}$ are s_i and t_i. Hence $\hat{C}$ has degree 2. This completes the proof.

We are now ready for the next proposition.

PROPOSITION 6.9. *Assume* (c) *holds and that* $\hat{C}_1$ *is a complementary n-distinguished circuit of* $\hat{\mathcal{M}}$ *that is not positive. Then there is a unique maximal chain* $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_k$ *of distinct n-distinguished circuits of* $\hat{\mathcal{M}}$ *with each consecutive pair adjacent. Moreover,* $\hat{C}_k$ *is a positive complementary circuit of* $\hat{\mathcal{M}}$.

Proof. By Lemma 6.8 the graph of n-distinguished circuits of $\hat{\mathcal{M}}$ is a disjoint union of paths and cycles, and $\hat{C}_1$ is the endpoint of a path. Let the nodes on the path be $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_k$. This is clearly the unique maximal chain desired. Moreover, $\hat{C}_k$ is complementary and unequal to $\hat{C}_1$. If $s_n \in \hat{C}_k^-$ then $\hat{C}_k \backslash \{s_n\}$ is a positive complementary circuit in $\hat{\mathcal{M}}_n$ other than $\hat{C}_1 \backslash \{s_n\}$. But (c) and hence (d) and (a) hold for $\mathcal{M}_n$, and thus there cannot be two positive complementary circuits in $\hat{\mathcal{M}}_n$. Hence $s_n \notin \hat{C}_k^-$ and thus $\hat{C}_k$ is a positive complementary circuit as claimed.

Theorem A follows from (6.1), Lemmas 6.2, 6.3, 6.5, 6.6 and Proposition 6.9.

7. Proof of Theorem B and algorithms. Here we will abbreviate conditions (3.4a)–(3.4c) by (a)–(c). We first have this lemma.

LEMMA 7.1. *If* (b) *holds for* $\mathcal{M}$, *it also holds for all principal submatroids* $\mathcal{M}_I$.

The *proof* follows that of Lemma 6.4 and is omitted.

LEMMA 7.2. (b)$\Rightarrow$(c).

Proof. By Lemma 7.1 it is sufficient to prove that (b) implies the truth of (c) for $I = \emptyset$. Suppose (c) fails. Then there is an extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ that contains the positive complementary circuits C_1 and C_2 where $\underline{C}_1 \cap T = \emptyset$. Now apply the signed elimination property to C_2 and $-C_1$, eliminating p. The result is a sign-reversing circuit in $\mathcal{M}$ that is positive on T, contradicting (b).

LEMMA 7.3. (c)$\Rightarrow$(b).

Proof. By induction on n. If $n = 1$ and (b) fails, then either $\mathcal{M}$ contains a positive complementary circuit in which case the zero-extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ violates (c), or $C = (\{t_1\}, \{s_1\})$ is a circuit of $\mathcal{M}$. In the latter case, let $p = -\text{lex}(s_1)$ extend $\mathcal{M}$ to $\hat{\mathcal{M}}$. Then $C_1 = (\{p, s_1\}, \emptyset)$ is a positive complementary circuit by Proposition 5.4. Moreover, the signed elimination property applied to C_1 and C to eliminate s_1 gives $C_2 = (\{p, t_1\}, \emptyset)$ as another positive complementary circuit, violating (c). This proves the case $n = 1$.

Now suppose the lemma is true for $n < k$ and consider the case $n = k$. Suppose (c) holds but (b) fails, so that $\mathcal{M}$ contains a sign-reversing circuit C that is positive on T. If for any $i = 1, 2, \cdots, n$, $t_i \notin \underline{C}$, then $C \backslash \{s_i\}$ contains a sign-reversing circuit in $\mathcal{M}_i$ that is positive on T. But since (c) trivially holds for $\mathcal{M}_i$ we have a contradiction to the induction hypothesis. Thus $T \subseteq \underline{C}$. If $T = \underline{C}$, then the zero-extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ violates (c), and so $\underline{C} = T \cup \{s_i\}$ for some s_i, and $s_i \in C^-$, $T = C^+$. Now let $p = -\text{lex}(s_i)$ extend $\mathcal{M}$ to $\hat{\mathcal{M}}$. Then $C_1 = (\{p, s_1\}, \emptyset)$ is a positive complementary circuit in $\hat{\mathcal{M}}$, and applying the signed elimination property to C_1 and C eliminating s_1 gives another positive complementary circuit C_2; but this contradicts (c).

LEMMA 7.4. (a)$\Rightarrow$(b).

Proof. Again, by induction on n. Suppose $n = 1$. Let $\mathcal{M}$ contain the complementary circuit $(\{t_1\}, \emptyset)$. Then if $p = \text{lex}(s_1)$ extends $\mathcal{M}$ to $\hat{\mathcal{M}}$, the only circuits in $\hat{\mathcal{M}}$ are $(\{p\}, \{s_1\})$, $(\{t_1\}, \emptyset)$ and their negatives, so that (a) fails. If $\mathcal{M}$ contains the circuit $(\{t_1\}, \{s_1\})$ then let $p = \text{lex}(t_1)$ extend $\mathcal{M}$ to $\hat{\mathcal{M}}$. Then the only circuits in $\hat{\mathcal{M}}$ are $(\{p\}, \{t_1\})$, $(\{t_1\}, \{s_1\})$ and (by the signed elimination property applied to these) $(\{p\}, \{s_1\})$ together with their negatives, and again (a) fails. Thus (a)$\Rightarrow$(b) if $n = 1$.

Assume the lemma is true for $n < k$ and consider the case $n = k$. Suppose (a) is satisfied but C is a sign-reversing circuit that is positive on T. If for some $i = 1, 2, \cdots, n$, $t_i \notin \underline{C}$, then $C \backslash \{s_i\}$ contains a sign-reversing circuit in $\mathcal{M}_i$ that is positive on T. But then our inductive hypothesis implies that (a) fails for $\mathcal{M}_i$, hence for $\mathcal{M}$. Thus $T \subseteq \underline{C}$.

Case 1. For some j, $s_j \in C^-$. Then T is a base of $\mathcal{M}$. If for every $i = 1, 2, \cdots, n$, we have $t_1 \notin C_i^+$, where C_i is the fundamental circuit associated with the base T and s_i, then the fundamental cocircuit D associated with the cobase S and t_1 is positive. But then if $p = \text{lex}(t_1)$ extends $\mathcal{M}$ to $\hat{\mathcal{M}}$, $(D^+ \cup \{p\}, \varnothing)$ is a positive cocircuit in $\hat{\mathcal{M}}$, which then has no positive circuit containing p by orthogonality, contradicting (a). Thus for some i, we have $t_1 \in C_i^+$. Now apply the unique pivot result Theorem 4.2 to C and $-C_i$ to get a sign-reversing circuit C' in $\mathcal{M}$ that is positive on T—note that $t_1 \in C^+ \cap (-C_i)^-$. But $T \not\subseteq \underline{C}'$, and this leads to a contradiction to (a) as above.

Case 2. $\underline{C} = T$. Then $\underline{C} \backslash \{t_1\}$ is independent, so there is some s_i, $i = 1, 2, \cdots, n$ such that $B = (\underline{C} \cup \{s_i\}) \backslash \{t_1\}$ is a base of $\mathcal{M}$. Let C_j, $j \neq i$, be the fundamental circuit associated with B and s_j. If $s_i \notin C_j^+$ for every j, then the fundamental cocircuit D associated with the cobase $E \backslash B$ and s_i is positive, and we are led to a contradiction to (a) as in Case 1. So let $s_i \in C_j^+$. If $T \cap C_j^+ = \varnothing$, then $-C_j$ is a sign-reversing circuit that is positive on T, and $t_1 \notin (-XC_j)$. This leads to a contradiction to (a) as above. If $T \cap C_j^+ \neq \varnothing$, then we may apply the first part of Theorem 4.2 to $C_1 = C$ and $C_2 = -C_j$ to obtain a circuit C_3 that is sign-reversing but does not contain T. Once again a contradiction to (a) results. This completes the inductive step and hence the proof.

It is now only necessary to prove (b)$\Rightarrow$(a), and thus we do constructively. The algorithm is that of § 6, suitably extended, and corresponds to the method of Van der Heyden [26] in the representable case. Note that it is only necessary to prove (a) for $I = \varnothing$, in light of Lemma 7.1. Thus let p extend $\mathcal{M}$ to $\hat{\mathcal{M}}$; we seek a positive complementary circuit involving p. As in § 6, we may assume without loss of generality that every circuit of $\hat{\mathcal{M}}$ containing p contains $n+1$ elements. While we do not prove (a) completely with just one application of the algorithm, we will in fact construct positive complementary circuits in each $\hat{\mathcal{M}}_I$ (the appropriate minor of $\hat{\mathcal{M}}$) for a nested set of n I's, which we may take to be $\{i+1, i+2, \cdots, n\}$ for $1 \leqq i \leqq n$, the last set being $\varnothing$. For the following definition, it is convenient to let S_k and T_k denote $\{s_1, s_2, \cdots, s_k\}$ and $\{t_1, t_2, \cdots, t_k\}$ respectively. Also, let $\hat{\mathcal{M}}_{(i)}$ denote $\hat{\mathcal{M}}_I$ as above, i.e. the minor of $\hat{\mathcal{M}}$ obtained by deleting all t_j's and contracting all s_j's for $i < j \leqq n$.

DEFINITION 7.5. A signed subset $\hat{X}$ of $\hat{E}$ is *i-distinguished*, $1 \leqq i \leqq n$, if $s_i \in \hat{X}^- \subseteq \{s_i, s_{i+1}, \cdots, s_n\} \subseteq \underline{\hat{X}}$, $p \in \hat{X}^+$, and $\hat{X} \backslash \{s_i\}$ is complementary. It is $(n+1)$*-distinguished* if it is positive and complementary and contains p. An i-distinguished set $\hat{X}$ and a j-distinguished set $\hat{X}'$ are *adjacent* if $\hat{X} \neq \hat{X}'$, $(\underline{\hat{X}} \cup \underline{\hat{X}}') \backslash \{s_k\}$ is complementary, and no element of $S_k \cup T_k$ appears in $\hat{X}$ and in $\hat{X}'$ with opposite signs, where $k = \min\{i, j\}$.

We seek an $(n+1)$-distinguished circuit $\hat{C}$ of $\hat{\mathcal{M}}$. Let $\hat{C}_1$ be the fundamental circuit of $\hat{\mathcal{M}}$ associated with the base S and p; we call $\hat{C}_1$ the initial circuit. If $\hat{C}_1$ is positive, it is $(n+1)$-distinguished and we are done. Otherwise, it is i-distinguished, where s_i is the first s_j in $\hat{C}_1^-$. Note that any i-distinguished circuit $\hat{C}$, since it contains p, has cardinality $n+1$. Thus either it is complementary, or there is exactly one $h < i$ with $\hat{C}$ containing neither s_h nor t_h.

The algorithm to generate an $(n+1)$-distinguished circuit proceeds to construct a sequence of adjacent distinguished circuits from $\hat{C}_1$. It can only fail when the next circuit cannot be found. As we shall see, the following definition captures this possibility.

DEFINITION 7.6. Let $\hat{C}$ be an i-distinguished circuit of $\hat{\mathcal{M}}$ and C a circuit of $\mathcal{M}$. We say that $\hat{C}$ is the *endpoint of the ray* C if either

(i) $\hat{C}$ is complementary, and either C is the fundamental circuit associated with the base $\underline{\hat{C}} \backslash \{p\}$ and t_i with $s_i \notin C^+$ and C positive on $S_{i-1} \cup T_{i-1}$, or, where t_h is the last t_j in $\hat{C}$, C is the negative of the fundamental circuit associated with the base $\underline{\hat{C}} \backslash \{p\}$ and s_h with C positive on $(S_h \cup T_h) \backslash \{s_h\}$; or

(ii) $\hat{C}$ contains neither s_h nor t_h, and C is the fundamental circuit associated with the base $\hat{C}\backslash\{p\}$ and either s_h or t_h with $s_i \notin C^+$ and C positive on $(S_i \cup T_i)\backslash\{s_i\}$.

Note that in each case the ray C is sign-reversing and positive on T. Hence we obtain the following proposition.

PROPOSITION 7.7. *If* (b) *holds there are no rays.*

LEMMA 7.8. *Consider the graph whose nodes are all i-distinguished circuits of* $\hat{\mathcal{M}}$, $1 \leqq i \leqq n+1$, *with adjacency as defined in Definition* 7.5. *If* $\hat{C}_1$ *is positive or the endpoint of a ray, it has degree* 0, *and otherwise its degree is* 1. *If a node* (*not* $\hat{C}_1$) *is* $(n+1)$*-distinguished, it is either the endpoint of a ray with degree* 0 *or it has degree* 1. *All other nodes are either endpoints of rays, with degree* 0 *or* 1, *or have degree* 2.

Proof. First consider the initial circuit $\hat{C}_1$. If $\hat{C}_1$ is positive, it can be adjacent to no other $(n+1)$-distinguished circuit (which would necessarily be contained in $S \cup \{p\}$) nor to any j-distinguished circuit with $j \leqq n$, since $s_j \in \hat{C}_1^+$. Hence it must have degree 0 in this case. Suppose it is i-distinguished, $i \leqq n$. By the argument above, it can be adjacent to no j-distinguished circuit with $j<i$ $(s_j \in \hat{C}_1^+)$. If it is adjacent to a j-distinguished circuit, $j \geqq i$, then this must be contained in $\hat{C}_1 \cup \{t_i\}$. Thus let C be the fundamental circuit of $\mathcal{M}$ associated with $\hat{C}_1\backslash\{p\}$ and t_i. If C is positive on $(S_i \cup T_i)\backslash\{s_i\}$ and $s_i \notin C^+$, then $\hat{C}_1$ is the endpoint of the ray C, and it is easy to see that $\hat{C}_1$ has degree 0. Otherwise, $\hat{C}_1$ and C differ in sign somewhere in $S_i \cup T_i$. Let $\hat{C}_{1(i)}$ and $C_{(i)}$ result from removing elements $s_{i+1}, s_{i+2}, \cdots, s_n$ from $\hat{C}_1$ and C. Then $\hat{C}_{1(i)}$ and $C_{(i)}$ are circuits in $\hat{\mathcal{M}}_{(i)}$ (in fact, fundamental circuits with respect to the base S_i) which disagree in sign somewhere. Hence we may apply the unique pivot Theorem 4.2 to $\hat{C}_{1(i)}$ and $C_{(i)}$ to get a circuit $\hat{C}_{2(i)}$. The corresponding circuit $\hat{C}_2$ of $\hat{\mathcal{M}}$ is i-distinguished if it contains s_i and j-distinguished for some $j>i$ otherwise. Moreover, $\hat{C}_2$ is adjacent to $\hat{C}_1$ and is the unique distinguished circuit adjacent to $\hat{C}_1$.

Next consider an $(n+1)$-distinguished circuit $\hat{C} \neq \hat{C}_1$. Since $\hat{C} \cap T \neq \varnothing$, let t_i be the last t_j in $\hat{C}$ and let C be the negative of the fundamental circuit associated with the base $\hat{C}\backslash\{p\}$ and s_i. If C is positive on $(S_i \cup T_i)\backslash\{s_i\}$, then $\hat{C}$ is the endpoint of the ray C and one can check that it has degree 0. Otherwise, define $\hat{C}_{(i)}$ and $C_{(i)}$ as above, and apply the unique pivoting Theorem 4.2 to $\hat{C}_{(i)}$ and $C_{(i)}$ in $\hat{\mathcal{M}}_{(i)}$ to get $\hat{C}'_{(i)}$. Then the corresponding circuit $\hat{C}'$ of $\hat{\mathcal{M}}$ is the unique distinguished circuit adjacent to $\hat{C}$.

If $\hat{C}$ is i-distinguished and complementary, where $i<n+1$ and $\hat{C} \neq \hat{C}_1$, then a combination of the arguments above shows that $\hat{C}$ is either the endpoint of (at least) one ray, with degree 0 or 1, or has degree 2.

Finally, suppose $\hat{C}$ is i-distinguished but not complementary, and let $h<i$ be chosen so that $\hat{C}$ contains neither s_h nor t_h. It is easy to see that any adjacent distinguished circuit must contain either s_h or t_h but not both. Thus let C_s and C_t be the fundamental circuits associated with the base $\hat{C}\backslash\{p\}$ and s_h and t_h respectively. If $C=C_s$ or $C=C_t$ satisfies $s_i \notin C^+$ with C positive on $(S_i \cup T_i)\backslash\{s_i\}$ then $\hat{C}$ is the endpoint of (at least) one ray. In this case, $\hat{C}$ has degree 0 or 1. Otherwise, by defining $\hat{C}_{(i)}$, $C_{s(i)}$ and $C_{t(i)}$ as above and performing pivots in $\mathcal{M}_{(i)}$, we find exactly two adjacent distinguished circuits. This completes the proof.

PROPOSITION 7.9. *There is a unique maximal chain* $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_k$ *of distinct distinguished circuits of* $\hat{\mathcal{M}}$ *with each consecutive pair adjacent. Moreover,* $\hat{C}_k$ *is either the endpoint of a ray or a positive complementary circuit involving p, and in the latter case, for each* $1 \leqq i < n$, *there is some l with* $\hat{C}_l\backslash\{s_{i+1}, s_{i+2}, \cdots, s_n\}$ *a positive complementary circuit of* $\hat{\mathcal{M}}_{(i)}$ *involving p.*

Proof. By Lemma 7.8 the graph of distinguished circuits of $\hat{\mathcal{M}}$ is a disjoint union of paths and loops, and $\hat{C}_1$ is the endpoint of a path (possibly trivial). Let the nodes on this path be $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_k$. This is clearly the unique maximal chain. Since $\hat{C}_k$ has degree 1 (0 if $k=1$), it must be the endpoint of a ray or $(n+1)$-distinguished, i.e. positive and complementary and containing p. Finally, we may take for $\hat{C}_l$ the first $\hat{C}_m$ that is j-distinguished for some $j>i$.

Theorem B follows from Lemmas 7.2–7.4 and Propositions 7.7 and 7.9.

Another corollary of Propositions 7.7 and 7.9 is that, when (b) holds and p is "nondegenerate", i.e., lies in no circuit of size smaller than $n+1$, then there is an odd number of positive complementary circuits involving p. This follows from the fact that any graph has an even number of nodes with odd degree.

Clearly, the algorithm to seek such a positive complementary circuit is to trace the sequence $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_k$ of distinguished circuits. The proof of Lemma 7.8 shows how each is obtained from its predecessor by a pivot. This is precisely the extension of the algorithm of Van der Heyden [26] to the oriented matroid setting. The algorithm terminates either with the endpoint of a ray or with a solution. If (b) holds, the latter must occur.

Note that it is possible for the algorithm to proceed from an i-distinguished circuit to a j-distinguished circuit, where $j<i$, i.e., it need not be monotonic. If $\mathcal{M}$ satisfies any of the conditions (3.2b)–(3.2d′), however, the algorithm will be monotonic; this follows from the uniqueness of the solution to all subproblems—see the proof of Proposition 6.9, which shows that regression cannot take place for n-distinguished circuits. Given that regression does not take place, our algorithm can also be viewed as an extension of the principal pivoting algorithm of Cottle and Dantzig [6], in which the possibly negative variables $w_1, w_2, \cdots, w_n$ are made nonnegative in that order. However, our algorithm has a wider validity, in that it requires only condition (3.4b) or (3.4c) and allows regression.

Another algorithm that can be extended is that of Lemke [18]. For this we are given an extension $\hat{\mathcal{M}}$ of $\mathcal{M}$ to $\hat{E}=E\cup\{p\}$. It is not necessary to assume that p is nondegenerate. Let $\tilde{\mathcal{M}}$ be an extension of $\hat{\mathcal{M}}$ to $\tilde{E}=\hat{E}\cup\{q\}$, $q\notin\hat{E}$, such that every circuit of $\tilde{\mathcal{M}}$ involving q contains at least $n+1$ elements and that $S\cup\{q\}$ is the underlying set of a positive circuit. For example, we can let $q=-\text{lex}\,(s_1, s_2, \cdots, s_n)$. Now we call a circuit $\tilde{C}$ of $\tilde{\mathcal{M}}$ *special* if it is positive, complementary and includes p. We can find one as follows. Let $\tilde{C}_p$ and $\tilde{C}_q$ be the fundamental circuits associated with the base S and p and q. Then $\tilde{C}_q$ is positive. If $\tilde{C}_p$ is positive, it is special and moreover is a positive complementary circuit involving p in $\hat{\mathcal{M}}$, so we are done. Otherwise, $\tilde{C}_p^-\cap\tilde{C}_q^+\neq\varnothing$, and we may apply the unique pivot Theorem 4.2 to $\tilde{C}_q$ and $\tilde{C}_p$ to obtain a special circuit $\tilde{C}_1$. Note that, for some i, $\underline{\tilde{C}}_1\cap\{s_i, t_i\}=\varnothing$. Let $\tilde{B}_1$ be the base $\underline{\tilde{C}}_1\backslash\{p\}$ and $\tilde{C}$ the fundamental circuit associated with $\tilde{B}_1$ and t_i. If $\tilde{C}$ is positive then $\tilde{C}$ and $\tilde{C}_q$ show that (c) is violated. Otherwise we may apply the unique pivot results to $\tilde{C}_1$ and $\tilde{C}$ to obtain $\tilde{C}_2$. If $q\notin\underline{\tilde{C}}_2$ we are done; otherwise there is some j with $\underline{\tilde{C}}_2\cap\{s_j, t_j\}=\varnothing$. One of s_j, t_j just left $\underline{\tilde{C}}_2$; we proceed to bring in the other. Thus we may continue to generate a sequence $\tilde{C}_1, \tilde{C}_2, \cdots, \tilde{C}_k$ of special circuits; as long as (c) holds we must terminate with a special circuit not containing q, which is our desired circuit $\hat{C}$. More formally,

DEFINITION 7.10. A circuit $\tilde{C}$ of $\tilde{\mathcal{M}}$ is *special* if it is positive and complementary and contains p. Two such are *adjacent* if the union of their underlying sets is complementary. $\tilde{C}_k$ is an *endpoint of a q-ray* $\tilde{C}$ if both are positive circuits of $\tilde{\mathcal{M}}$, $p\in\underline{\tilde{C}}_k\backslash\underline{\tilde{C}}$, $\underline{\tilde{C}}\cap T\neq\varnothing$ and $\underline{\tilde{C}}_k\cup\underline{\tilde{C}}$ is complementary.

One can prove:

PROPOSITION 7.11. *If* (c) *holds there are no q-rays.*

LEMMA 7.12. *Consider the graph whose nodes are all special circuits of* $\hat{\mathcal{M}}$ *with adjacency as defined in Definition* 7.10. *If* $\tilde{C}_p^-\neq\varnothing$, *then node* $\tilde{C}_1$ *is the endpoint of a*

q-ray or has degree 1, *and every other node containing q is the endpoint of a q-ray or has degree* 2.

PROPOSITION 7.13. *Suppose $\tilde{C}_p^- \neq \emptyset$. Then there is a chain $\tilde{C}_1, \tilde{C}_2, \cdots, \tilde{C}_k$ of distinct special circuits of $\tilde{\mathcal{M}}$ with each consecutive pair adjacent and $\tilde{C}_k$ the endpoint of a q-ray or a positive complementary circuit of $\hat{\mathcal{M}}$ containing p.*

The proofs are similar to those of Propositions 7.7–7.9 and are omitted. Note that, if (c) holds, then $\tilde{C}_k$ is our desired circuit. If, furthermore, p is nondegenerate then any special circuit that does not contain q has degree 1 (we must force in q) and we can stipulate that the chain in Proposition 7.13 is unique and maximal; moreover, Lemma 7.12 implies that there is an odd number of positive complementary circuits of $\hat{\mathcal{M}}$ involving p in this case.

Clearly, the algorithm to find such a circuit is to trace the sequence $\tilde{C}_1, \tilde{C}_2, \cdots, \tilde{C}_k$ of special circuits. It terminates either with the endpoint of a q-ray or with a solution, and if (c) holds the latter must occur. This method is a generalization of Lemke's algorithm [18]; here the element q indexes the column $-d$, where $w = Az + b$ is changed to $w = dz_0 + Az + b$ to assure feasibility, and d is positive. The two algorithms seem to be rather different, prompting Cottle [5] to ask for the classes of linear complementarity problem which Van der Heyden's algorithm will process (i.e., find a solution or demonstrate that none exists); much work has been done on this question for Lemke's algorithm, see for instance Eaves [8]. In fact, we will show below that when it succeeds or a strong nondegeneracy condition holds, Van der Heyden's algorithm is just a special case of Lemke's, corresponding to the choice $d = (\delta^n, \delta^{n-1}, \cdots, \delta)$ for all sufficiently small positive δ. Note that such a d corresponds to an extension $\tilde{\mathcal{M}}$ of $\hat{\mathcal{M}}$ by $q = -\text{lex}\,(s_n, s_{n-1}, \cdots, s_1)$.

THEOREM 7.14. *Let $\hat{\mathcal{M}}$ be an extension of $\mathcal{M}$ to $\hat{E}$, where p lies in no circuit of $\hat{\mathcal{M}}$ of size smaller than $n+1$. Let $q = -\text{lex}\,(s_n, s_{n-1}, \cdots, s_1)$ extend $\hat{\mathcal{M}}$ to $\tilde{\mathcal{M}}$. Pick $1 \leqq i \leqq n$ and let $G \subseteq S_{i-1} \cup T_i$ be complementary. Then* (i) *implies* (ii) *below. Moreover, if $\mathcal{M}$ has no almost complementary circuit of size smaller than $n+1$, the converse is true.*

(i) *There is a circuit $\hat{C}$ of $\hat{\mathcal{M}}$ with $p \in \hat{C}^+$, $s_i \in \hat{C}^- \subseteq \{s_i, s_{i+1}, \cdots, s_n\} \subseteq \underline{\hat{C}}$, $\underline{\hat{C}} \setminus \{s_i\}$ complementary and $\underline{\hat{C}} \cap (S_{i-1} \cup T_i) = G$.*

(ii) *There is a positive complementary circuit $\tilde{C}$ of $\tilde{\mathcal{M}}$ with $p \in \tilde{C}^+$, $q \in \tilde{C}^+$, $s_i \notin \tilde{C}^+$, $\{s_{i+1}, s_{i+2}, \cdots, s_n\} \subseteq \tilde{C}^+$ and $\underline{\tilde{C}} \cap (S_{i-1} \cup T_i) = G$.*

Proof. Suppose (i) holds. Then $B = \underline{\hat{C}} \setminus \{p\}$ is a base of $\hat{\mathcal{M}}$, hence of $\tilde{\mathcal{M}}$. Let $\tilde{C}_q$ be the fundamental circuit of $\tilde{\mathcal{M}}$ associated with B and q, and let $\tilde{D}$ be the fundamental cocircuit of $\tilde{\mathcal{M}}$ associated with the cobase $\tilde{E} \setminus B$ and s_i. Then s_i is the last s_j in $\tilde{D}$, so $q \in \tilde{D}^-$. Thus by orthogonality, $s_i \in \tilde{C}_q^+$. We may therefore apply the signed elimination property to $\hat{C}$ and $\tilde{C}_q$, eliminating s_i, to get a circuit $\tilde{C}$. Thus $p \in \tilde{C}^+$, $q \in \tilde{C}^+$ and $\underline{\tilde{C}}$ contains exactly the elements desired. It remains to show that it is positive.

Consider s_j for $j > i$. Since $\tilde{C} \setminus \{q\}$ is a base, there is a cocircuit $\tilde{D}$ meeting $\tilde{C}$ in just q and s_j. By choice of q, q and s_j appear in $\tilde{D}$ with opposite signs, hence in $\tilde{C}$ with similar signs. Thus $s_j \in \tilde{C}^+$.

Next consider $e \in \underline{\tilde{C}} \cap (S_{i-1} \cup T_i) = G$. There is a cocircuit $\tilde{D}$ meeting $\tilde{C}$ in just q and e. Thus $\tilde{D}$ meets $\hat{C}$ in just s_i and e. By orthogonality, s_i and e appear in $\tilde{D}$ with the same sign. By definition of q, s_i and q appear in $\tilde{D}$ with opposite signs. Thus q and e appear in $\tilde{D}$ with opposite signs, hence in $\tilde{C}$ with the same sign. We have established that $\tilde{C}$ is positive, i.e. (ii) holds.

Conversely assume that (ii) holds and that $\mathcal{M}$ has no almost complementary circuit of size smaller than $n+1$. Then $\tilde{B} = \tilde{C} \setminus \{p\}$ is a base of $\tilde{\mathcal{M}}$. Let $\tilde{C}_i$ be the fundamental circuit associated with $\tilde{B}$ and s_i and note that the nondegeneracy condition implies that $q \in \underline{\tilde{C}}_i$. Let $\tilde{D}$ be the fundamental cocircuit associated with the cobase $\tilde{E} \setminus \tilde{B}$ and

q. Then by orthogonality of $\tilde{C}_i$ and $\tilde{D}$, $s_i \in \underline{\tilde{D}}$; hence $s_i \in \tilde{D}^-$ and so $q \in \tilde{C}_i^+$. Now apply the signed elimination property to $\tilde{C}$ and $-\tilde{C}_i$, eliminating to q, to get $\hat{C}$. We have $p \in \hat{C}^+$, $s_i \in \hat{C}^-$ and $\underline{\hat{C}}$ contains exactly the desired elements. We must show that $\hat{C}$ is positive on $S_{i-1} \cup T_i$.

Choose $e \in \underline{\hat{C}} \cap (S_{i-1} \cup T_i) = G$. There is a cocircuit $\tilde{D}$ meeting $\hat{C}$ in just e and s_i. Thus $\tilde{D}$ meets $\tilde{C}$ in just e and q, and hence $\tilde{D}$ contains e and q with opposite signs. By definition of q, $\tilde{D}$ contains q and s_i with opposite signs, hence e and s_i with the same sign. Now orthogonality of $\hat{C}$ and $\tilde{D}$ gives $e \in \hat{C}^+$ as desired. This completes the proof.

If we extend the correspondence of distinguished circuits of $\hat{\mathcal{M}}$ and special circuits of $\tilde{\mathcal{M}}$ in the theorem so that positive complementary circuits of $\hat{\mathcal{M}}$ containing p correspond to themselves, we obtain by a straightforward argument the following result.

COROLLARY 7.15. *Let $\mathcal{M}$, $\hat{\mathcal{M}}$ and $\tilde{\mathcal{M}}$ be as in the theorem. Then the chain of distinguished circuits in $\hat{\mathcal{M}}$ in Proposition* 7.9 *corresponds to an initial segment (possibly all) of the chain of special circuits of $\tilde{\mathcal{M}}$ in Proposition* 7.13. *If $\mathcal{M}$ has no almost complementary circuit of size smaller than $n+1$, then the chains correspond exactly.*

Note, however, that it is possible (if the latter condition does not hold) for the distinguished-circuit algorithm (Van der Heyden's) to fail while the special-circuit algorithm (Lemke's) succeeds. For an example, take the matroids for the linear complementarity problem with $A = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, $b = \left(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right)$.

To conclude, we note that the concept of orientation in complementary pivot algorithms extends to the oriented matroid setting—see Eaves and Scarf [9], Todd [23] and Van der Heyden [26]. Indeed, the general proofs of [23] apply directly, and one obtains:

THEOREM 7.16. *Let the positive complementary circuit $\hat{C}$ of $\hat{\mathcal{M}}$ be generated by one of the two algorithms above, and suppose $\hat{C}$ contains $n+1$ elements. Then the complementary base $U = \underline{\hat{C}} \backslash \{p\}$ satisfies $\varepsilon(\upsilon) = (-1)^{|U \cap T|}$, where ε is the orientation of bases that satisfies $\varepsilon(s_1, s_2, \cdots, s_n) = 1$ and υ is the natural ordering $(u_1, u_2, \cdots, u_n)$ of U with $u_i \in \{s_i, t_i\}$ for each i.*

REFERENCES

[1] R. G. BLAND, *Complementary orthogonal subspace of R^n and orientability of matroids*, Ph.D. Thesis, Cornell University, Ithaca, NY, 1974.

[2] ———, *A combinatorial abstraction of linear programming*, J. Comb. Thy (B), 23 (1977), pp. 33–57.

[3] ———, *Linear programming duality and Minty's lemma*, Technical Report No. 449, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1980.

[4] R. G. BLAND AND M. LAS VERGNAS, *Orientability of matroids*, J. Comb. Thy (B), 24 (1978), pp. 94–123.

[5] R. W. COTTLE, *Completely-Q matrices*, Math. Programming, 19 (1980), pp. 347–351.

[6] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Linear Algebra and Appl., 1 (1968), pp. 103–125.

[7] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, NJ, 1963.

[8] B. C. EAVES, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 612–634.

[9] B. C. EAVES AND H. SCARF, *The solution of systems of piecewise linear equations*, Math. Oper. Res., 1 (1976), pp. 1–27.

[10] J. FOLKMAN AND J. LAWRENCE, *Oriented matroids*, J. Comb. Thy (B), 25 (1978), pp. 199–236.

[11] K. FUKUDA, *Oriented matroid programming*, Ph.D. Thesis, Univ. Waterloo, Waterloo, Ontario, 1981.

[12] D. GALE AND H. NIKAIDO, *The Jacobian matrix and global univalence of mappings*, Math. Ann., 159 (1965), pp. 81–93.

[13] A. W. INGLETON, *A problem in linear inequalities*, Proc. London Math. Soc., 16 (1966), pp. 519–536.

[14] M. LAS VERGNAS, *Matroides orientables*, preprint, April 1974, announced in C.R. Acad. Sci. Paris, 280 (1975), pp. 61–64.

[15] M. LAS VERGNAS, *Extensions ponctuelles d'une géometrie combinatoire orientée*, in Problèmes combinatoires et théorie des graphes, Actes du Colloque International C.N.R.S., No. 260, Orsay 1976, Paris, 1978, pp. 263–268.

[16] ———, *Bases in oriented matroids*, J. Comb. Thy (B), 25 (1978), pp. 283–289.

[17] J. LAWRENCE, *Oriented matroids*, Ph.D. Thesis, Univ. Washington, Seattle, WA, 1975.

[18] C. E. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.

[19] C. E. LEMKE AND J. T. HOWSON, JR., *Equilibrium points of bimatrix games*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 413–423.

[20] G. J. MINTY, *On the axiomatic foundations of the theories of directed linear graphs, electrical networks, and network programming*, J. Math. Mech., 15 (1966), pp. 485–520.

[21] K. G. MURTY, *On the number of solutions to the complementarity problem and spanning properties of complementary cones*, Linear Algebra and Appl., 5 (1972), pp. 65–108.

[22] H. SAMELSON, R. M. THRALL AND O. WESLER, *A partition theorem for Euclidean n-space*, Proc. Amer. Math. Soc., 9 (1958), pp. 805–807.

[23] M. J. TODD, *Orientation in complementary pivot algorithms*, Math. of Operations Res., 1 (1976), pp. 54–66.

[24] ———, *Linear and quadratic programming in oriented matroids*, Technical Report No. 565, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1983.

[25] A. W. TUCKER, *Principal pivotal transforms of square matrices*, SIAM Rev., 5 (1963), p. 305.

[26] L. VAN DER HEYDEN, *A variable dimension algorithm for the linear complementarity problem*, Math. Programming, 19 (1980), pp. 328–346.

[27] H. WHITNEY, *On the abstract properties of linear dependence*, Amer. J. Math., 57 (1935), pp. 507–553.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

004

NONTESTABLE HYPOTHESES IN LINEAR MODELS*

SHAYLE R. SEARLE†, WILLIAM H. SWALLOW‡ AND CHARLES E. McCULLOCH†

Abstract. Nontestable hypotheses in linear models are formally defined and proof given that they cannot be tested. Consequences of carrying out calculations as if they were testable are considered.

AMS(MOS) subject classifications. 62F03, 15A09, 15A03

1. Introduction. We represent the general linear model under normality as

$$\mathbf{y} \sim N(\mathbf{Xb}, \sigma^2 \mathbf{I}_N) \tag{1}$$

where $\mathbf{y}$ is a vector of N observations with mean $\mathbf{Xb}$ and dispersion matrix $\sigma^2\mathbf{I}_N$, where $\mathbf{I}_N$ is an identity matrix of order N. Estimation of the parameter vector $\mathbf{b}$ is often made by using the principle of least squares, or using maximum likelihood, both of which lead in the case of (1) to what are called the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b}^0 = \mathbf{X}'\mathbf{y}$. A solution of them is taken as

$$\mathbf{b}^0 = \mathbf{G}\mathbf{X}'\mathbf{y} \tag{2}$$

for $\mathbf{G}$ being a generalized inverse of $\mathbf{X}'\mathbf{X}$, meaning that it satisfies

$$\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}, \quad \text{from which } \mathbf{X} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}. \tag{3}$$

$\mathbf{G}'$ is also a generalized inverse of $\mathbf{X}'\mathbf{X}$. If $\mathbf{G}$ satisfies both (3) and

$$\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G} = \mathbf{G} \tag{4}$$

it is said to be a reflexive generalized inverse; if $\mathbf{G}$ is not reflexive then

$$\mathbf{G}^* = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G} \tag{5}$$

is.

2. Defining testable hypotheses. A linear function $\mathbf{q}'\mathbf{b}$ of the elements of the parameter vector $\mathbf{b}$ is said to be an estimable function when $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ for some $\mathbf{t}'$, i.e., when $\mathbf{q}'$ lies in the row space of $\mathbf{X}$. The best linear unbiased estimator of the estimable function $\mathbf{q}'\mathbf{b}$ is $\mathbf{q}'\mathbf{b}^0$, a function of $\mathbf{b}^0$ that is invariant to whatever $\mathbf{G}$ is used for $\mathbf{b}^0$ in (2), i.e., to whatever particular solution, $\mathbf{b}^0$, of the normal equations is used.

A linear hypothesis concerning elements of $\mathbf{b}$ is defined as

$$H: \mathbf{K}'\mathbf{b} = \mathbf{m} \tag{6}$$

for a known matrix $\mathbf{K}'$ and vector $\mathbf{m}$. It will be assumed throughout that $\mathbf{K}'$ is of full row rank. There is no loss of generality in making this assumption (Scheffé (1959, p. 29)) and it is also a practical requirement. $\mathbf{K}'$ is derived from the statistical problem at hand and redundant or inconsistent equations would not normally be specified as part of any hypothesis of interest.

When all elements of $\mathbf{K}'\mathbf{b}$ are estimable, $\mathbf{K}'\mathbf{b}$ is said to be estimable and

$$\mathbf{K}' = \mathbf{T}'\mathbf{X} \tag{7}$$

for some $\mathbf{T}'$. The hypothesis (6) is said to be testable when $\mathbf{K}'\mathbf{b}$ is estimable. The

* Received by the editors May 13, 1983, and in revised form August 9, 1983.

† Biometrics Unit, Cornell University, Ithaca, New York 14853.

‡ Department of Statistics, North Carolina State University, Raleigh, North Carolina 27650.

F-statistic for testing (6) is well known to be

$$F(H)=Q/r_{\mathbf{K}}\hat{\sigma}^2 \tag{8}$$

where

$$Q=(\mathbf{K}'\mathbf{b}^0-\mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0-\mathbf{m}), \tag{9}$$

$$r_{\mathbf{K}}=\text{rank of } \mathbf{K}'=\text{number of rows in } \mathbf{K}', \tag{10}$$

and

$$\hat{\sigma}^2=\mathbf{y}'(\mathbf{I}-\mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}/(N-r_{\mathbf{X}}). \tag{11}$$

All of the preceding development is well known. What is not so well known is the treatment of the hypothesis H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ when not all elements of $\mathbf{K}'\mathbf{b}$ are estimable, i.e., when $\mathbf{K}'\mathbf{b}$ is not estimable. Generally speaking we simply say that such a hypothesis is not testable and leave it at that. The purpose of this paper is twofold. First, nontestable and partially testable hypotheses are formally defined and proof given that a nontestable hypothesis cannot be tested. Second, in cases where $\mathbf{K}'$ and $\mathbf{G}$ are such that $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ exists, thus enabling Q of (9) to be calculated and used in $F(H)$ in (8), it is shown what *is* being tested.

3. The general procedure for hypothesis testing. The residual sum of squares after fitting the model (1) is well known to be

$$SSE=(\mathbf{y}-\mathbf{X}\mathbf{b}^0)'(\mathbf{y}-\mathbf{X}\mathbf{b}^0). \tag{12}$$

To derive the residual sum of squares under the hypothesis we fit the model

$$\mathbf{y}\sim N(\mathbf{X}\mathbf{b},\sigma^2\mathbf{I}) \quad\text{and}\quad \mathbf{K}'\mathbf{b}=\mathbf{m}. \tag{13}$$

Using Lagrange multipliers $2\boldsymbol{\theta}$ to account for $\mathbf{K}'\mathbf{b}=\mathbf{m}$ in minimizing $(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$ subject to $\mathbf{K}'\mathbf{b}=\mathbf{m}$, leads to the well-known equations

$$\mathbf{X}'\mathbf{X}\mathbf{b}_H^0+\mathbf{K}\boldsymbol{\theta}=\mathbf{X}'\mathbf{y} \tag{14}$$

and

$$\mathbf{K}'\mathbf{b}_H^0=\mathbf{m} \tag{15}$$

for $\mathbf{b}_H^0$ and $\boldsymbol{\theta}$. The error sum of squares under the hypothesis is then

$$SSE_H=(\mathbf{y}-\mathbf{X}\mathbf{b}_H^0)'(\mathbf{y}-\mathbf{X}\mathbf{b}_H^0), \tag{16}$$

and the F-statistic for testing H is

$$F=(SSE_H-SSE)/r_{\mathbf{K}}\hat{\sigma}^2. \tag{17}$$

4. Testable hypotheses. H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ is defined as testable when $\mathbf{K}'\mathbf{b}$ is estimable. A solution to the minimizing equations (14) and (15) is then (see, e.g., Searle (1971, p. 191))

$$\boldsymbol{\theta}=(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0-\mathbf{m}) \tag{18}$$

and

$$\mathbf{b}_H^0=\mathbf{G}\mathbf{X}'\mathbf{y}-\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0-\mathbf{m}) \tag{19}$$

where, when $\mathbf{K}'$ has full row rank and $\mathbf{K}'\mathbf{b}$ is estimable, $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ always exists, ensuring that $\mathbf{b}_H^0$ of (19) and Q of (9) can be calculated. On substituting $\mathbf{b}_H^0$ into (16) it will be found that

$$SSE_H = SSE + (\mathbf{K}'\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m}).$$

Hence for (17)

$$SSE_H - SSE = (\mathbf{K}'\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m}),$$

i.e.,

$$SSE_H - SSE = Q \tag{20}$$

for Q of (9). Thus, when H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is testable, F of (17) is $F(H)$ introduced in (8).

5. Nontestable hypotheses.

5.1. Definition. We define a linear hypothesis H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ as being nontestable when every element, and every linear combination of elements, of $\mathbf{K}'\mathbf{b}$ is nonestimable. For example, in the case where $\mathbf{b} = [\mu \;\; \alpha_1 \;\; \alpha_2 \;\; \alpha_3]'$ as in the subsequent example (§ 6), this precludes describing

$$H: \begin{cases} \alpha_1 = 7 \\ \alpha_2 = 4 \end{cases} \tag{21}$$

as nontestable in a situation where α_1 and α_2 are each nonestimable, but $\alpha_1 - \alpha_2$ is estimable.

In our definition of a nontestable hypothesis the nonestimability of elements of $\mathbf{K}'\mathbf{b}$ means that $\mathbf{K}'$ must be such that there is no $\mathbf{T}'$ such that $\mathbf{K}' = \mathbf{T}'\mathbf{X}$; and, more generally, nonestimability of linear combinations of elements of $\mathbf{K}'\mathbf{b}$ means that there must be *no* $\mathbf{R}$ and $\mathbf{T}'$ other than both null such that

$$\mathbf{R}\mathbf{K}' = \mathbf{T}'\mathbf{X}. \tag{22}$$

The reason for defining a nontestable hypothesis this way is that otherwise a hypothesis such as (21), written in terms of nonestimable functions, could be rewritten as

$$H: \begin{cases} \alpha_1 = 7 \\ \alpha_1 - \alpha_2 = 3 \end{cases}$$

where it is now in terms of a mixture of nonestimable and estimable functions. We call such a mixture a partially testable hypothesis and formally define it in § 5.6.

5.2. A nontestable hypothesis cannot be tested. Equations (14), (15), and (16) apply whether or not H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is testable; when it is, (18) and (19) are a solution and (20) follows.

But when H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is nontestable, we find the solution differs from (18) and (19). To see this consider the following equations in elements of a vector $\mathbf{u}$:

$$\mathbf{K}'(\mathbf{I} - \mathbf{H})\mathbf{u} = \mathbf{m} - \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}, \tag{23}$$

where $\mathbf{H}$ is defined as $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$, with $\mathbf{X}\mathbf{H} = \mathbf{X}$ from (3). First note that when $\mathbf{K}'\mathbf{b}$ is estimable, $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ and $\mathbf{K}'(\mathbf{I} - \mathbf{H}) = \mathbf{T}'(\mathbf{X} - \mathbf{X}\mathbf{H}) = \mathbf{0}$ and equations (23) are inconsistent. More generally, when $\mathbf{K}'$ is such that (22) is true, pre-multiplication of (23) by $\mathbf{R}$ leads to $\mathbf{0} = \mathbf{R}\mathbf{m} - \mathbf{R}\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}$ and again the equations are inconsistent. Indeed, $\mathbf{R}\mathbf{K}'(\mathbf{I} - \mathbf{H}) = \mathbf{0}$, if and only if (22) holds. Since (22) is assumed not to hold, we consider (23).

For $\mathbf{X}$ having order $N \times p$ and rank r, the matrices $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent of order p, and $\mathbf{I} - \mathbf{H}$ has rank $p - r$. Therefore $(\mathbf{I} - \mathbf{H})\mathbf{u}$ in (23) has $p - r$ arbitrary elements. Furthermore, $\mathbf{X}$ has rows of order p and its rank is r; and $\mathbf{K}'$, though having full row rank, has $r_{\mathbf{K}}$ rows of order p which, by (22), have no linear dependencies with the rows of $\mathbf{X}$. Therefore $r_{\mathbf{K}} \leqq p-r$. Hence (23) consists of no more than $p-r$ equations in $p-r$ unknowns, the $p-r$ arbitrary elements of $(\mathbf{I}-\mathbf{H})\mathbf{u}$. Thus (23) has a solution when H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ is nontestable.

Now consider equations (14) and (15) for determining $\mathbf{b}_H^0$, when H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ is nontestable. It is easily shown that they are satisfied by

$$\boldsymbol{\theta}=\mathbf{0} \tag{24}$$

and

$$\mathbf{b}_H^0=\mathbf{G}\mathbf{X}'\mathbf{y}+(\mathbf{I}-\mathbf{H})\mathbf{u} \tag{25}$$

for $\mathbf{u}$ being a solution of (23), i.e.,

$$\mathbf{X}'\mathbf{X}\mathbf{b}_H^0+\mathbf{K}\boldsymbol{\theta}=\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}(\mathbf{I}-\mathbf{H})\mathbf{u}+\mathbf{K}\mathbf{0}=\mathbf{X}'\mathbf{y}+\mathbf{0}+\mathbf{0}=\mathbf{X}'\mathbf{y},$$

thus satisfying (14), and, using (23),

$$\mathbf{K}'\mathbf{b}_H^0=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}+\mathbf{K}'(\mathbf{I}-\mathbf{H})\mathbf{u}=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}+\mathbf{m}-\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}=\mathbf{m},$$

which satisfies (15).

We now use (25) in (16); because $\mathbf{X}=\mathbf{X}\mathbf{H}$ we have $\mathbf{X}\mathbf{b}_H^0=\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}=\mathbf{X}\mathbf{b}^0$ and so in (16)

$$SSE_H=(\mathbf{y}-\mathbf{X}\mathbf{b}^0)'(\mathbf{y}-\mathbf{X}\mathbf{b}^0)=SSE \text{ of } (12).$$

Therefore

$$SSE_H-SSE=0$$

and so F of (17) is zero, i.e., there is no test of H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ for nontestable hypotheses.

5.3. Alternative forms for $\mathbf{b}_H^0$. To eliminate $\mathbf{u}$ from $\mathbf{b}_H^0$ of (25) we solve (23) as

$$\mathbf{u}=[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-(\mathbf{m}-\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y})$$

and get

$$\mathbf{b}_H^0=\mathbf{G}\mathbf{X}'\mathbf{y}+(\mathbf{I}-\mathbf{H})[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-(\mathbf{m}-\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}) \tag{26}$$

where, in general, $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$. Rearranging (26) gives

$$\mathbf{b}_H^0=\{\mathbf{G}-(\mathbf{I}-\mathbf{H})[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-\mathbf{K}'\mathbf{G}\}\mathbf{X}'\mathbf{y}+(\mathbf{I}-\mathbf{H})[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-\mathbf{m}. \tag{27}$$

Define the coefficient of $\mathbf{X}'\mathbf{y}$ in (27) as $\mathbf{G}_r$:

$$\mathbf{G}_r=\mathbf{G}-(\mathbf{I}-\mathbf{H})[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-\mathbf{K}'\mathbf{G}. \tag{28}$$

Then, because $\mathbf{X}(\mathbf{I}-\mathbf{H})=\mathbf{0}$, it is easily seen that $\mathbf{G}_r$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Furthermore, for

$$\mathbf{I}-\mathbf{H}_r=\mathbf{I}-\mathbf{G}_r\mathbf{X}'\mathbf{X},$$

$$(\mathbf{I}-\mathbf{H}_r)(\mathbf{I}-\mathbf{H})=\mathbf{I}-\mathbf{H}-\mathbf{G}_r\mathbf{X}'\mathbf{X}+\mathbf{G}_r\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}=\mathbf{I}-\mathbf{H}$$

so that in (27)

$$\mathbf{b}_H^0=\mathbf{G}_r\mathbf{X}'\mathbf{y}+(\mathbf{I}-\mathbf{H}_r)\{(\mathbf{I}-\mathbf{H})[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-\}\mathbf{m}=\mathbf{G}_r\mathbf{X}'\mathbf{y}+(\mathbf{I}-\mathbf{H}_r)\mathbf{w} \tag{29}$$

for

$$\mathbf{w}=(\mathbf{I}-\mathbf{H})[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-\mathbf{m}. \tag{30}$$

Thus in (29) we have $\mathbf{b}_H^0$ expressed in the general form of any solution to equations $\mathbf{X}'\mathbf{X}\mathbf{b}^0 = \mathbf{X}'\mathbf{y}$; and so we know at once that $\mathbf{X}\mathbf{b}_H^0 = \mathbf{X}\mathbf{b}^0$, and hence $SSE_H = SSE$ and so $F = 0$ as earlier determined.

Another form for $\mathbf{b}_H^0$ is obtained by dealing with equations (14) and (15) directly, after writing them as

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_H^0 \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{m} \end{bmatrix}. \tag{31}$$

Then, on defining

$$\mathbf{L} = (\mathbf{X}'\mathbf{X} + \mathbf{K}\mathbf{K}')^-, \tag{32}$$

which is a generalized inverse of $\mathbf{X}'\mathbf{X}$, it is easily verified that

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{bmatrix}^- = \begin{bmatrix} \mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L} & \mathbf{L}'\mathbf{K} \\ \mathbf{K}'\mathbf{L} & \mathbf{0} \end{bmatrix}. \tag{33}$$

If, without loss of generality, we assume that the first $r_{\mathbf{X}}$ rows of $\mathbf{X}$ are linearly independent and we represent them by $\mathbf{Z}$, then $[\mathbf{Z}' \ \ \mathbf{K}]'$ has full row rank. Augment this with some matrix $\mathbf{M}$ of full row rank chosen so that $[\mathbf{Z}' \ \ \mathbf{K} \ \ \mathbf{M}']'$ is nonsingular. Then, on defining

$$[\mathbf{Z}' \ \ \mathbf{K} \ \ \mathbf{M}']'^{-1} = [\mathbf{A} \ \ \mathbf{B} \ \ \mathbf{C}],$$

it is easily shown that

$$\mathbf{B} = \mathbf{L}\mathbf{K}, \quad \mathbf{X}\mathbf{L}\mathbf{K} = \mathbf{0} \quad \text{and} \quad \mathbf{K}'\mathbf{L}\mathbf{K} = \mathbf{I}. \tag{34}$$

It will then be found that a solution to (31) is

$$\boldsymbol{\theta} = \mathbf{0} \quad \text{and} \quad \mathbf{b}_H^0 = \mathbf{L}\mathbf{X}'\mathbf{y} + \mathbf{L}\mathbf{K}\mathbf{m} + [\mathbf{I} - \mathbf{L}(\mathbf{X}'\mathbf{X} + \mathbf{K}\mathbf{K}')]\mathbf{z} \tag{35}$$

for arbitrary $\mathbf{z}$; and by using (34), $\mathbf{X}\mathbf{b}_H^0 = \mathbf{X}\mathbf{b}^0$, so that again $SSE_H = SSE$ and $F = 0$. In this case a convenient solution for $\mathbf{b}_H^0$ from (35) is

$$\mathbf{b}_H^0 = \mathbf{L}(\mathbf{X}'\mathbf{y} + \mathbf{K}\mathbf{m}) = (\mathbf{X}'\mathbf{X} + \mathbf{K}\mathbf{K}')^-(\mathbf{X}'\mathbf{y} + \mathbf{K}\mathbf{m}). \tag{36}$$

5.4. Comparing solutions for $\mathbf{b}_H^0$. The solution for $\mathbf{b}_H^0$ when H is testable is (19), which can also be written as

$$\mathbf{b}_H^0 = [\mathbf{G} - \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{G}]\mathbf{X}'\mathbf{y} + \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{m}.$$

But in this expression the coefficient of $\mathbf{X}'\mathbf{y}$ is not a generalized inverse of $\mathbf{X}'\mathbf{X}$ as it is in (29) and (36) for nontestable hypotheses. Thus $\mathbf{b}_H^0$ for testable hypotheses does not have the same general form that occurs for nontestable hypotheses. This difference in solution to equations (14) and (15) is further accentuated by noting that for testable hypotheses the solution for $\boldsymbol{\theta}$ is $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m})$ of (18) which is nonnull, whereas for nontestable hypotheses the solution in (24) and (35) is $\boldsymbol{\theta} = \mathbf{0}$.

5.5. $F(H)$ when H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is nontestable. We showed in § 5.2 that when H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is nontestable it cannot be tested. Despite this, there are occasions when, even for $\mathbf{K}'\mathbf{b}$ nonestimable, $\mathbf{G}$ and $\mathbf{K}'$ are such that $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ exists. This is so because, although $\mathbf{K}'$ having full row rank is a necessary condition for $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ to exist, $\mathbf{K}'\mathbf{b}$ being estimable is only a sufficient condition; it is not necessary. Therefore, for $\mathbf{K}'\mathbf{b}$ nonestimable there can exist full row rank matrices $\mathbf{K}'$ with $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ existing and, when this is so, Q of (9) can be computed as can $F(H) = Q/r_{\mathbf{K}}\hat{\sigma}^2$ of (8). We emphasize that one should not be calculating Q and $F(H)$ when H: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is nontestable.

However, in cases where Q and $F(H)$ can be (and perhaps have been) calculated despite H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ being nontestable, the existence of $F(H)$ prompts the question "What hypothesis is $F(H)$ testing?". A partial answer is contained in the following theorem.

THEOREM. *If, when H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ is nontestable $\mathbf{K}'$ and $\mathbf{G}$ are nonetheless such that*

$$Q=(\mathbf{K}'\mathbf{b}^0-\mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0-\mathbf{m}) \tag{37}$$

can be computed, then, provided $\mathbf{G}$ is symmetric and reflexive,

$$F(H)=Q/r_{\mathbf{K}}\hat{\sigma}^2 \text{ is testing } H\text{: } \mathbf{K}'\mathbf{H}\mathbf{b}=\mathbf{m}. \tag{38}$$

Proof. Because $\mathbf{H}=\mathbf{G}\mathbf{X}'\mathbf{X}$, $\mathbf{K}'\mathbf{H}=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}$ and so $\mathbf{K}'\mathbf{H}\mathbf{b}$ is estimable. Therefore, from (20), we can use (8) and (9) to have

$$Q=(\mathbf{K}'\mathbf{H}\mathbf{b}^0-\mathbf{m})'(\mathbf{K}'\mathbf{H}\mathbf{G}\mathbf{H}'\mathbf{K})^{-1}(\mathbf{K}'\mathbf{H}\mathbf{b}^0-\mathbf{m}) \tag{39}$$

with

$$\mathbf{K}'\mathbf{H}\mathbf{b}^0=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y}=\mathbf{K}'\mathbf{b}^0. \tag{40}$$

Similarly,

$$\begin{aligned}\mathbf{K}'\mathbf{H}\mathbf{G}\mathbf{H}'\mathbf{K}&=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{K}\\&=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{K}\\&=\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}\quad \text{for } \mathbf{G}=\mathbf{G}'\\&=\mathbf{K}'\mathbf{G}\mathbf{K}\quad \text{for } \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}=\mathbf{G}.\end{aligned} \tag{41}$$

Substituting (40) and (41) into (39) gives (37) when $\mathbf{G}$ is symmetric and reflexive. Q.E.D.

This theorem corrects an error implicit in Searle (1971, p. 195), which fails to mention the requirements of $\mathbf{G}$ to be symmetric and reflexive.

Of course, the question will be asked "When, for a nontestable hypothesis, will $\mathbf{K}'\mathbf{G}\mathbf{K}$ nevertheless be nonsingular?", thus permitting computation of Q and interpretation thereof by means of the theorem. There seems to be no universal answer to this: nonsingularity of $\mathbf{K}'\mathbf{G}\mathbf{K}$ depends completely on the particular forms of $\mathbf{K}'$ and $\mathbf{G}$ being used. Illustrations of $\mathbf{K}'\mathbf{G}\mathbf{K}$ being nonsingular and being singular are shown in the example at the end of the paper.

5.6. Partially testable hypotheses. As illustrated in § 5.1 there are some hypotheses in which some elements of $\mathbf{K}'\mathbf{b}$ are nonestimable but for which some linear combinations are estimable. To exclude this mixture from the definition of a nontestable hypothesis we defined it as being a hypothesis for which no linear combination of elements of $\mathbf{K}'\mathbf{b}$ is estimable. Since linear combinations of estimable functions are estimable, a similar definition, equivalent to that in § 4, can be made for testable hypotheses: A hypothesis H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ is defined as testable when all linear combinations of elements of $\mathbf{K}'\mathbf{b}$ are estimable. The remaining forms of $\mathbf{K}'\mathbf{b}=\mathbf{m}$ lead naturally to a definition of a partially testable hypothesis: A hypothesis H: $\mathbf{K}'\mathbf{b}=\mathbf{m}$ is defined as partially testable when at least one linear combination of elements of $\mathbf{K}'\mathbf{b}$ is estimable and at least one linear combination of elements of $\mathbf{K}'\mathbf{b}$ is nonestimable. Note that this requires at least one element of $\mathbf{K}'\mathbf{b}$ to be nonestimable. These three definitions provide a complete categorization of the possible forms of $\mathbf{K}'\mathbf{b}$ vis-a-vis estimable and nonestimable elements.

Often a partially testable hypothesis can be partitioned as

$$H\colon \begin{cases} \mathbf{K}_1'\mathbf{b} = \mathbf{m}_1 \\ \mathbf{K}_2'\mathbf{b} = \mathbf{m}_2, \end{cases}$$

where H_1: $\mathbf{K}_1'\mathbf{b} = \mathbf{m}_1$ is testable and H_2: $\mathbf{K}_2'\mathbf{b} = \mathbf{m}_2$ is nontestable. Suppose that $\mathbf{K}'\mathbf{G}\mathbf{K}$ is nonsingular and $\mathbf{G}$ is symmetric and reflexive; then $F(H)$ is testing $\mathbf{K}'\mathbf{H}\mathbf{b} = \mathbf{m}$ which in this case is

$$(42) \qquad H\colon \begin{cases} \mathbf{K}_1'\mathbf{H}\mathbf{b} = \mathbf{m}_1 \\ \mathbf{K}_2'\mathbf{H}\mathbf{b} = \mathbf{m}_2 \end{cases} \text{ equivalent to } H\colon \begin{cases} \mathbf{K}_1'\mathbf{b} = \mathbf{m}_1 \\ \mathbf{K}_2'\mathbf{H}\mathbf{b} = \mathbf{m}_2. \end{cases}$$

This is so because for H_1 testable,

$$\mathbf{K}_1' = \mathbf{T}_1'\mathbf{X} \quad \text{and} \quad \mathbf{K}_1'\mathbf{H} = \mathbf{T}_1'\mathbf{X}\mathbf{H} = \mathbf{T}_1'\mathbf{X} = \mathbf{K}_1'.$$

Result (42) corrects an error implicit in Searle (1971, pp. 194–195) that if Q can be calculated for a partially testable hypothesis then $F(H)$ is testing only the testable part of it. That is not so; it is testing (42).

6. Example. To illustrate the preceding results we use the simple example of a 1-way classification (completely randomized design) given in Searle (1971, p. 165). The model is

$$E(y_{ij}) = \mu + \alpha_i$$

for $i = 1, 2, 3$ with $n_1 = 3$, $n_2 = 2$ and $n_3 = 1$, where y_{ij} is the jth observation in the ith class, μ is a general mean and α_i is the effect of the ith class. Data are as follows:

	Class	
1	2	3
101	84	32
105	88	
94		
300	172	32

Then, with $\mathbf{y} = [101 \;\; 105 \;\; 94 \;\; 84 \;\; 88 \;\; 32]'$ we have

$$\mathbf{X}\mathbf{b} = \begin{bmatrix} 1 & 1 & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot \\ 1 & \cdot & 1 & \cdot \\ 1 & \cdot & 1 & \cdot \\ 1 & \cdot & \cdot & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 504 \\ 300 \\ 172 \\ 32 \end{bmatrix},$$

with a dot in a matrix representing zero. Thus

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix}6&3&2&1\\3&3&\cdot&\cdot\\2&\cdot&2&\cdot\\1&\cdot&\cdot&1\end{bmatrix},\quad \mathbf{G}=\begin{bmatrix}\cdot&\cdot&\cdot&\cdot\\\cdot&\frac{1}{3}&\cdot&\cdot\\\cdot&\cdot&\frac{1}{2}&\cdot\\\cdot&\cdot&\cdot&1\end{bmatrix},\quad \mathbf{b}^0=\mathbf{G}\mathbf{X}'\mathbf{y}=\begin{bmatrix}0\\100\\86\\32\end{bmatrix}$$

$$\text{and}\quad \mathbf{H}=\mathbf{G}\mathbf{X}'\mathbf{X}=\begin{bmatrix}\cdot&\cdot&\cdot&\cdot\\1&1&\cdot&\cdot\\1&\cdot&1&\cdot\\1&\cdot&\cdot&1\end{bmatrix}. \tag{43}$$

And

$$\mathbf{X}\mathbf{b}^0=[100\quad 100\quad 100\quad 86\quad 86\quad 32]'.$$

The functions $\mu+\alpha_i$ for $i=1$, 2, 3 are estimable, as are linear combinations of them; and, for example, the numerator sum of squares, namely Q of (9), for testing the testable hypothesis

$$H\colon\begin{cases}\alpha_1-\alpha_2=0\\\alpha_2-\alpha_3=0\end{cases}$$

involves

$$\mathbf{K}'\mathbf{b}^0=\begin{bmatrix}0&1&-1&0\\0&0&1&-1\end{bmatrix}\begin{bmatrix}0\\100\\86\\32\end{bmatrix}=\begin{bmatrix}14\\54\end{bmatrix}$$

and

$$(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}=\left\{\begin{bmatrix}0&1&-1&0\\0&0&1&-1\end{bmatrix}\begin{bmatrix}\cdot&\cdot&\cdot&\cdot\\\cdot&\frac{1}{3}&\cdot&\cdot\\\cdot&\cdot&\frac{1}{2}&\cdot\\\cdot&\cdot&\cdot&1\end{bmatrix}\begin{bmatrix}0&0\\1&0\\-1&1\\0&-1\end{bmatrix}\right\}^{-1}$$

$$=\begin{bmatrix}\frac{5}{6}&-\frac{1}{2}\\-\frac{1}{2}&1\frac{1}{2}\end{bmatrix}^{-1}=\tfrac{1}{6}\begin{bmatrix}9&3\\3&5\end{bmatrix}$$

so that, from (9)

$$Q=[14\quad 54]\tfrac{1}{6}\begin{bmatrix}9&3\\3&5\end{bmatrix}\begin{bmatrix}14\\54\end{bmatrix}=3{,}480.$$

Now consider the nontestable hypothesis

$$H\colon \alpha_1=7$$

with $\mathbf{K}'=[0\ \ 1\ \ 0\ \ 0]$. Because it is nontestable we know that its $SSE_H=SSE$ and $F=0$. This is confirmed from (28) and (29) as follows. For (28)

$$\mathbf{K}'\mathbf{G}=[0\ \ \tfrac{1}{3}\ \ 0\ \ 0]\quad\text{and}$$

$$[\mathbf{K}'(\mathbf{I}-\mathbf{H})]^-=\left\{[0\quad 1\quad 0\quad 0]\begin{bmatrix}1&\cdot&\cdot&\cdot\\-1&\cdot&\cdot&\cdot\\-1&\cdot&\cdot&\cdot\\-1&\cdot&\cdot&\cdot\end{bmatrix}\right\}^-=\begin{bmatrix}-1\\0\\0\\0\end{bmatrix}.$$

Hence (28) is

$$\mathbf{G}_r=\begin{bmatrix}\cdot&\cdot&\cdot&\cdot\\ \cdot&\frac{1}{3}&\cdot&\cdot\\ \cdot&\cdot&\frac{1}{2}&\cdot\\ \cdot&\cdot&\cdot&1\end{bmatrix}-\begin{bmatrix}1&\cdot&\cdot&\cdot\\ -1&\cdot&\cdot&\cdot\\ -1&\cdot&\cdot&\cdot\\ -1&\cdot&\cdot&\cdot\end{bmatrix}\begin{bmatrix}-1\\ \cdot\\ \cdot\\ \cdot\end{bmatrix}[0\quad \tfrac{1}{3}\quad 0\quad 0]=\begin{bmatrix}\cdot&\frac{1}{3}&\cdot&\cdot\\ \cdot&\cdot&\cdot&\cdot\\ \cdot&-\frac{1}{3}&\frac{1}{2}&\cdot\\ \cdot&-\frac{1}{3}&\cdot&1\end{bmatrix}.$$

Then for (29)

$$\mathbf{w}=\begin{bmatrix}1&\cdot&\cdot&\cdot\\ -1&\cdot&\cdot&\cdot\\ -1&\cdot&\cdot&\cdot\\ -1&\cdot&\cdot&\cdot\end{bmatrix}\begin{bmatrix}-1\\ \cdot\\ \cdot\\ \cdot\end{bmatrix}7=\begin{bmatrix}-7\\ 7\\ 7\\ 7\end{bmatrix}\quad\text{and}\quad \mathbf{H}_r=\begin{bmatrix}1&1&\cdot&\cdot\\ \cdot&\cdot&\cdot&\cdot\\ \cdot&-1&1&\cdot\\ \cdot&-1&\cdot&1\end{bmatrix},$$

so that (29) is

$$\mathbf{b}^0_H=\begin{bmatrix}\cdot&\frac{1}{3}&\cdot&\cdot\\ \cdot&\cdot&\cdot&\cdot\\ \cdot&-\frac{1}{3}&\frac{1}{2}&\cdot\\ \cdot&-\frac{1}{3}&\cdot&1\end{bmatrix}\begin{bmatrix}504\\ 300\\ 172\\ 32\end{bmatrix}+\begin{bmatrix}\cdot&-1&\cdot&\cdot\\ \cdot&1&\cdot&\cdot\\ \cdot&1&\cdot&\cdot\\ \cdot&1&\cdot&\cdot\end{bmatrix}\begin{bmatrix}-7\\ 7\\ 7\\ 7\end{bmatrix}=\begin{bmatrix}93\\ 7\\ -7\\ -61\end{bmatrix}.$$

Hence

$$\mathbf{Xb}^0_H=[100\quad 100\quad 100\quad 86\quad 86\quad 32]'=\mathbf{Xb}^0$$

and so $SSE_H=SSE$ and thus $F=0$.

Similarly for (36)

$$(\mathbf{X'X}+\mathbf{KK'})^-=\left\{\mathbf{X'X}+\begin{bmatrix}0\\ 1\\ 0\\ 0\end{bmatrix}[0\quad 1\quad 0\quad 0]\right\}^-=\begin{bmatrix}6&3&2&1\\ 3&4&\cdot&\cdot\\ 2&\cdot&2&\cdot\\ 1&\cdot&\cdot&1\end{bmatrix}^-$$

$$=\frac{1}{6}\begin{bmatrix}8&-6&-8&-8\\ -6&6&6&6\\ -8&6&11&8\\ -8&6&8&14\end{bmatrix}$$

so that (36) is

$$\mathbf{b}^0_H=\frac{1}{6}\begin{bmatrix}8&-6&-8&-8\\ -6&6&6&6\\ -8&6&11&8\\ -8&6&8&14\end{bmatrix}\begin{bmatrix}504+0\\ 300+7\\ 172+0\\ 32+0\end{bmatrix}=\begin{bmatrix}93\\ 7\\ -7\\ -61\end{bmatrix}.$$

This is the same $\mathbf{b}^0_H$ as yielded by (29), but in general this need not be so. By the nature of this simple example, $\mathbf{K}'$ cannot have more than one row; but in cases where $\mathbf{K}'$ has less than its maximum number of rows, i.e., less than $p-r_{\mathbf{X}}$ rows, (29) and (36) do not necessarily yield the same $\mathbf{b}^0_H$; indeed $\mathbf{X'X}+\mathbf{KK'}$ may not be nonsingular, as it is here.

Although H: $\alpha_1=7$ is nontestable and hence has $F=0$, we find that with $\mathbf{K}'=[0\quad 1\quad 0\quad 0]$ and $\mathbf{G}$ of (43), the Q of (9) can be computed because $\mathbf{K'GK}=\frac{1}{3}$ is nonsingular.

Hence, with $\mathbf{K}'\mathbf{b}^0-\mathbf{m}=100-7=93$, $Q=93^2(3)=25{,}947$. This is confirmed from (38):

$$\text{(44)}\qquad \mathbf{K}'\mathbf{Hb}=[0\quad 1\quad 0\quad 0]\begin{bmatrix}\cdot & \cdot & \cdot & \cdot\\ 1 & 1 & \cdot & \cdot\\ 1 & \cdot & 1 & \cdot\\ 1 & \cdot & \cdot & 1\end{bmatrix}\begin{bmatrix}\mu\\ \alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix}=\mu+\alpha_1$$

so that

$$\mathbf{K}'\mathbf{Hb}^0-\mathbf{m}=0+100-7=93$$

and

$$\mathbf{K}'\mathbf{HG}(\mathbf{K}'\mathbf{H})'=[1\quad 1\quad 0\quad 0]\mathbf{G}[1\quad 1\quad 0\quad 0]'=\tfrac{1}{3}$$

and hence

$$Q=93^2(3)=25{,}947$$

as before.

In contrast to $\mathbf{K}'\mathbf{GK}$ being nonsingular for the choice of $\mathbf{G}$ above, if we choose

$$\mathbf{G}=\tfrac{1}{6}\begin{bmatrix}2 & \cdot & -2 & -2\\ \cdot & \cdot & \cdot & \cdot\\ -2 & \cdot & 5 & 2\\ -2 & \cdot & 2 & 8\end{bmatrix}$$

then $\mathbf{K}'\mathbf{GK}$ is zero. This illustrates that for a given $\mathbf{K}'$, the existence of Q depends on the choice of $\mathbf{G}$.

Finally, consider the partially testable hypothesis

$$H:\begin{cases}\alpha_1=7\\ \alpha_1-\alpha_2=3\end{cases}$$

for which

$$\mathbf{K}'=\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 1 & -1 & 0\end{bmatrix},\quad \mathbf{K}'\mathbf{b}^0-\mathbf{m}=\begin{bmatrix}100\\ 14\end{bmatrix}-\begin{bmatrix}7\\ 3\end{bmatrix}=\begin{bmatrix}93\\ 11\end{bmatrix}$$

and

$$(\mathbf{K}'\mathbf{GK})^{-1}=\begin{bmatrix}\frac{1}{3} & \frac{1}{3}\\ \frac{1}{3} & \frac{5}{6}\end{bmatrix}^{-1}=\begin{bmatrix}5 & -2\\ -2 & 2\end{bmatrix}.$$

Then

$$Q=[93\quad 11]\begin{bmatrix}5 & -2\\ -2 & 2\end{bmatrix}\begin{bmatrix}93\\ 11\end{bmatrix}=39{,}395.$$

From (42), with H_1: $\mathbf{K}_1'\mathbf{b}=\mathbf{m}_1$ being H_1: $\alpha_1-\alpha_2=3$ the testable part, and H_2: $\mathbf{K}_2'\mathbf{b}=\mathbf{m}_2$ being H_2: $\alpha_1=7$ the nontestable part, (42) gives, using (44)

$$H:\begin{cases}\alpha_1-\alpha_2=3\\ \mu+\alpha_1=7\end{cases}$$

with

$$\mathbf{K}'\mathbf{b}-\mathbf{m}=\begin{bmatrix}0&1&-1&0\\1&1&0&0\end{bmatrix}\begin{bmatrix}0\\100\\86\\32\end{bmatrix}-\begin{bmatrix}3\\7\end{bmatrix}=\begin{bmatrix}11\\93\end{bmatrix}$$

and

$$(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}=\begin{bmatrix}\frac{5}{6}&\frac{1}{3}\\\frac{1}{3}&\frac{1}{3}\end{bmatrix}^{-1}=\begin{bmatrix}2&-2\\-2&5\end{bmatrix}$$

so that

$$Q=[11\quad 93]\begin{bmatrix}2&-2\\-2&5\end{bmatrix}\begin{bmatrix}11\\93\end{bmatrix}=39{,}395$$

as before. Again, using the **G** in (45) gives a singular **K**′**GK** and leads to the nonexistence of Q.

Acknowledgment. Grateful thanks go to a referee for his perspicacious question which led to the definitions of nontestable and partially testable hypotheses.

REFERENCES

H. SCHEFFÉ, *The Analysis of Variance*, John Wiley, New York, 1959.
S. R. SEARLE, *Linear Models*, John Wiley, New York, 1971.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

005

THE APPROXIMATION OF ARBITRARY CLUSTERING FUNCTIONS BY CLUSTERING FUNCTIONS WHICH SATISFY OVERLAPPING CRITERIA*

GERHARD HERDEN†

Abstract. Let (H, h) be an arbitrary hierarchy and let f be its corresponding clustering function. In order to eliminate "chaining" complete characterizations of all minimal dominating hierarchies of (H, h) and all minimal dominating clustering functions of f which satisfy some overlapping criterion in the sense of Hubert are presented. These characterizations may be used—at least in principle—to construct all minimal dominating hierarchies of (H, h) and all minimal dominating clustering functions of f which satisfy a desired overlapping criterion. Finally the "classical" case that f is the diameter-function of some dissimilarity coefficient is considered. All overlapping criteria are characterized such that at least one of the minimal dominating clustering functions of f which satisfy one of these criteria is the diameter-function of some (weakly) k-ultrametric in the sense of Jardine and Sibson.

Introduction. In 1977 Hubert [5] studied extensively real valued (isotone) functions on a finite set S of data which measure the homogeneity of the subsets of S. These "clustering functions" correspond in a natural way bijectively to the "generalized hierarchies" on S (cf. Herden [3, Thm. 1.4]). Often one is interested in *clusters* ("maximal linked" sets, elements of the generalized hierarchies on S) which satisfy some *overlapping criterion.* In order to avoid "chaining", Jardine and Sibson [8] introduced for example the already very well-known and often discussed cluster-methods "B_k" which approximate a given dissimilarity coefficient by its uniquely determined maximal subdominating (weakly) k-ultrametric.

There are two natural "lattice-theoretical" possibilities to approximate an arbitrary clustering function f by some clustering function f^+ which satisfies an overlapping criterion A_k in the sense of Hubert [5] or more generally 0K in the sense of Herden [3].

1. Determine all maximal subdominating clustering functions $f^+ \leqq f$ of f which satisfy A_k or 0K!
2. Determine all minimal dominating clustering functions $f^+ \geqq f$ of f which satisfy A_k or 0K!

If f is the "diameter function" of some dissimilarity coefficient the first possibility was studied extensively by Jardine and Sibson [8] (cf. the cluster-methods B_k). Their results may easily be generalized to arbitrary clustering functions. Hence we study—for the sake of brevity—only the second possibility in this paper.

Throughout the literature a special cluster-method to construct a minimal dominating clustering function is well known in only one case, namely the "complete linkage method" (CLM). The CLM produces for a given diameter function f some minimal dominating clustering function of f which is the diameter-function of an ultrametric. But even the CLM does not allow constructing *all* possible minimal dominating diameter functions of f which are diameter functions of ultrametrics (cf. Bock [2, § 39]). No examples of other *dominant* methods are well known. In [8, p. 75], Jardine and Sibson wrote:

> Another way in which other types of non-hierarchic methods could arise is by the consideration of methods which are not subdominant methods. No examples of methods of this type are known, and it seems likely that they would be difficult to work with.

* Received by the editors December 14, 1981, and in revised form August 10, 1983.

† Fachbereich für Mathematik und Didaktik der Mathematik, Universität/Gesamthochschule, Essen, West Germany.

In contrast to the terminology of this paper Jardine and Sibson called a method "nonhierarchic" if the clusters may satisfy some overlapping criterion.

Theorem 2.4 characterizes all hierarchies and all clustering functions which satisfy some overlapping criterion and which are minimal dominating hierarchies or clustering functions on S. These characterizations allow us—at least in principle—to construct all these minimal dominating hierarchies or clustering functions.

We should mention the most important points in which subdominant and dominant methods differ:

1. The clustering functions which may be constructed by subdominant methods are uniquely determined (cf. Jardine and Sibson [8]) while many different clustering functions may be constructed by dominant methods. This fact is at least implicitly well known (cf. Bock [2, § 38]).

2. The subdominant methods which were introduced by Jardine and Sibson do not really cut down on chaining. This fact has been pointed out for example by Matula [9] and Rohlf [10]. To be more precise we consider an arbitrary clustering function f. The clusters of some maximal subdominating clustering function f^- of f are always unions of clusters of f. Hence chaining cannot be eliminated by any subdominant methods.

In contrast, the clusters which are obtained by some minimal dominating clustering function f^+ of f are always contained in the original clusters of f. Hence chaining of clusters is surely eliminated if one uses dominant methods as characterized in this paper instead of subdominant methods.

In the last section we study the classical case where f is the diameter function of some dissimilarity coefficient on S. In general *no* minimal dominating clustering function of f which satisfies some overlapping criterion is the diameter function of some (weakly) k-ultrametric (cf. the example of § 2).

Hence it is an interesting problem to determine all overlapping criteria such that at least one minimal dominating clustering function of f which satisfies one of these criteria is the diameter function of some (weakly) k-ultrametric. This problem will be solved completely in Theorem 3.1 of this paper.

The reader is assumed to be familiar with the basic concepts of hierarchical clustering due to Jardine and Sibson [7], [8], Hubert [4], [5] and Janowitz [6]. Many papers on the more classical concepts of hierarchical clustering may be found in the bibliography of Hubert [4].

We develop our theory within the mainstream of the theory of partially ordered sets. The reader should consult the account of Blyth and Janowitz [1] or the paper of Herden [3] whenever necessary.

1. The background.

1.1. The basic situation. Let $(L, \leqq)$ be a finite lattice and let $(M, \leqq)$ be a linearly ordered complete meet semilattice. In all *classical* concepts of hierarchical clustering $(L, \leqq)$ is the power set of a finite set S of data partially ordered by set inclusion and $(M, \leqq)$ is the set of nonnegative reals partially ordered in the usual manner.

Of particular interest within our concept of hierarchical clustering is the lattice $(\bar{L}, \subset)$ of order ideals of L. For every nonempty subset $K \subsetneq L$ we denote by I_K the order ideal of L which is generated by K.

Following the notation of Janowitz [6], $\mathrm{Res}^+(M, \bar{L})$ denotes the set of residual mappings $g: M \to \bar{L}$ and $\mathrm{Res}(\bar{L}, M)$ denotes the set of residuated mappings $\bar{g}: \bar{L} \to M$.

One may conclude from Herden [3] that every isotone mapping $f: L \to M$ with $f(0) = 0$ is a clustering function and that a pair $(H, h)(H \subset L, h: H \to M)$ is a hierarchy iff it satisfies the following conditions:

H0: $\{a \in H | h(a) = 0\} \neq \emptyset$,

H^+1: $1 \in H$,

H3: $a < b \Rightarrow h(a) < h(b)$ for all $a, b \in H$.

Let Clus (L, M) be the set of clustering functions $f: L \to M$ and let Hier (L) be the set of hierarchies on L. Because of Herden [3, Thm. 1.4] there exist natural bijections between any pairs of the following sets: Clus (L, M), Res $(\bar{L}, M)$, Res^+ $(M, \bar{L})$ and Hier (L).

1.2. Overlapping criteria. Let K be a nonempty subset of L. In order to study a generalized version of Hubert's monotone K-clustering functions (cf. [5]) within our model of hierarchical clustering we consider the order filter F_K of L which is generated by K. It is easy to see that a clustering function $f: L \to M$ is a K-clustering function in the sense of Herden [3] iff for all $a, b \in L$ the following condition holds:

0K: if $\inf\{a, b\} \in F_K$ then $f(\sup\{a, b\}) \leqq \max\{f(a), f(b)\}$.

Furthermore a hierarchy (H, h) is a K-hierarchy in the sense of Herden [3] iff it satisfies the following additional condition:

HK: if $\inf\{a, b\} \in F_K$ then $a \leqq b$ or $b \leqq a$ for all $a, b \in H$.

We now consider the set $\bar{L}_K := \{I \in \bar{L} | \sup\{a, b\} \in I$ for all $a, b \in I$ such that $\inf\{a, b\} \in F_K\}$. $(\bar{L}_K, \subset)$ is a sublattice of $(\bar{L}, \subset)$ and for every $I \in \bar{L}$ there exists a uniquely determined smallest order ideal $I^K \in \bar{L}_K$ such that $I \subset I^K$. Moreover the mapping $I \to I^K$ is a closure operator on $\bar{L}$.

Let Clus_K (L, M) be the set of K-clustering functions $f: L \to M$ and let Hier_K (L) be the set of K-hierarchies on L. In addition to Herden [3, Thm. 2.3] we now present a short proof of the following proposition.

PROPOSITION 1.1. *There are natural bijections between any pairs of the following sets*: Clus_K (L, M), Res $(\bar{L}_K, M)$, Res^+ $(M, \bar{L}_K)$ *and* Hier_K (L).

Proof. Because of Herden [3, Thm. 2.3] it is sufficient to establish a natural bijection between Clus_K (L, M) and Res^+ $(M, \bar{L}_K)$. The reader may easily verify that this bijection is given by the following maps:

$g: \text{Clus}_K$ $(L, M) \to \text{Res}^+$ $(M, \bar{L}_K)$ defined by

$g_f(m) := \{a \in L | f(a) \leqq m\}$ for all $f \in \text{Clus}_K$ (L, M) and all $m \in M$;

$f: \text{Res}^+$ $(M, \bar{L}_K) \to \text{Clus}_K$ (L, M) defined by

$f_g(a) := \inf\{m \in M | a \in g(m)\}$ for all $g \in \text{Res}^+$ $(M, \bar{L}_K)$ and all $a \in L$.

1.3. *K*-cluster-methods. We define a K-cluster-method to be a function T: Clus $(L, M) \to \text{Clus}_K$ (L, M). Because of Proposition 1.1 a K-cluster-method may also be thought of as a function $F: \text{Res}^+$ $(M, \bar{L}) \to \text{Res}^+$ $(M, \bar{L}_K)$ or $\bar{F}$: Res $(\bar{L}, M) \to$ Res $(\bar{L}_K, M)$ or as a function G: Hier $(L) \to \text{Hier}_K$ (L).

As we already mentioned in the introduction two types of K-cluster-methods are of particular interest:

The first associate every $f \in$ Clus (L, M) with its uniquely determined maximal subdominating K-clustering function or respectively every $F \in \text{Res}^+$ $(M, \bar{L})$ with its uniquely determined minimal dominating residual mapping $F^+: M \to \bar{L}_K$.

If f is the diameter-function of some dissimilarity coefficient these methods were studied extensively by Jardine and Sibson [8] (cf. their methods B_k). Their results may easily be extended to the general case.

Hence we study in this paper only the second type of K-cluster-methods which associate every $f \in$ Clus (L, M) with some, not necessarily uniquely determined, minimal dominating K-clustering function or respectively every $F \in \mathrm{Res}^+ (M, \bar{L})$ with some maximal subdominating residual mapping $F^- : M \to \bar{L}_K$.

2. The approximation theorem. Let (H_1, h_1), (H_2, h_2) be arbitrary hierarchies with corresponding clustering functions $f_1, f_2 : L \to M$ and corresponding residual mappings $g_1, g_2 : M \to \bar{L}$. We define $(H_1, h_1) \leqq (H_2, h_2) \Leftrightarrow f_1 \leqq f_2 \Leftrightarrow g_2 \leqq g_1$.

In order to prove the main result of this section we need three lemmas.

LEMMA 2.1. *The following conditions are equivalent*:

(i) $(H_1, h_1) \leqq (H_2, h_2)$.

(ii) *For all* $b \in H_2$ *there exists some* $a \in H_1$ *such that* $b \leqq a$ *and* $h_1(a) \leqq h_2(b)$.

Proof. Recall from Herden [3, Thm. 1.4] that $(H_1, h_1) \leqq (H_2, h_2)$ iff $f_1(c) = \min \{h_1(a) | a \in H_1, c \leqq a\} \leqq f_2(c) = \min \{h_2(b) | b \in H_2, c \leqq b\}$ for all $c \in L$.

LEMMA 2.2. *Let* (H, h) *be a hierarchy with corresponding clustering function* $f : L \to M$; *then* $(H, h) \leqq (H', h')$ *for every hierarchy* (H', h') *such that* $h' = f|_{H'}$.

Proof. Let $f' : L \to M$ be the (H', h') corresponding clustering function. Because of $f(a) \leqq \min \{f(b) | b \in H', a \leqq b\} = f'(a)$ for all $a \in L$ the desired inequality follows.

LEMMA 2.3 (key lemma). *Let* $(H_1, h_1) \leqq (H_2, h_2)$ *be hierarchies and let* $f_1 : L \to M$ *be the* (H_1, h_1) *corresponding clustering function. If* $(H_2, h_2) \in \mathrm{Hier}_K (L)$ *then there exists a hierarchy* (H_3, h_3) *which satisfies the following conditions*:

(i) $(H_3, h_3) \in \mathrm{Hier}_K (L)$,

(ii) $(H_1, h_1) \leqq (H_3, h_3) \leqq (H_2, h_2)$,

(iii) $h_3 = f_{1|H_3}$.

Proof. Let a be an arbitrary element of H_2. We set $O(a) := \{b < a | f_1(a) \leqq f_1(b)\}$. Then we define (H_3, h_3) by $H_3 := H_2 \setminus \bigcup_{a \in H_2} O(a)$ and $h_3 := f_{1|H_3}$.

We now show in a first step:

(+) For every $b \in H_2$ there exists some $a \in H_3$ such that $b \leqq a$ and $f_1(a) \leqq f_1(b) \leqq h_2(b)$.

Let b be an arbitrary element of H_2. We set $b_1 := b$. Let $b_t \geqq b$ be defined for $t \in \mathbb{N}$. We consider the following two cases:

Case 1. $b_t \in H_3$. Then we set $b_{t+1} := b_t$.

Case 2. $b_t \notin H_3$. Then there exists some $b_{t+1} \in H_2$ such that $b_t \in O(b_{t+1})$.

A routine induction argument shows that $b_t = b_{t+1}$ if $b_t \in H_3$ and that $b_t < b_{t+1}$ if $b_t \notin H_3$ for all $t \in \mathbb{N}$. Moreover we may conclude that $f_1(b_t) \leqq f_1(b) \leqq h_2(b)$ for all $t \in \mathbb{N}$.

L is a finite lattice. Hence there exists some $t \in \mathbb{N}$ such that $b_t \in H_3$ and (+) follows.

In a second step we show that $(H_3, h_3) \in \mathrm{Hier}_K (L)$.

H0: This is an immediate consequence of (+).

H^+1: Trivial.

H3: Let $a < b$ for $a, b \in H_3$. If $f_1(b) \leqq f_1(a)$ then $a \in O(b)$. Hence $a \notin H_3$ which contradicts our assumption.

HK: This follows from the inclusion $H_3 \subset H_2$.

It remains to verify condition (ii). The inequality $(H_1, h_1) \leqq (H_3, h_3)$ follows with the help of Lemma 2.2 from the definition of h_3. On the other hand the inequality $(H_3, h_3) \leqq (H_2, h_2)$ is an immediate consequence of Lemma 2.1 and (+).

Before we are able to formulate and prove the "approximation theorem" we need some more definitions and notation:

DEFINITION 1. Let $(V, \leqq)$ be an arbitrary partially ordered set and let $(U, \leqq)$, $(W, \leqq)$ be subsets of V. An element $w \in W$ is a *maximal subdominating* element of U if $w \leqq u$ for all $u \in U$ and if $w = w'$ for all $w \in W$ such that $w \leqq w' \leqq u$ for all $u \in U$. If $w \in W$ satisfies the "dual" conditions it is called a *minimal dominating* element of U.

DEFINITION 2. Let $f: L \to M$ be an arbitrary clustering function. $\mathcal{M}_K$ denotes the set of all subsets $N \subset L$ such that the following conditions hold:

MK1: For all $a, b \in N$ such that $\inf\{a, b\} \in F_K$ there exists some $c \in N$ such that $\sup\{a, b\} \leqq c$ and $f(c) \leqq \max\{f(a), f(b)\}$.

MK2: $N = N'$ for all subsets $N' \subset L$ such that $f|_{N'}$ satisfies condition MK1 and $N \subset N'$.

DEFINITION 3. Let $g: M \to \bar{L}$ be an arbitrary residual mapping. An *isotone* family $\{I_m\}_{m \in M}$ of order ideals $I_m \in \bar{L}_K$ is called *maximal subdominating* with respect to g iff it satisfies the following conditions:

LK1: $I_m \subset g(m)$ for all $m \in M$.

LK2: If $\{J_m\}_{m \in M}$ is another *isotone* family of order ideals $J_m \in \bar{L}_K$ which satisfies condition LK1 and if $I_m \subset J_m$ for all $m \in M$ then $I_m = J_m$ for all $m \in M$.

We now consider the hierarchies (H_1, h_1) and (H_2, h_2) with corresponding clustering functions $f_1, f_2: L \to M$ and corresponding residual mappings $g_1, g_2: M \to \bar{L}$ and prove the following theorem.

THEOREM 2.4 (approximation theorem). *The following conditions are equivalent*:

(i) (H_2, h_2) *is a minimal dominating K-hierarchy of* (H_1, h_1).

(ii) f_2 *is a minimal dominating K-clustering function of* f_1.

(iii) $g_2 \in \mathrm{Res}^+(M, \bar{L}_K)$ *is a maximal subdominating residual mapping of* g_1.

(iv) (H_2, h_2) *satisfies condition* HK *and the following conditions*:

(a) $h_2 = f_{1|H_2}$;

(b) *an element a of L is in* H_2 *provided it satisfies the following pair of conditions*:

(b1) $f_1(a) < h_2(b)$ *for all* $a < b \in H_2$,

(b2) *if* $\inf\{a, b\} \in F_K$ *then* $a \leqq b$ *or* $b \leqq a$ *for all* $b \in H_2$.

(v) *There exists some* $N \in \mathcal{M}_K$ *such that* $f_2(a) = \min\{f_1(b) | b \in N, a \leqq b\}$ *for all* $a \in L$.

(vi) *There exists some maximal subdominating family* $\{I_m\}_{m \in M}$ *of order ideals* $I_m \in \bar{L}_K$ *with respect to* g_1 *such that* $g_2(m) = I_m$ *for all* $m \in M$.

Proof. It is sufficient to prove the equivalence of (i) and (iv), (ii) and (v) and (iii) and (vi).

(i)$\Rightarrow$(iv). We have to verify the conditions HK, (a) and (b):

HK: definition.

(a): Use Lemma 2.3.

(b): Let $a \in L$ satisfy the conditions (b1) and (b2). We define a K-hierarchy (H_3, h_3) by $H_3 := (H_2 \cup \{a\}) \backslash \{b < a | f_1(a) \leqq f_1(b)\}$ and $h_3 := f_{1|H_3}$. (H_3, h_3) is indeed a K-hierarchy as the reader will immediately verify. With the help of Lemmas 2.2 and 2.1 it is easy to see that $(H_1, h_1) \leqq (H_3, h_3) \leqq (H_2, h_2)$. On the other hand (H_2, h_2) is a minimal dominating K-hierarchy of (H_1, h_1). Hence we may conclude that $(H_2, h_2) = (H_3, h_3)$ and that $a \in H$.

(iv)$\Rightarrow$(i). With the help of Lemma 2.2 condition (a) implies that $(H_1, h_1) \leqq (H_2, h_2)$.

We now assume the existence of some K-hierarchy (H_3, h_3) such that $(H_1, h_1) \leqq (H_3, h_3) < (H_2, h_2)$. Let $f_3: L \to M$ be the (H_3, h_3) corresponding clustering function. We consider the set $N := \{c \in L | f_3(c) < f_2(c)\}$. Because of our assumption N is not empty. On the other hand L is finite. Hence there exists some $a \in N \subset L$ such that $f_3(c) \leqq f_3(a)$ for all $c \in N$. We may assume without loss of generality that $a \in H_3$ (cf. the proof of Herden [3, Thm. 1.4]). Because of $f_{2|H_2} = h_2 = f_{1|H_2}$ the inequalities $f_1(a) \leqq f_3(a) < f_2(a)$ imply that $a \notin H_2$. But $a \in H_3$. Hence the following inequalities hold for all $a < b \in L$: $f_1(a) \leqq f_3(a) < f_3(b) \leqq f_2(b)$. Hence a satisfies condition (b1).

But $a \notin H_2$. Hence there exists some $b \in H_2$ such that $\inf\{a, b\} \in F_K$ but $a \nleqq b$ and $b \nleqq a$.

We now consider an element $a_1 \in H_2$ such that $a \leqq a_1$ and $f_2(a) = f_2(a_1)$ (cf. the proof of Herden [3, Thm. 1.4] and we may conclude that $\inf\{a_1, b\} \in F_K$. Hence $a_1 \leqq b$ or $b \leqq a_1$. If $a_1 \leqq b$ then $a \leqq a_1 \leqq b$ which is a contradiction. Thus we have proved the inequalities $b < a_1$ and $(+) f_2(b) < f_2(a_1)$. In the next step we consider some $b_2 \in H_3$ such that $b \leqq b_2$ and $f_3(b) = f_3(b_2)$.

Because $\inf\{a, b_2\} \in F_K$ we may conclude that $a \leqq b_2$ or $b_2 \leqq a$. But $b \nleqq a$. Hence $a < b_2$. This implies the inequalities $f_2(a) \leqq f_2(b_2)$ and $f_3(a) < f_3(b_2)$. On the other hand we just proved the following inequalities: $f_3(b_2) = f_3(b) \leqq f_2(b) < f_2(a_1) = f_2(a) \leqq f_2(b_2)$ (cf. (+)). Hence $b_2 \in N$ and $f_3(a) < f_3(b_2)$ which contradicts our assumption on a.

(ii)$\Rightarrow$(v). The characterization of (H_2, h_2) implies together with Herden [3, Thm. 1.4] that $f_2(a) = \min\{f_1(b) \mid b \in H_2, a \leqq b\}$ for all $a \in L$. Because of condition HK $f_{1|_{H_2}}$ satisfies condition MK1. Hence there exists some set $N \in \mathcal{M}_K$ such that $H_2 \subset N$.

We now define an isotone mapping $f_3 : L \to M$ by

$$f_3(a) := \min\{f_1(b) \mid b \in N, a \leqq b\} \quad \text{for all } a \in L.$$

This definition implies immediately the following inequalities: $f_1 \leqq f_3 \leqq f_2$. With the help of condition MK1 one may easily verify that f_3 is a K-clustering function. But f_2 is a minimal dominating K-clustering function of f_1. Hence $f_2 = f_3$.

(v)$\Rightarrow$(ii). Let N be an arbitrary element of $\mathcal{M}_K$. Because of MK1 the clustering function $f_3 : L \to M$ defined by $f_3(a) := \min\{f_1(b) \mid b \in N, a \leqq b\}$ for all $a \in L$ is a K-clustering function.

We now define a K-hierarchy (H_4, h_4) by $H_4 := \{b \in N \mid f_1(b) < f_1(c)$ for all $b < c \in N\}$ and $h_4 := f_{1|_{H_4}}$.

The reader will immediately verify that (H_4, h_4) is indeed a K-hierarchy. Let $f_4 : L \to M$ be the (H_4, h_4) corresponding K-clustering function then $f_1 \leqq f_3 \leqq f_4$. Hence it remains to prove that (H_4, h_4) is a minimal dominating K-hierarchy of (H_1, h_1). As we already know that (H_4, h_4) is a K-hierarchy and because of the definition of h_4 we only have to verify condition (b) of (iv). Therefore we consider an arbitrary element a of L which satisfies the conditions (b1) and (b2). In a first step we show that $f_{1|_{N \cup \{a\}}}$ satisfies condition MK1. Hence we consider an arbitrary element b of N such that $\inf\{a, b\} \in F_K$. The definition of H_4 implies the existence of some $b_1 \in H_4$ such that $b \leqq b_1$ and $f_1(b) = f_1(b_1)$. Because of $\inf\{a, b\} \in F_K$ condition (b2) implies that $a \leqq b_1$ or $b_1 \leqq a$.

If $a \leqq b_1$ then $\sup\{a, b\} \leqq b_1$ and $f_1(b_1) = f_1(b) \leqq \max\{f_1(a), f_1(b)\}$. On the other hand the inequality $b_1 \leqq a$ implies that $b \leqq a$. Hence $\sup\{a, b\} \leqq a$ and $f_1(a) \leqq \max\{f_1(a), f_1(b)\}$.

Condition MK2 now implies that $N \cup \{a\} = N$. Hence we may conclude that $a \in N$. With the help of condition (b1) the definition of H_4 implies that $a \in H_4$. Hence (H_4, h_4) is indeed a minimal dominating K-hierarchy of (H_1, h_1).

(iii)$\Leftrightarrow$(vi). The reader may immediately verify that the equivalence of the conditions (iii) and (vi) follows if we are able to prove that for every maximal subdominating family $\{I_m\}_{m \in M}$ of order ideals $I_m \in \bar{L}_K$ with respect to g_1 the isotone mapping $g_3 : M \to \bar{L}_K$ defined by $g_3(m) := I_m$ for all $m \in M$ is actually residual. Because of Blyth and Janowitz [1] or Herden [3] we thus have to show that $L \in \operatorname{Im}(g_3)$ and that arbitrary meets are preserved by g_3: The relations $L \in \operatorname{Im}(g_1)$ and $L \in \bar{L}_K$ imply with the help of condition LK2 that $L \in \operatorname{Im}(g_3)$. It thus remains to prove that arbitrary meets are preserved by g_3: Let $\{m_j\}_{j \in J}$ be an arbitrary family of elements of M. We set $n := \inf_{j \in J} m_j$. Clearly $I_n = g_3(n) \subset \bigcap_{j \in J} g_3(m_j)$. If we assume that $g_3(n) \not\subseteq \bigcap_{j \in J} g_3(m_j)$ then we may define an isotone family $\{J_m\}_{m\in M}$ of order ideals $J_m \in \bar{L}_K$ by

$$J_m := I_m \quad \text{for all } m \in M\backslash\{n\} \quad \text{and} \quad J_n := \bigcap_{j\in J} g_3(m_j).$$

It is easy to see that $\{J_m\}_{m\in M}$ satisfies condition LK1. Hence condition LK2 implies immediately the equality of $g_3(n)$ and $\bigcap_{j\in J} g_3(m_j)$ and nothing remains to prove.

COROLLARY 2.5.

(i) *For every K-hierarchy* $(H_3, h_3) \geqq (H_1, h_1)$ *there exists some minimal dominating K-hierarchy* (H_2, h_2) *of* (H_1, h_1) *such that* $(H_3, h_3) \geqq (H_2, h_2) \geqq (H_1, h_1)$.

(ii) *For every K-clustering function* $f_3 \geqq f_1$ *there exists some minimal dominating K-clustering function* f_2 *of* f_1 *such that* $f_3 \geqq f_2 \geqq f_1$.

(iii) *For every* $g_1 \geqq g_3 \in \mathrm{Res}^+(M, \bar{L}_K)$ *there exists some maximal subdominating* $g_2 \in \mathrm{Res}^+(M, \bar{L}_K)$ *of* g_1 *such that* $g_3 \leqq g_2 \leqq g_1$.

Proof. It is sufficient to prove assertion (ii). In order to prove this assertion one may use Lemma 2.3 and the implication (ii)⇒(iv) of Theorem 2.4.

COROLLARY 2.6.

(i) $\inf\{(H', h') \mid (H', h')$ *is a minimal dominating K-hierarchy of* $(H_1, h_1)\} = (H_1, h_1)$.

(ii) $\inf\{f' \mid f'$ *is a minimal dominating K-clustering function of* $f_1\} = f_1$.

(iii) $\sup\{g' \mid g' : M \to \bar{L}_K$ *is a maximal subdominating residual mapping of* $g_1\} = g_1$.

Proof. We prove equality (ii). For every $a \in L$ the mapping $f_{1|\{a,1\}}$ satisfies condition MK1. Hence there exists some $N \in \mathcal{M}_K$ which contains a and we may conclude that $f_1(a) = \min\{f_1(b) \mid b \in N, a \leqq b\}$.

Remark and example.

1. A K-cluster-method which associates every clustering function $f: L \to M$ with some minimal dominating K-clustering function f' of f or every $g \in \mathrm{Res}^+(M, \bar{L})$ with some maximal subdominating $g^- \in \mathrm{Res}^+(M, \bar{L}_K)$ of g clearly selects only *one* minimal dominating K-clustering function of f or *one* maximal subdominating $g^- \in \mathrm{Res}^+(M, \bar{L}_K)$ of g.

2. We illustrate Theorem 2.4 by the following example: Let M be the set of nonnegative reals. We consider the set $S := \{0_1, 0_2, 0_3, 0_4\}$ and define a *dissimilarity coefficient* $d: S \times S \to M$ by

$$d(0_i, 0_j) := |i - j| \quad \text{for all } 1 \leqq i, j \leqq 4.$$

The d corresponding clustering function $\mathrm{diam}_d : P(S) \to M$ is defined by

$$\mathrm{diam}_d(A) := \begin{cases} 0 & \text{if } A = \varnothing, \\ \max\{d(0_i, 0_j) \mid 0_i, 0_j \in A\} & \text{else} \end{cases}$$

(cf. Jardine and Sibson (8)).

The diam_d corresponding hierarchy (H_d, h_d) may be described in the following way:

$h_d = 3$: $\{0_1, 0_2, 0_3, 0_4\}$,

$h_d = 2$: $\{0_1, 0_2, 0_3\}, \{0_2, 0_3, 0_4\}$,

$h_d = 1$: $\{0_1, 0_2\}, \{0_2, 0_3\}, \{0_3, 0_4\}$,

$h_d = 0$: $\{0_1\}, \{0_2\}, \{0_3\}, \{0_4\}$.

Hence the diam_d corresponding residual mapping $g_d: M \to \overline{P(S)}$ is defined by

$$(+) \qquad g_d(m) := \begin{cases} P(S) & \text{if } 3 \leqq m, \\ I_{\{\{0_1,0_2,0_3\},\{0_2,0_3,0_4\}\}} & \text{if } 2 \leqq m < 3, \\ I_{\{\{0_1,0_2\},\{0_2,0_3\}\{0_3,0_4\}\}} & \text{if } 1 \leqq m < 2, \\ I_{\{\{0_1\},\{0_2\},\{0_3\},\{0_4\}\}} & \text{if } 0 \leqq m < 1. \end{cases}$$

Let K be the set of all subsets $A \subset S$ such that $|A| = 2$. In order to determine all minimal dominating K-clustering functions, all minimal dominating K-hierarchies and all maximal subdominating residual mappings $g': M \to \bar{L}_K$ of diam_d, (H_d, h_d) and g_d respectively, we determine the set $\mathcal{M}_K$ of all subsets $N \subset P(S)$ which satisfy the conditions MK1 and MK2 with respect to diam_d. $\mathcal{M}_K$ consists of

$$N_1 = P(S) \backslash \{\{0_1, 0_2, 0_3\}\} \quad \text{and} \quad N_2 = P(S) \backslash \{\{0_2, 0_3, 0_4\}\}.$$

Hence the minimal dominating K-clustering functions $f_1, f_3: P(S) \to M$ of diam_d are defined by

$$f_{1|N_1} := \mathrm{diam}_{d|N_1} \quad \text{and} \quad f_1(\{0_1, 0_2, 0_3\}) = 3,$$

$$f_{2|N_2} := \mathrm{diam}_{d|N_2} \quad \text{and} \quad f_2(\{0_2, 0_3, 0_4\}) = 3.$$

The minimal dominating K-hierarchies (H_1, h_1) and (H_2, h_2) of (H_d, h_d) may be described by

$$h_1 = 3\colon \{0_1, 0_2, 0_3, 0_4\},$$

$$h_1 = 2\colon \{0_2, 0_3, 0_4\}, \{0_1, 0_3\},$$

$$h_1 = 1\colon \{0_1, 0_2\}, \{0_2, 0_3\},\{0_3, 0_4\},$$

$$h_1 = 0\colon \{0_1\}, \{0_2\}, \{0_3\}, \{0_4\},$$

and

$$h_2 = 3\colon \{0_1, 0_2, 0_3, 0_4\},$$

$$h_2 = 2\colon \{0_1, 0_2, 0_3\}, \{0_2, 0_4\},$$

$$h_2 = 1\colon \{0_1, 0_2\}, \{0_2, 0_3\}, \{0_3, 0_4\},$$

$$h_2 = 0\colon \{0_1\}, \{0_2\}, \{0_3\}, \{0_4\}.$$

Finally the maximal subdominating residual mappings $g_1, g_2: M \to \overline{P(S)}$ of g_d are defined analogously (cf. (+)).

3. Characterization of order filters. Henceforth we assume that $(L, \leqq)$ is a finite locally atomic boolean algebra, i.e. $(L, \leqq)$ is isomorphic to the power set of a finite set S of data. Let $AL \subset L$ consist of all atoms and the least element of L. We set $ALV := \{a_{ij} \in L |$ there exist elements $a_i, a_j \in AL$ such that $a_{ij} = \sup\{a_i, a_j\}\}$. Since AL and ALV are subsets of L they are canonically partially ordered. We repeat from Herden [3] that a clustering function $d: ALV \to M$ is called a *dissimilarity coefficient* iff $d(a) = 0$ for all $a \in AL$.

Let $DC(L)$ be the set of dissimilarity coefficients on L. There exists a canonical *order monomorphism* $\mathrm{diam}: DC(L) \to \mathrm{Clus}\,(L, M)$ defined by $\mathrm{diam}_d := \max\{d(b) | b \in ALV, b \leqq a\}$ for all $a \in L$. If $DC_K(L)$ is especially the set of (weakly) K-ultrametrics on L (cf. Herden [3]) the image of $\mathrm{diam}: DC_K(L) \to \mathrm{Clus}\,(L, M)$ is denoted by $\mathrm{Clus}_K^d\,(L, M)$.

We now consider an arbitrary dissimilarity coefficient $d: ALV \to M$. The example of § 2 demonstrates, as is easily verified, that it may happen that none of the minimal dominating K-clustering functions of diam_d is the diameter function of some (weakly) K-ultrametric d'. Hence we may formulate the following problem (cf. the introduction):

Determine all order filters F_K such that for every $d \in DC(L)$ there exist minimal dominating K-clustering functions $f \in \mathrm{Clus}_K^d(L,M)$ of diam_d!

In order to approach this problem we set:

$$K^+ := F_K \cap AL, \; V := AL \backslash F_K, \quad v := \sup(V), \quad E_V := \{a \in L \mid a \leqq v\}.$$

Furthermore we consider the following conditions on F_K:

F1: $|K^+| \geqq 1$ and $a = v$ or $a \in F_{K^+}$ for all $a \in F_K$.

F2: $|K^+| \geqq 1$, $a \in F_K$ for some $a < v$ and $a = v$ or a is a coatom of $(E_V, \leqq)$ for all $a \in F_K \cap E_V$.

F3: $|K^+| \geqq 1$ and there exists some $a \in F_K \cap E_V$ such that $a < v$ and a is not a coatom of $(E_V, \leqq)$.

Some remarks.

1. In order to avoid trivial cases we assume that $|AL| \geqq 4$ and that M contains at least two elements $0 < m$.

2. If $F_K = L$ then the only one (weakly) K-ultrametric $d: ALV \to M$ *is defined by* $d(a) := 0$ for all $a \in ALV$. Hence in this case no dissimilarity coefficient which is not (weakly) K-ultrametric can be approximated by a minimal dominating (weakly) K-ultrametric and we assume therefore for the remainder of this paragraph that $F_K \not\subseteq L$, i.e. $0 \notin K$.

3. If $|K^+| \geqq 1$ then it is easy to see that there exist always dissimilarity coefficients $d: ALV \to M$ which are not (weakly) K-ultrametric such that all minimal dominating K-clustering functions f of diam_d are diameter functions of (weakly) K-ultrametrics.

Indeed, let a_t be an arbitrary element of K^+. We may define a dissimilarity coefficient $d: ALV \to M$ by

$$d(a_{ij}) := \begin{cases} 0 & \text{if } i = j \text{ or } j = t \\ m & \text{else} \end{cases} \qquad \text{for all } a_{ij} \in ALV.$$

The reader will immediately verify that $d \notin DC_K(L)$ and that every minimal dominating K-clustering function f of diam_d is in $\mathrm{Clus}_K^d(L, M)$.

The above remarks present together with the following theorem a complete solution of our problem.

THEOREM 3.1 (characterization theorem). *Let $d: ALV \to M$ be an arbitrary dissimilarity coefficient which is not (weakly) K-ultrametric then the following properties are satisfied*:

(i) *If $|V| \leqq 3$ or if F_K satisfies condition* F1 *then $f \in \mathrm{Clus}_K^d(L, M)$ for all minimal dominating K-clustering functions f of* diam_d.

(ii) *If F_K satisfies condition* F2 *then*

(a) *there exist minimal dominating K-clustering functions f of* diam_d *which are in* $\mathrm{Clus}_K^d(L, M)$,

(b) *there exist dissimilarity coefficients $d': ALV \to M$ such that $f \notin \mathrm{Clus}_K^d(L, M)$ for at least one minimal dominating K-clustering function f of* diam_d.

(iii) *If F_K satisfies condition* F3 *then there exist dissimilarity coefficients $d': ALV \to M$ such that no minimal dominating K-clustering function f of* diam_d *is in* $\mathrm{Clus}_K^d(L, M)$.

(iv) *If* $|K^+| = 0$ *then no minimal dominating* K*-clustering function* f *of* diam_d *is in* $\operatorname{Clus}^d_K(L, M)$.

Proof. (i) If we assume that $|V| \leqq 3$ then Herden [3, Thm. 4.1] implies that $f \in \operatorname{Clus}^d_K(L, M)$ for all minimal dominating K-clustering functions f of diam_d.

Hence we now assume that F_K satisfies condition F1.

Let N be an arbitrary element of $\mathscr{M}_K$ and let f be its corresponding minimal dominating K-clustering function of diam_d. Condition MK2 implies together with our assumption on $(L, \leqq)$ and condition F1 that $E_V \subset N$. Hence $f(a) \leqq \max\{d(b) \mid b \in ALV, b \leqq a\}$ for all $a \in E_V$.

We now consider an arbitrary element a of L such that $a_p \leqq a$ for some $a_p \in K^+$. Let $a_t \leqq a$ be an arbitrary atom then we consider the element $a' := \sup\{a_i \mid a_i \in AL, a_i \leqq a, a_i \neq a_t\}$. f is a K-clustering function. Hence the following inequality holds: $f(a) \leqq \max\{f(a_{pt}), f(\sup\{a_p, a'\})\}$. Moreover L is a finite set. Hence a routine induction argument implies that $f(a) \leqq f(a_{ps})$ for some $a_s \leqq a$. Especially we may thus conclude that $f(a) \leqq \max\{f(b) \mid b \in ALV, b \leqq a\}$.

(ii)(a) Condition F2 implies that E_V satisfies condition MK1. Hence there exists some $N \in \mathscr{M}_K$ such that $E_V \subset N$. As in the proof of property (i) we may thus conclude that the N corresponding minimal dominating K-clustering function f of diam_d is an element of $\operatorname{Clus}^d_K(L, M)$.

(b) Condition F2 implies that $|V| \geqq 4$. Furthermore there exists some $a \in F_K$ such that $a < v$. We set $A := \{a_i \in AL \mid a_i \leqq a\}$. Let a_1 be an atom of $V \backslash A$ and let a_2 be an atom of $AL \backslash V$. We define a dissimilarity coefficient $d' : ALV \to M$ by

$$d'(a_{ij}) := \begin{cases} m & \text{if } i = 1 \text{ and } a_j \in AL \backslash V \\ 0 & \text{else} \end{cases} \quad \text{for all } a_{ij} \in ALV.$$

This definition implies the following inequality:

$$(+) \qquad \operatorname{diam}_{d'}(\sup\{v, a_2\}) = m > \max\{\operatorname{diam}_{d'}(v), \operatorname{diam}_{d'}(\sup\{a, a_2\})\} = 0.$$

The proof of Corollary 2.6 shows the existence of some $N \in \mathscr{M}_K$ which contains $\sup\{a, a_2\}$. With the help of (+) we may conclude that $v \notin N$. Hence $f(v) = m$ for the N corresponding minimal dominating K-clustering function f of $\operatorname{diam}_{d'}$. On the other hand it is easy to see that $b \in N$ for all $b \in ALV \cap E_V$. Hence we have proved the inequality $f(v) = m > \max\{f(b) \mid b \in ALV, b \leqq v\} = 0$ and we may conclude that $f \notin \operatorname{Clus}^d_K(L, M)$.

(iii) Let $a < v$ be an arbitrary element of F_K which is not a coatom of $(E_V, \leqq)$. As in the proof of property (iib) we set $A := \{a_i \in AL \mid a_i \leqq a\}$. Now we define a dissimilarity coefficient $d' : ALV \to M$ by

$$d'(a_{ij}) := \begin{cases} m & \text{if } a_i, a_j \in AL \backslash A \\ 0 & \text{else} \end{cases} \quad \text{for all } a_{ij} \in ALV.$$

Our assumption on $(L, \leqq)$ implies together with the properties of a that there exist at least two atoms $a_1, a_2 \in V \backslash A$. Let N be an arbitrary element of $\mathscr{M}_K$; then the reader may easily verify that $\sup\{a, a_1\} \notin N$ or that $\sup\{a, a_2\} \notin N$. On the other hand $a_{it} \in N$ for all atoms $a_i, a_t \in V$. Hence $f(\sup\{a, a_1\}) = m > \max\{f(b) \mid b \in ALV, b \leqq \sup\{a, a_1\}\} = 0$ or $f(\sup\{a, a_2\}) = m > \max\{f(b) \mid b \in ALV, b \leqq \sup\{a, a_2\}\} = 0$ for all minimal dominating K-clustering functions f of diam_d, and we may conclude that $f \notin \operatorname{Clus}^d_K(L, M)$ for all minimal dominating K-clustering functions f of $\operatorname{diam}_{d'}$.

(iv) If $|K^+| = 0$ then $a_{ij} \in N$ for all $a_i, a_j \in AL$ and all $N \in \mathscr{M}_K$ (cf. the example of § 2). Hence for every $a \in L$ and every minimal dominating K-clustering function f of

diam_d the following equality holds:

$$\mathrm{diam}_d(a)=\max\{f(b)|b\in ALV,\ b\leqq a\}.$$

This implies that $f\notin \mathrm{Clus}_K^d\ (L,M)$ for all minimal dominating K-clustering functions of diam_d.

Supplementary remarks. 1. Let $d:ALV\rightarrow M$ be an arbitrary dissimilarity coefficient. Because of the characterization theorem the following problem is very natural:

Determine all minimal dominating (*weakly*) *K-ultrametrics of d.* This problem will be solved completely in a forthcoming paper.

2. The reader, especially the user of clustering techniques, may miss the consideration of algorithms which realize methods which produce for a given clustering function its minimal dominating K-clustering functions.

On the other hand the reader may perhaps notice that Theorem 2.4, especially the conditions (iv) and (vi), can actually be used to develop the desired algorithms the author is looking forward to describe these methods in a forthcoming paper.

REFERENCES

[1] T. S. Blyth and M. F. Janowitz, *Residuation Theory*, Pergamon Press, London, 1972.

[2] H. Bock, *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Göttingen, 1974.

[3] G. Herden, *Some aspects of clustering functions*, this Journal, 5 (1984), pp. 101–116.

[4] L. Hubert, *Some applications of graph theory to clustering*, Psychometrika 39 (1974), pp. 283–309.

[5] ———, *A set theoretical approach to the problem of hierarchical clustering*, J. Math. Psychol., 15 (1977), pp. 70–88.

[6] M. F. Janowitz, *An order theoretic model for cluster analysis*, SIAM J. Appl. Math., 34 (1978), pp. 55–72.

[7] N. Jardine and R. Sibson, *A model for taxonomy*, Math. Biosci., 2 (1968), pp. 465–482.

[8] ———, *Mathematical Taxonomy*, John Wiley, New York, 1971.

[9] D. W. Matula, *Graph theoretic techniques for cluster analysis algorithms*, in Classification and Clustering, Van Ryzin, ed., Academic Press, New York, 1977, pp. 95–129.

[10] F. J. Rohlf, *Graphs implied by the Jardine–Sibson overlapping methods*, B_k, J. Amer. Statist. Assoc., 69 (1974), pp. 705–710.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

006

ON THE MAXIMAL NUMBER OF STRONGLY INDEPENDENT VERTICES IN A RANDOM ACYCLIC DIRECTED GRAPH*

AMNON B. BARAK† AND PAUL ERDÖS‡

Abstract. Let $\mathscr{A}_n$ denote a random acyclic directed graph which is obtained from a random graph with vertex set $\{1, 2, \cdots, n\}$, such that each edge is present with a prescribed probability p and all the edges are directed from higher to lower indexed vertices. Define a subset of vertices in $\mathscr{A}_n$ to be *strongly independent* if there is no directed path between any pair of vertices in the subset. We show that the sequence $\mathscr{I}(\mathscr{A}_n)$, the number of vertices in the largest strongly independent vertex subset of $\mathscr{A}_n$ satisfies with probability tending to 1,

$$\frac{\mathscr{I}(\mathscr{A}_n)}{\sqrt{\log n}} \to \frac{\sqrt{2}}{\sqrt{\log 1/q}} \quad \text{as } n \to \infty,$$

where $q = 1 - p$.

Key words. acyclic directed graphs, random graphs, independent vertices

CR categories. 5.32, 5.5

1. Introduction. A random graph is a graph with vertex set $\mathbb{N}$, the set of natural numbers, such that each pair of vertices is joined by an edge with a prescribed probability p, independently of the presence or absence of any other edges. We assume no loops or multiple edges. A random acyclic directed graph is a random graph in which all the edges are directed such that there are no directed cycles.

In this paper we consider the class $\mathscr{A}$ of random acyclic directed graphs which are obtained from random graphs by directing all the edges from higher to lower indexed vertices. In other words the random variables e_{ij}, $1 \leqq j < i$, defined by

$$e_{ij} = \begin{cases} 1 & \text{if there is an edge from vertex } i \text{ to vertex } j \text{ in } \mathscr{A}, \\ 0 & \text{otherwise,} \end{cases}$$

are independent random variables with $P\{e_{ij} = 1\} = p$ and $P\{e_{ij} = 0\} = 1 - p = q$. Let $\mathscr{A}_n$ denote a subgraph of $\mathscr{A}$ spanned by the vertices $\{1, 2, \cdots, n\}$.

DEFINITION. Two vertices i, j of $\mathscr{A}_n$ are called *strongly independent* (independent) if there is no directed path (edge) from i to j $(i > j)$.

Notice that the transitive closure of our random graph is a partially ordered set (poset). Two vertices in the graph are strongly independent iff they are incomparable in this poset. A set of vertices which are pairwise strongly independent correspond to an antichain in the poset and vice versa.

Let $\mathscr{I}(\mathscr{A}_n)$ denote the number of vertices in the largest strongly independent subset of $\mathscr{A}_n$. Then in this paper we prove that with probability tending to 1, the sequence $\mathscr{I}(\mathscr{A}_n)$ satisfies:

$$\frac{\mathscr{I}(\mathscr{A}_n)}{\sqrt{\log n}} \to \frac{\sqrt{2}}{\sqrt{\log 1/q}} \quad \text{as } n \to \infty.$$

The applications of these results could be in the fields of operation research, scheduling theory and parallel computation, since several problems which may be formulated in terms of acyclic directed graphs have solutions which are specified by the maximal number of strongly independent vertices.

* Received by the editors April 9, 1981, and in final revised form September 6, 1983.

† Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem 91904, Israel.
‡ Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary.

We note that random (undirected) graphs of the kind used in this paper were investigated in connection with cliques [1], coloring [3] and complete subgraphs [4]. Random graphs of a slightly different kind were investigated in detail in [2].

2. Strongly independent vertex sets. In this section we find lower and upper bounds for k, the number of vertices in strongly independent subsets of $\mathcal{A}_n$.

Lower bound. Consider the following subsets of *k consecutive* vertices in $\mathcal{A}_n$: $\{1, 2, \cdots, k\}$, $\{k+1, k+2, \cdots, 2k\}$, $\{2k+1, 2k+2, \cdots, 3k\}, \cdots$. Then the number of these subsets is $\lceil n/k \rceil$. The probability that a subset is independent is $q^{\binom{k}{2}}$. Note that in this case the subset is also strongly independent. The probability that a subset is not independent is $1-q^{\binom{k}{2}}$. Also, the probability that none of the subsets are independent is:

$$(1-q^{\binom{k}{2}})^{\lceil n/k \rceil} \approx (1-q^{\binom{k}{2}})^{n/k}.$$

Since $1-x \leqq e^{-x}$ if $x \geqq 0$, we have

$$(1-q^{\binom{k}{2}})^{n/k} \leqq \exp(-q^{\binom{k}{2}} n/k).$$

This probability tends to zero if

$$q^{\binom{k}{2}} n/k \to \infty \quad \text{as } n \to \infty,$$

which is implied if

$$\log n - \log k + \binom{k}{2} \log q \to \infty \quad \text{as } n \to \infty,$$

or

$$\log n - \log k - \frac{k(k-1)}{2} \log \frac{1}{q} \to \infty \quad \text{as } n \to \infty.$$

Let

$$k = \lfloor K_n - \varepsilon \rfloor,$$

where ε is a positive constant and

$$K_n = \sqrt{\frac{2 \log n}{\log(1/q)} + \frac{1}{4}} + \frac{1}{2}. \tag{2.1}$$

Then

$$\log n - \frac{K_n(K_n-1)}{2} \log \frac{1}{q} = 0,$$

therefore

$$\log n - \log k - \frac{k(k-1)}{2} \log \frac{1}{q} \geqq \varepsilon K_n \log \frac{1}{q} - \frac{\varepsilon^2+\varepsilon}{2} \log \frac{1}{q} - \log(K_n - \varepsilon) \to \infty \quad \text{as } n \to \infty,$$

for every fixed value $\varepsilon > 0$. We have proved:

THEOREM 1. *Let $\mathcal{A}_n$ be a random acyclic directed graph. Then the probability that $\mathcal{A}_n$ has no strongly independent vertex subset of size $k < K_n$ tends to zero as $n \to \infty$.*

Upper bound. Let $a_1 < a_2 < \cdots < a_k$ be a subset of k vertices in $\mathscr{A}_n$ and let $E_p(n, k)$ denote the expectation for the number of subsets with k strongly independent vertices in $\mathscr{A}_n$.

The strategy of the proof is to consider four different cases of distances between a_k and a_1. First, we assume $a_k - a_1 > k^4$; next we consider $Ck \log k \leqq a_k - a_1 \leqq k^4$ where C is a constant. In the third case we consider $a_k - a_1 < Ck \log k$ and finally in the fourth case $a_k - a_1 \leqq Mk$, where M is a positive constant. In each case we prove that the probability that there is a strongly independent vertex subset of size $k \geqq K_n$, tends to zero as $n \to \infty$.

Case 1. Let $a_k - a_1 > k^4$. In this case the number of subsets of k vertices in $\mathscr{A}_n$ is bounded by $\binom{n}{k}$. The probability that each subset is strongly independent is bounded by the product of the probability that there is no directed path of length 1 from a_k to a_1 and the probability that there is no directed paths of length 2 from a_k to a_1, through at least $k^4 - k$ vertices which are not in the subset. These probabilities are q and $1 - p^2$ respectively. Therefore,

$$E_p(n, k) \leqq \binom{n}{k} q(1-p^2)^{k^4-k} \leqq \frac{n^k}{k!} q(1-p^2)^{k^4-k} \leqq n^k(1-p^2)^{k^4-k}.$$

This expectation tends to zero as $n \to \infty$ if

$$k \log n + (k^4 - k) \log(1-p^2) \to -\infty \quad \text{as } n \to \infty.$$

Since

$$\log(1-p^2) = -|\log(1-p^2)|,$$

we must have

$$(k^3 - 1)|\log(1-p^2)| - \log n \to \infty \quad \text{as } n \to \infty,$$

which is satisfied if

$$k > \sqrt[3]{\frac{\log n}{|\log(1-p^2)|} + 1}.$$

Conclusion. If $a_k - a_1 > k^4$, then even for values of k which are smaller than K_n the probability that a_1 and a_k are strongly independent tends to zero as $n \to \infty$.

Case 2. Let $Ck \log k \leqq a_k - a_1 \leqq k^4$ where C is a constant. First, we find a bound for the possible number of different subsets of k vertices. Clearly, for each vertex of $\mathscr{A}_n$ there are at most k^4 different subsets, from which we can choose k vertices. As a result, the number of subsets with k verices in $\mathscr{A}_n$ is bounded by $\binom{k^4}{k} nk^4$.

Next, we find the probability that each subset is strongly independent. This probability is bounded by the product of the probability that the subset is independent and the probability that there is no directed path of length 2 from a_k to a_1, through any vertex j which is not in the subset $\{a_1, a_2, \cdots, a_k\}$, for $a_1 < j < a_k$.

The expectation for the number of strongly independent vertex subsets of size k is bounded in this case by:

$$E_p(n, k) \leqq \binom{k^4}{k} nk^4 q^{\binom{k}{2}} (1-p^2)^{Ck \log k - k}$$

$$\leqq k^{4k} nq^{k(k-1)/2}(1-p^2)^{Ck \log k - k}.$$

This expectation tends to zero as $n \to \infty$ if

$$4k \log k + \log n - \frac{k(k-1)}{2} \log \frac{1}{q} + (k - Ck \log k)|\log(1-p^2)| \to -\infty \quad \text{as } n \to \infty.$$

Let $k \geqq K_n$, where K_n is defined in (2.1). Then

$$\log n - \frac{k(k-1)}{2} \log \frac{1}{q} \leqq 0.$$

Thus

$$4k \log k + (k - Ck \log k)|\log(1-p^2)| \to -\infty \quad \text{as } n \to \infty,$$

provided that

$$C > \frac{4}{|\log(1-p^2)|}. \tag{2.2}$$

Conclusion. If $Ck \log k \leqq a_k - a_1 \leqq k^4$, where C is defined in (2.2), then the probability that there is a strongly independent vertex subset of size $k \geqq K_n$ tends to zero as $n \to \infty$.

Case 3. Let $a_k - a_1 < Ck \log k$, where $C > 4/|\log(1-p^2)|$.

(a) Suppose that the interval between the first r vertices and the last r vertices in the subset $\{a_1, a_2, \cdots, a_n\}$ includes at least $(1+\alpha)k$ vertices, where α is a positive constant. In other words, we assume that

$$a_{k-r+1} - a_r \geqq (1+\alpha)k,$$

of which clearly, at least αk vertices of $\mathcal{A}_n$ are not in the subset $\{a_1, a_2, \cdots, a_k\}$.

This subset is strongly independent if it is independent and for each pair (a_i, a_{k-i+1}) for $i = 1, 2, \cdots, r$, there is no directed path of length 2 from a_i to a_{k-i+1} through αk vertices j, $a_r < j < a_{k-r+1}$, which are not in the subset.

The expectation $E_p(n, k)$ is bounded in this case by:

$$E_p(n,k) \leqq \binom{Ck \log k}{k} nCk \log k\, q^{\binom{k}{2}} (1-p^2)^{\alpha rk}$$

$$\leqq (Ck \log k)^k n q^{k(k-1)/2} (1-p^2)^{\alpha rk}.$$

This expectation tends to zero as $n \to \infty$ if

$$k \log(Ck \log k) + \log n - \frac{k(k-1)}{2} \log \frac{1}{q} - \alpha rk |\log(1-p^2)| \to -\infty \quad \text{as } n \to \infty.$$

Suppose that we choose $k \geqq K_n$. Then

$$\log n - \frac{k(k-1)}{2} \log \frac{1}{q} \leqq 0,$$

and

$$k(\log C + \log k + \log\log k - \alpha r|\log(1-p^2)|) \to -\infty \quad \text{as } n \to \infty,$$

provided that

$$\alpha r |\log(1-p^2)| > \log k + \log\log k.$$

It is therefore sufficient to choose a value of r such that

$$r \geqq (\log k)^{1+\sigma}, \tag{2.3}$$

where σ is a positive constant.

(b) Suppose that the conditions of (a) are not satisfied i.e., if $r \geqq (\log k)^{1+\sigma}$, where σ is a positive constant, then

$$a_{k-r+1} - a_r \leqq (1+\alpha)k,$$

for every positive value of α. Suppose however, that $a_r - a_1 \geqq r + Qk$ or $a_k - a_{k-r+1} \geqq r + Qk$, where Q is a positive constant to be defined. Then the subset $\{a_1, a_2, \cdots, a_n\}$ is strongly independent if it is independent and there is no directed path of length 2 from a_k to a_1, through at least Qk vertices of $\mathscr{A}_n$ which are not in the subset.

The expectation $E_p(n, k)$ is bounded in this case by:

$$E_p(n, k) \leqq \binom{Ck \log k}{2r}\binom{(1+\alpha)k}{k} nq^{\binom{k}{2}}(1-p^2)^{Qk}.$$

Note that as $n \to \infty$, for every $0 < \alpha < 1$ we can choose a positive constant β such that

$$\binom{(1+\alpha)k}{k} \leqq (1+\beta)^k. \tag{2.4}$$

Thus

$$E_p(n, k) \leqq (Ck \log k)^{2r}(1+\beta)^k nq^{k(k-1)/2}(1-p^2)^{Qk}.$$

This expectation tends to zero as $n \to \infty$ if

$$2r \log (Ck \log k) + k \log (1+\beta) + \log n - \frac{k(k-1)}{2} \log \frac{1}{q} - Qk|\log (1-p^2)| \to -\infty.$$

Suppose that we choose $k \geqq K_n$. Then

$$Qk|\log (1-p^2)| - 2r \log (Ck \log k) - k \log (1+\beta) \to \infty \quad \text{as } n \to \infty,$$

provided that

$$Q > \frac{\log (1+\beta)}{|\log (1-p^2)|}, \tag{2.5}$$

where β is a given positive constant.

Conclusion. If $a_k - a_1 < Ck \log k$ where $C > 4/|\log (1-p^2)|$, and $a_{k-r+1} - a_r \geqq (1+\alpha)k$, where r is defined in (2.3) and α is a positive constant, or $a_{k-r+1} - a_r \leqq (1+\alpha)k$ and $a_r - a_1 \geqq r + Qk$ or $a_k - a_{k-r+1} \geqq r + Qk$, where Q is defined in (2.5), then the probability that there is a strongly independent vertex subset of size $k \geqq K_n$ tends to zero as $n \to \infty$.

Case 4. Let $a_k - a_1 \leqq Mk$, where M is a positive constant to be defined. Suppose that $a_r - a_1 \leqq r + Qk$, and $a_k - a_{k-r+1} \leqq r + Qk$, where r and Q are defined in (2.3) and (2.5) respectively and that $a_{k-r+1} - a_r \leqq (1+\alpha)k$ for every value of $\alpha > 0$. Then

$$a_k - a_1 \leqq 2r + 2Qk + (1+\alpha)k < k + 2Qk + (1+\alpha)k = k(2 + 2Q + \alpha).$$

Thus $M = 2 + 2Q + \alpha$.

DEFINITION. Let $a_1 < a_2 < \cdots < a_k$ be a subset of k vertices in $\mathscr{A}_n$ such that $a_k - a_1 \leqq Mk$. Then the subset is called *nearly consecutive.*

THEOREM 2. *Let $\mathcal{A}_n$ be a random acyclic directed graph and let K_n be defined in* (2.1). *Then the probability that $\mathcal{A}_n$ has a strongly independent vertex subset of size $k \geqq K_n$ tends to zero as $n \to \infty$.*

Proof. First, note that by the previous conclusions, it is sufficient to consider only nearly consecutive subsets.

The expectation for the number of subsets with k strongly independent vertices in $\mathcal{A}_n$ is bounded in this case by

$$E_p(n, k) \leqq \binom{(1+\alpha)k}{k}\left[\binom{Qk+r}{r}\right]^2 nq^{\binom{k}{2}}.$$

Using (2.4) for some $\beta > 0$, we get

$$E_p(n, k) \leqq (1+\beta)^k (Qk+r)^{2r} nq^{k(k-1)/2}.$$

This expectation tends to zero as $n \to \infty$ if

$$k \log (1+\beta) + 2r \log (Qk+r) + \log n - \frac{k(k-1)}{2} \log \frac{1}{q} \to -\infty \quad \text{as } n \to \infty.$$

Let

$$k = \lceil K_n + \delta \rceil,$$

where δ is a positive constant. Then

$$\begin{aligned} k \log (1+\beta) &+ 2r \log (Qk+r) + \log n - \frac{k(k-1)}{2} \log \frac{1}{q} \\ &\leqq (K_n + \delta) \log (1+\beta) + 2r \log (QK_n + Q\delta + r) - \frac{\delta^2 - \delta}{2} \log \frac{1}{q} - \delta K_n \log \frac{1}{q}. \end{aligned}$$

Thus, the expectation tends to zero as $n \to \infty$ provided that

$$\delta \log \frac{1}{q} > \log (1+\beta).$$

Given $\delta > 0$, it is now sufficient to choose $\beta > 0$ so that

$$\delta > \frac{\log (1+\beta)}{\log 1/q}.$$

and the theorem is proved.

3. A maximal strongly independent vertex set. We now prove the main result of this paper.

THEOREM 3. *Let $\mathcal{A}_n$ be a random acyclic directed graph with vertex set $\{1, 2, \cdots, n\}$ and let $\mathcal{I}(\mathcal{A}_n)$ be the number of vertices in the largest (maximal) stongly independent subset of $\mathcal{A}_n$. Then with probability tending to* 1, *the sequence $\mathcal{I}(\mathcal{A}_n)$ satisfies:*

$$\frac{\mathcal{I}(\mathcal{A}_n)}{\sqrt{\log n}} \to \frac{\sqrt{2}}{\sqrt{\log 1/q}} \quad \textit{as } n \to \infty.$$

Proof. Let

$$K_n = \sqrt{\frac{2 \log n}{\log 1/q} + \frac{1}{4}} + \frac{1}{2}.$$

Then by Theorem 1:

$$P\{\mathcal{J}(\mathcal{A}_n) < \lfloor K_n - \varepsilon \rfloor\} \to 0 \quad \text{as } n \to \infty,$$

or

$$P\{\mathcal{J}(\mathcal{A}_n) \geqq \lfloor K_n - \varepsilon \rfloor\} \to 1 \quad \text{as } n \to \infty, \tag{3.1}$$

for every $\varepsilon > 0$, and by Theorem 2,

$$P\{\mathcal{J}(\mathcal{A}_n) \geqq \lceil K_n + \delta \rfloor\} \to 0 \quad \text{as } n \to \infty,$$

or

$$P\{\mathcal{J}(\mathcal{A}_n) < \lceil K_n + \delta \rfloor\} \to 1 \quad \text{as } n \to \infty. \tag{3.2}$$

for every $\delta > 0$.

From (3.1), (3.2) and the Borel–Cantelli lemma follows that as $n \to \infty$

$$\limsup_{n\to\infty} \frac{\mathcal{J}(\mathcal{A}_n)}{\sqrt{\log n}} \leqq \frac{\sqrt{2}}{\sqrt{\log 1/q}},$$

and

$$\liminf_{n\to\infty} \frac{\mathcal{J}(\mathcal{A}_n)}{\sqrt{\log n}} \geqq \frac{\sqrt{2}}{\sqrt{\log 1/q}},$$

and the theorem follows.

COROLLARY. *Suppose that the interval* $[K_n - \varepsilon, K_n + \varepsilon]$ *does not include an integer, i.e., for every integer* I, $|K_n - I| > \varepsilon$, *for some* $1 > \varepsilon > 0$.

Then

$$P\{\mathcal{J}(\mathcal{A}_n) = \lfloor K_n \rfloor\} \to 1 \quad \textit{as } n \to \infty.$$

We note that there is an absolute constant D $(0 < D < 1)$, so that the probability that one of the maximal strongly independent vertex subsets is consecutive, is greater than D. If K_n is not close to an integer, this probability tends to 1, but if K_n is close to an integer then this probability does not tend to 1.

Finally, it is interesting to note that although the obtained results are asymptotic, they hold even for small values of n. For example, for $p = 0.5$, $n = 10$, and $K_n = 3.13$, less than 8% of a sample of random acyclic directed graphs had a maximal strongly independent vertex subset larger than 3.

REFERENCES

[1] B. BOLLOBAS AND P. ERDÖS, *Cliques in random graphs*, Math. Proc. Camb. Phil. Soc., 80 (1976), pp. 419–427.

[2] P. ERDÖS AND A. RENYI, *On the evolution of random graphs*, Publ. Math. Inst. Hung. Acad. Sci., 5A (1960), pp. 17–61.

[3] G. R. GRIMMETT AND C. J. H. MCDIARMID, *On colouring random graphs*, Math. Proc. Comb. Phil. Soc., 77 (1975), pp. 313–324.

[4] D. W. MATULA, *On the complete subgraphs of a random graphs*, Proc. 2nd Conference on Combinatory Theory and Applications, Chapel Hill, NC 1970, pp. 356–369.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

007

TRIANGULATIONS OF ORIENTED MATROIDS AND CONVEX POLYTOPES*

LOUIS J. BILLERA† AND BETH SPELLMAN MUNSON‡

Abstract. We define the notion of a triangulation of an oriented matroid and show that, in the representable case, oriented matroid triangulations yield triangulations of the underlying polytopes. We then discuss a class of oriented matroid triangulations which can be generated by means of matroid lifts of a specific form.

AMS(MOS) subject classifications. 05, 52

1. Introduction. Oriented matroids provide a generalization of the order properties of points in space and have recently been used to gain a better understanding of convex polytopes [1], [12] and linear programming methods [7]. In this paper we show how one can generate triangulations of convex polytopes by means of lifts of the related oriented matroids.

In § 2 we describe a lift (the opposite of a contraction) of an oriented matroid by means of extensions, using an approach of Mason [13]. In § 3 we define the notion of triangulation for oriented matroids, and show that in the representable case, oriented matroid triangulations yield directly triangulations of the underlying polytopes. Section 4 gives a class of triangulations obtainable by means of lifts. This class is given an alternative description in § 5.

An expository version of this paper has appeared in the *Proceedings of the Silver Jubilee Conference on Combinatorics* [2].

For definitions, notation and general results concerning oriented matroids, see [3], [6]. If $M=(E,\mathcal{O})$ is an oriented matroid, we will often denote the rank of M (usually denoted $\rho(E)$) by $\rho(M)$.

If $M=(E,\mathcal{O})$ is an acyclic oriented matroid, every element $e\in E$ is contained in some positive cocircuit of M, and hence $\mathcal{K}^+(\mathcal{O}^\perp)$, the set of all positive elements of the signed cocircuit span of M, is nonempty. Let $\mathcal{F}=\{F\subseteq E|E\backslash F=\mathbf{Y}$ for some $Y\in\mathcal{K}^+(\mathcal{O}^\perp)\}$, and partially order $\mathcal{F}$ by set inclusion. $\mathcal{F}$ is then a lattice with $F_1\wedge F_2=F_1\cap F_2$ for every F_1, F_2 in $\mathcal{F}$. $\mathcal{F}$ has some very nice properties, including many properties of polytopal face lattices. Las Vergnas [9] has studied some of these properties (see also [4]), and hence we will refer to this lattice, denoted $L(M)$, as the Las Vergnas lattice of a matroid. We will refer to the elements of $\mathcal{F}$ as *faces* of $L(M)$; in particular, the points of $L(M)$ will be called *vertices* and the copoints *facets.*

If M is an acyclic oriented matroid representable over some ordered field, then (Mandel [11]) it is representable over the reals. Thus the elements of E can be viewed as points in $\mathbb{R}^n$ for some n, and M will be the matroid of affine dependencies of the points. $L(M)$ is then isomorphic to the face lattice of the polytope which is the convex hull of the set of points, and hence $L(M)$ is *polytopal.* The converse is not true, i.e., it is possible to have a matroid M such that $L(M)$ is polytopal but M is not representable. An example of such will be discussed later.

In [9], Las Vergnas defines the notion of the convex hull of a subset of the elements of a matroid: if $M=(E,\mathcal{O})$ is an oriented matroid and A is a subset of E, $\operatorname{conv} A\equiv$

* Received by the editors October 7, 1982, and in revised form September 2, 1983. This research was supported in part by National Science Foundation grants MCS 77-28392 and MCS 81-02353 and by Office of Naval Research contract N00014-75-C-0678 at Cornell University.

† Cornell University, Ithaca, New York 14853.

‡ AT&T Bell Laboratories, Holmdel, New Jersey 07733.

$A \cup \{e \in E \backslash A |$there exists $X \in \mathcal{O}$ with $X^- = \{e\}$ and $X^+ \subseteq A\}$. One can show that if M is the oriented matroid defined by the affine dependencies among the elements of a finite subset E of $\mathbb{R}^n$ and A is any subset of E, then $x \in E\backslash A$ is in conv A if and only if $x = \sum_{i=1}^k \lambda_i a_i$, where $\sum_{i=1}^k \lambda_i = 1$ and for every $i = 1, \cdots, k$, $a_i \in A$ and $\lambda_i \geqq 0$. As an easy consequence of Las Vergnas' work we have:

PROPOSITION 1.1. *If $M = (E, \mathcal{O})$ is an acyclic oriented matroid and $V \subseteq E$ is a closed set of rank 1 in M, then V is a vertex of $L(M)$ if and only if there does not exist $A \subseteq E \backslash V$ and $e \in V$ such that $e \in \operatorname{conv} A$.*

2. Extensions and lifts. An oriented matroid $N = (E, \mathcal{O})$ is said to be an *extension* of the oriented matroid M if there exists $E' \subseteq E$ such that $N\backslash E' = M$. If $|E'| = 1$, N is called a *point extension* of M. The only such extension N of M with $\rho(N) = \rho(M) + 1$ is the direct sum of M and the free matroid on one element p where $E' = \{p\}$, i.e., $N = M \oplus \mathbb{F}_{\{p\}}$. In this case F is a face of $L(N)$ if and only if either $F = E(M)$ or $F = G \cup \{p\}$ where G is a face of $L(M)$. We note that this is a generalization to oriented matroids of the construction of a pyramid with apex p and whose base has $L(M)$ as its face lattice.

We consider a slight generalization of the principal extensions described by Las Vergnas ([10, Prop. 3.1]; see also Mandel [12]). For notation, see [10]. For any flat F of M and base $B = \{e_1, \cdots, e_k\}$ for F, the *principal extension determined by* $[e_1^{\alpha_1}, \cdots, e_k^{\alpha_k}]$, where $\alpha_i \in \{+, -\}$ for $i = 1, \cdots, k$, is the extension of M by a new point p in general position on F such that if H is a hyperplane of M not containing F, and j is the least index such that $e_j \notin H$, then $e_j \in c^{\alpha_j}(H)$, where c is the localization of the extension. Each principal extension of an oriented matroid M is associated with some unoriented principal extension of the underlying matroid $\mathbf{M}$.

Principal extensions can be used to form a large class of matroid constructions which can be defined by means of bipartite graphs. Mason discusses these constructions and looks at some specific examples for unoriented matroids in [13]; here we summarize the method for unoriented matroids and discuss an extension to the case of oriented matroids.

Let $\mathbf{M} = (E, \mathscr{C})$ be any (unoriented) matroid on E, and let Γ be any bipartite graph on $V \cup E$, with partition (V, E), for V some set of nodes disjoint from E. Then Γ defines a matroid $\mathbf{N}$ on $V \cup E$ in terms of independent sets: $I \subseteq V \cup E$ is independent in $\mathbf{N}$ if and only if I is matched in Γ to an independent set in $\mathbf{M}$, i.e., there exists $J \subseteq E \backslash I$ such that each element of $I \backslash E$ is linked in Γ to a distinct element of J and $J \cup (I \cap E)$ is independent in $\mathbf{M}$. (We consider each element $e \in E$ to be linked to itself and allow for these loops to be in the matching.) The matroid $\mathbf{N}$ can be shown to be the matroid that results from doing a series of unoriented principal extensions, $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_{|V|} = \mathbf{N}$, where $\mathbf{M} = \mathbf{M}_0 = (E_0, \mathscr{C}_0)$ and for each $i = 1, \cdots, |V|$, $\mathbf{M}_i$ is the matroid on $E_{i-1} \cup \{v_i\}$ defined by adding v_i freely on the flat of $\mathbf{M}_{i-1}$ which is spanned by $Nb(v_i) = \{e \in E | (v_i, e)$ is an edge of $\Gamma\}$.

Clearly, by assigning an order to the elements of $Nb(v_i)$ for each i and assigning to each edge from v_i to e_j an $\alpha_j \in \{+, -\}$, we can use Γ to define a sequence of principal extensions of oriented matroids. Although the resulting matroid N depends heavily on the α's, the orders given to the elements of the flats and the order in which the extensions are made, $\mathbf{N}$, the (unoriented) matroid underlying N, is always the matroid that would result from making the sequence of unoriented principal extensions defined by Γ and starting with the matroid $\mathbf{M}$ underlying the oriented matroid M. Thus the independent sets of N are defined just as in the unoriented case above.

As an example of this construction, suppose $M = (E, \mathcal{O})$ is the oriented matroid defined by the affine dependencies of the points a, b, c and d in Fig. 1(a). Let Γ be

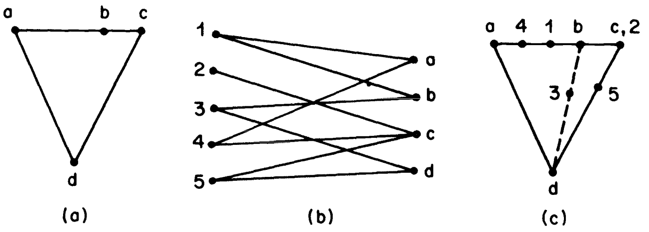


FIG. 1. *An example of using a bipartite graph to define a matroid extension.*

the bipartite graph in Fig. 1(b), and suppose for each principal extension we choose the base and the order for the base for the flat to be alphabetical and assign a positive α to each element of each flat. The resulting matroid is shown in Fig. 1(c).

This method of extending matroids can be combined with the operation of deletion to result in a lift of a given matroid.

DEFINITION 2.1. A *lift* of an oriented matroid $M=(E, \mathcal{O})$ is an oriented matroid N on $E\cup\{p\}$ for some $p\notin E$ such that $N/p=M$.

Just as any oriented matroid has many point extensions, it has many lifts. Here we give an example of a lift which will be useful to us with regard to another construction.

Let $M=(E, \mathcal{O})$ be an oriented matroid on $E=\{e_1, \cdots, e_n\}$. We define $\hat{M}$ to be $\hat{M}=N\backslash E$, where N is a multiple-element extension of M on $E\cup E'\cup\{p\}$, $E'=\{e'_1, \cdots, e'_n\}$, defined by the following sequence of extensions. Let $M_0=M\oplus\mathbb{F}_{\{p\}}$, where, as before, $\mathbb{F}_{\{p\}}$ denotes the free matroid on the element p. Let $M_1=(E_1, \mathcal{O}_1)$ be the extension of M_0 obtained by making the principal extension defined by $[p^+, e_1^+]$ with $E_1=E\cup\{p\}\cup\{e'_1\}$. In general, for $i=2, \cdots, n$, let $M_i=(E_i, \mathcal{O}_i)$ be obtained from M_{i-1} by means of a principal extension defined by $[p^+, e_i^+]$ with $E_i=E_{i-1}\cup\{e'_i\}$ and with localization c_i. Let $N=M_n$.

To see that $\hat{M}$ is a lift of M we need to show $\hat{M}/p=M$. Since $\hat{M}/p=(N\backslash E)/p=(N/p)\backslash E$, we can first consider the set of cocircuits of N/p. Since contraction in a matroid corresponds to deletion in its dual, the cocircuits of N/p are the cocircuits of N which do not contain p. The cocircuits of M_0 which do not contain p are precisely the cocircuits of M. Then (see [12], [14]) $\{Y\in\mathcal{O}^{\perp}(M_i)|p\notin\mathbf{Y}\}=A_i\cup B_i\cup -B_i\cup C_i$, where $A_i=\{Y\in\mathcal{O}^{\perp}(M_{i-1})|\mathbf{Y}\cap\{e_i, p\}=\varnothing\}$, $B_i=\{Y+(e'_i)^+|Y\in\mathcal{O}^{\perp}(M_{i-1}),\ p\notin\mathbf{Y},\ e_i\in Y^+\}$, $-B_i=\{Z|-Z\in B_i\}$ and $C_i=\{Y\in\mathcal{K}(\mathcal{O}^{\perp}(M_{i-1}))|Y$ is the conformal union of Y_1 and Y_2 in $\mathcal{O}^{\perp}(M_{i-1})$, $\rho(E\backslash\mathbf{Y})=\rho(M_0)-2$, $c_i(E_{i-1}\backslash\mathbf{Y}_1)=Y_1$, $c_i(E_{i-1}\backslash\mathbf{Y}_2)=-Y_2$, and $p\notin\mathbf{Y}_1\cup\mathbf{Y}_2\}$. (For Y a signed subset of E and $e\in E\backslash\mathbf{Y}$, $Y+e^+$ denotes the signed set having $(Y+e^+)^+=Y^+\cup\{e\}$ and $(Y+e^+)^-=Y^-$. $Y+e^-$ can be defined analogously.) But if $(\mathbf{Y}_1\cup\mathbf{Y}_2)\cap\{p\}=\varnothing$, $c_i(E_{i-1}\backslash\mathbf{Y}_1)=Y_1$ implies $e_i\in Y_1^+$, while $c_i(E_{i-1}\backslash\mathbf{Y}_2)=-Y_2$ implies $e_i\in Y_2^-$. Hence Y_1 and Y_2 are not conformal, so their conformal union is not defined. Therefore $C_i=\varnothing$ and $\{Y\in\mathcal{O}^{\perp}(M_i)|p\notin Y\}=A_i\cup B_i\cup -B_i$. This leads to the result that $Y\in\mathcal{O}^{\perp}(N/p)$ if and only if $Y\in\mathcal{O}^{\perp}(N)$ and $p\notin\mathbf{Y}$ which is the case if and only if $Y^+=\bar{Y}^+\cup\{e'_i|e_i\in\bar{Y}^+\}$ and $Y^-=\bar{Y}^-\cup\{e'_i|e_i\in\bar{Y}^-\}$ for some cocircuit $\bar{Y}$ of M. If $Y\in\mathcal{O}^{\perp}(N/p)$ is of this form, $Y\backslash E\in\mathcal{K}(\mathcal{O}^{\perp}((N/p)\backslash E))$ has $(Y\backslash E)^+=\{e'_i|e_i\in\bar{Y}^+\}$ and $(Y\backslash E)^-=\{e_i|e_i\in\bar{Y}^-\}$ for some $\bar{Y}\in\mathcal{O}^{\perp}(M)$, and we can identify these elements of $\mathcal{K}(\mathcal{O}^{\perp}(\hat{M}/p))$ with the cocircuits of M. Hence these elements must be all the cocircuits of $\hat{M}/p$, and $\hat{M}/p=M$. Thus $\hat{M}$ is a lift of M.

Note also that $(e_1, \cdots, e_n, p, e'_1, \cdots, e'_n)$ is in the cocircuit span of N, so $(p, e'_1, \cdots, e'_n)$ is in the cocircuit span of $\hat{M}$, and hence $\hat{M}$ is acyclic.

It is easy to see that this lift can be induced by the graph Γ in Fig. 2. The matroid on the elements on the right is $M\oplus\mathbb{F}_{\{p\}}$. $\hat{M}$ is obtained from the matroid on $E\cup\{p\}\cup E'$ by deleting E. (See Mason [13].)

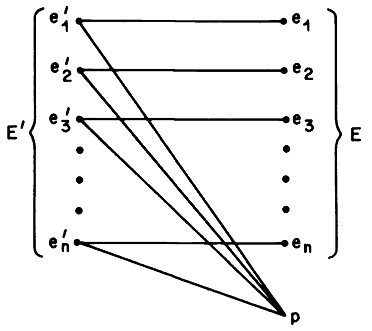


FIG. 2. *The graph* Γ *which induces the matroid lift as discussed in* § 2.

3. Triangulations. The lift defined above can be used to obtain a "triangulation" of the Las Vergnas lattice of an acyclic oriented matroid. Here we extend the definition of triangulation to this more general setting and show that in the case of a representable matroid, each matroid triangulation does correspond to a triangulation in the usual sense. We first recall the definition of a polytope triangulation.

DEFINITION 3.1. Let P be a d-dimensional polytope. A collection Δ of d-dimensional simplices is a *triangulation* of P if

(3.1.1) $\bigcup_{T\in\Delta} T=P$;

(3.1.2) T_1 and T_2 in Δ implies $T_1\cap T_2$ is a face of both T_1 and T_2.

To parallel more closely the geometric terminology we will call an independent set of cardinality n in M an *n-simplex* of M. (Note that if M is a representable matroid of rank ρ defined by the affine dependencies of a set of points in $\mathbb{R}^{\rho-1}$, $A\subseteq E$ is an n-simplex of M if and only if the points of $\mathbb{R}^{\rho-1}$ corresponding to the elements of A are affinely independent, in which case their convex hull is an $(n-1)$-dimensional simplex in $\mathbb{R}^{\rho-1}$.) Any subset of a simplex A will be called a *face* of A and is itself an m-simplex of M for some m.

DEFINITION 3.2. Let $M=(E,\mathcal{O})$ be an acyclic oriented matroid of rank ρ such that $e\in E$ implies e is a vertex of $L(M)$. A collection τ of ρ-simplices of M is a *matroid triangulation* of $L(M)$ if it satisfies

(3.2.1) $\bigcup_{A\in\tau} A=E$;

(3.2.2) for every extension N of M on $E\cup\{q\}$ and for every A, B in τ, $q\in \operatorname{conv}_N A\cap\operatorname{conv}_N B$ implies $q\in\operatorname{conv}_N(A\cap B)$;

(3.2.3) if D is a face of rank $\rho-1$ of a simplex of τ, then if D is not contained in any facet of $L(M)$, D is contained in precisely two elements of τ.

The first two properties of Definition 3.2 could be viewed as analogues in the matroid case of properties (3.1.1) and (3.1.2). However, in order that a matroid triangulation be a generalization of a polytope triangulation, the third property is also required. (In the Euclidean case, the corresponding property follows from Definition

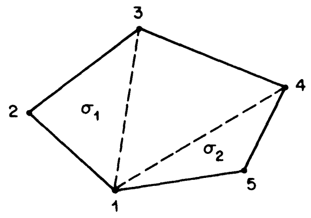


FIG. 3. *A partial triangulation of the pentagon.*

3.1.) To see this, consider the following example. Suppose $M=(E, \mathcal{O})$ is the oriented matroid of affine dependencies of the vertices of a pentagon in $\mathbb{R}^2$, and let $\sigma_1=\{1,2,3\}$ and $\sigma_2=\{1,4,5\}$ as shown in Fig. 3. Then $\{\sigma_1, \sigma_2\}$ is a collection of 3-simplices of M satisfying (3.2.1) and (3.2.2), but in the polytopal case, the pentagon P is not covered by $\operatorname{conv} \sigma_1 \cup \operatorname{conv} \sigma_2$.

Another property exhibited by polytopal triangulations is an immediate result of the definition of a matroid triangulation and is the content of the following proposition.

PROPOSITION 3.3. *Suppose τ is a matroid triangulation of the rank ρ oriented matroid $M=(E, \mathcal{O})$. Let D be a face of rank $\rho-1$ of some element of τ. Then if D is contained in some facet of $L(M)$, D is contained in precisely one simplex of τ.*

Proof. By the choice of D, D must be contained in at least one simplex of τ, say A. Suppose there exists $B \in \tau$ such that $D \subseteq B$ and $B \neq A$. Assume $D=\{e_1, e_2, \cdots, e_{\rho-1}\}$, $A=D \cup \{e_\rho\}$, and $B=D \cup \{e_{\rho+1}\}$. Since D spans a facet of $L(M)$, $E \backslash \operatorname{cl} D$ is the underlying set of a positive cocircuit Y of M.

Now form the principal extension $N=(E \cup \{p\}, \mathcal{O}')$ determined by $[e_1^+, \cdots, e_{\rho-1}^+, e_\rho^+]$. Since D, A and B are all simplices, any subset of D, A or B is again a simplex. Therefore any $(\rho-1)$-set of A or B spans a hyperplane of M and hence of N.

By the construction of N, $\{p, e_1, \cdots, e_\rho\}=A \cup \{p\}$ is the underlying set of a circuit X. For any $i \in \{1, \cdots, \rho\}$, the hyperplane $H_i=\operatorname{cl}_N \{e_1, \cdots, e_{i-1}, e_{i+1}, \cdots, e_\rho\}$ corresponds to a cocircuit of N, in which, by construction, e_i and p must agree in sign. Then since $(A \cup \{p\}) \cap [(E \cup \{p\}) \backslash H_i]=\{e_i, p\}$, e_i and p must disagree in sign in X. Thus, without loss of generality, $X^+=A$, $X^-=\{p\}$, so $p \in \operatorname{conv}_N A$.

Now consider the set $B \cup \{p\}$. As before, since p is placed in general position with respect to M, and B is a simplex, this is the underlying set of a circuit X'. Defining H_i as before for $i=1, \cdots, \rho-1$, we have $X' \cap [(E \cup \{p\}) \backslash H_i]=\{e_i, p\}$, so again e_i and p must have opposite signs in X'. Since $(\mathbf{Y} \cup \{p\}) \cap \mathbf{X}'=\{e_{\rho+1}, p\}$ and $\mathbf{Y} \cup \{p\}$ is the underlying set of a positive cocircuit in N, $e_{\rho+1}$ and p must also have opposite signs in X'. Thus we have $X'^+=B$ and $X'^-=\{p\}$, so $p \in \operatorname{conv}_N B$.

Then by (3.2.2) $p \in \operatorname{conv}_N A \cap \operatorname{conv}_N B$ implies $p \in \operatorname{conv}_N (A \cap B)=\operatorname{conv}_N D$. But this is impossible since by the construction of N, $D \cup \{p\}$ is independent. Therefore we have a contradiction, so B can not exist as supposed and D must be contained in precisely one simplex of τ. □

One could define a "triangulation" on an arbitrary lattice L to be any lattice L' with $r(L')=r(L)$ such that the set of points of L' is isomorphic to the set of points of L, and L' is the face lattice of some simplicial complex. However, in this abstract setting, property (3.2.2) has no meaning, and it is (3.2.2) which makes a triangulation interesting. We note here that being a matroid triangulation of $L(M)$ depends on M and not just on the lattice $L(M)$. This will be illustrated by an example at the end of the next section.

A perhaps more obvious parallel to Definition 3.1 would be obtained by replacing properties (3.2.1) and (3.2.3) with the following:

(3.2.1)′ for every extension N of M on $E \cup \{q\}$,
$q \in \operatorname{conv}_N E$ implies $q \in \operatorname{conv}_N A$ for some A in τ.

In fact, this definition can be shown to imply (3.2.1)–(3.2.3), but so far it is not known whether the converse is true. Thus this alternative definition may be stronger than Definition 3.2. Also, the lift construction described in the next section yields triangulations in the weaker sense which have not been shown to satisfy this stronger property. On the other hand, as will be seen in Theorem 3.4, even this weaker definition, when applied to a representable acyclic oriented matroid, is sufficient to guarantee that a triangulation of the associated polytope results.

Suppose P is a d-dimensional polytope and M is the oriented matroid defined by the affine dependencies of the vertices of P. It is fairly straightforward to show that every triangulation Δ of P such that the vertices of every d-dimensional simplex in Δ are vertices of P corresponds to a matroid triangulation of $L(M)$. One might hope that if $L(M)$ is polytopal, any matroid triangulation of $L(M)$ corresponds to a triangulation of any polytope P whose face lattice is isomorphic to $L(M)$. As we will show later, this is not generally the case; however, for representable matroids, we do have the following:

THEOREM 3.4. *Let P be a $(\rho-1)$-dimensional polytope with vertices $\{x_1, x_2, \cdots, x_m\}$, and let the oriented matroid defined by the affine dependencies of $\{x_1, x_2, \cdots, x_m\}$ be the matroid M on $\{e_1, e_2, \cdots, e_m\}$ with e_i corresponding to x_i for $i = 1, 2, \cdots, m$. Suppose τ is a matroid triangulation of $L(M)$, and let*

$$\Delta = \{\operatorname{conv}\{x_{i_1}, \cdots, x_{i_\rho}\} | \{e_{i_1}, \cdots, e_{i_\rho}\} \textit{ is a } \rho\textit{-simplex of } \tau\}.$$

Then Δ is a triangulation of P.

Proof. First note that the elements of Δ are $(\rho-1)$-dimensional simplices corresponding to ρ-simplices of τ. Suppose $T_1 = \operatorname{conv}\{x_{i_1}, \cdots, x_{i_\rho}\}$ and $T_2 = \operatorname{conv}\{x_{j_1}, \cdots, x_{j_\rho}\}$ are in Δ and $x \in T_1 \cap T_2$. Then the matroid of affine dependencies on $\{x_1, \cdots, x_m, x\}$ is an extension N of M on $E \cup \{p_x\}$ such that $p_x \in \operatorname{conv}_N \{e_{i_1}, \cdots, e_{i_\rho}\} \cap \operatorname{conv}_N \{e_{j_1}, \cdots, e_{j_\rho}\}$. By (3.2.2), $p_x \in \operatorname{conv}_N (\{e_{i_1}, \cdots, e_{i_\rho}\} \cap \{e_{j_1}, \cdots, e_{j_\rho}\}) = \operatorname{conv}_N \{e_{i_1}, \cdots, e_{i_k}\}$, say. Then, in $\mathbb{R}^{\rho-1}$, $x \in \operatorname{conv}\{x_{i_1}, \cdots, x_{i_k}\}$. Since $\{x_{i_1}, \cdots, x_{i_k}\}$ is a subset of the vertex sets of simplices T_1 and T_2, $\operatorname{conv}\{x_{i_1}, \cdots, x_{i_k}\}$ is a face of both T_1 and T_2. Note, in particular, that every $x \in T_1 \cap T_2$ must be in $\operatorname{conv}\{x_{i_1}, \cdots, x_{i_k}\}$ for the same set of x_{i_j}'s, and so $T_1 \cap T_2$ is a face of each.

Suppose $P \neq \bigcup_{T \in \Delta} T$. Then since P is the closure of the interior of P and each $T \in \Delta$ is closed in P, there exists $x \in (\operatorname{int} P) \setminus \bigcup_{T \in \Delta} T$. Choose $y \in (\bigcup_{T \in \Delta} \operatorname{int} T)$ such that y does not lie in any hyperplane H spanned by $\{x\} \cup F$, where F is a $(\rho$-$2)$-dimensional face of some $T \in \Delta$. Since each T is full dimensional, $y \in \operatorname{int} P$, and thus $[xy] \subseteq P$, where $[xy] = \{\lambda x + (1-\lambda) y | \lambda \in [0, 1]\}$. By the choice of y, $[xy]$ intersects facets of members of τ only at interior points. Also $y \in \operatorname{int} T$ for some T implies that $[xy]$ must intersect some facets of members of τ. Since there are only finitely many such facets, there exists one, say F, such that its point of intersection with $[xy]$, say z, is closest to x.

Now $z \in \operatorname{int} P$ so F cannot be contained in any facet of P. By (3.2.3) and the definition of M and Δ, F must be contained in two $(\rho-1)$-dimensional simplices of Δ, whose intersection, by the first part of this proof, must be F. Thus these two simplices lie on opposite sides of the hyperplane spanned by F; one of them, call it T, must lie on the same side as x. Since $x \notin T$, the segment $T \cap [xy]$ does not contain x. Furthermore, its endpoints are z and a point, call it w, which must lie between z and x on $[xy]$. This contradicts the choice of z, and thus $P = \bigcup_{T\in\Delta} T$. □

4. Triangulations from lifts. To show that the previously defined lift $\hat{M}$ of M can be used to obtain a matroid triangulation, we first prove the following lemmas. Recall that, by identifying the elements of E' with those of E, we may consider $\hat{M}$ to be a matroid on $E\cup\{p\}$, where $M=(E,\mathcal{O})$.

LEMMA 4.1. *H is a hyperplane of $\hat{M}$ not containing p if and only if H is a base of M.*

Proof. Let H be a hyperplane of $\hat{M}$ such that $p\notin H = \operatorname{cl} H$. Then $\rho_{\hat{M}}(H\cup\{p\}) = \rho(\hat{M})$, since $\rho_{\hat{M}}(H\cup\{p\}) > \rho_{\hat{M}}(H)$, and hence $H\cup\{p\}$ contains a base of $\hat{M}$, say $B=\{e_1,\cdots,e_k,p\}$, where $k=\rho(\hat{M})-1=\rho(M)$. Then in the graph Γ in Fig. 2, the elements of $\{e_1', e_2', \cdots, e_k', p\}$ must be linked to a base in $M_0 = M\oplus\mathbb{F}_{\{p\}}$. Since p can only be linked to itself, e_i' must be linked to e_i for each $i=1,\cdots,k$. Thus $\rho_{M_0}\{e_1,\cdots,e_k\}=\rho(M)$, and H contains a base $B=\{e_1,\cdots,e_k\}$ of M.

Suppose $e\in(E\cup\{p\})\backslash B$. In Γ, e_i' can be matched to e_i for $i=1,\cdots,\rho(M)$, and e' can be linked to p. Since $\{e_1,\cdots,e_k,p\}$ is independent in $M\oplus\mathbb{F}_{\{p\}}$, $B\cup\{e\}$ must be independent in $\hat{M}$. Thus $e\in(E\cup\{p\})\backslash B$ implies $e\notin H$, so $H=B$.

Conversely, suppose B is a base of M. Then $\rho_{\hat{M}}(B)=\rho(\hat{M})-1$, and, as above, $\rho_{\hat{M}}(B\cup\{e\})=\rho(\hat{M})$ for every $e\in(E\cup\{p\})\backslash B$. Hence B is closed in $\hat{M}$ and therefore is a hyperplane of $\hat{M}$ not containing p. □

LEMMA 4.2. *Let F be any facet of $L(\hat{M})$ not containing p, and let G be any face of F of rank $\rho(\hat{M})-2=\rho(M)-1$. Let F' be the facet of $\hat{M}$ such that $F\cap F'=G$. Then $p\in F'$ if and only if G is a subset of some facet of $L(M)$.*

Proof. By results of Mandel [12, Thm. 2.IV.13], F' is unique.

Suppose $p\in F'$. Then the positive cocircuit $Y_{F'}$ with $\mathbf{Y}_{F'}=(E\cup\{p\})\backslash F'$ of $\hat{M}$ does not contain p and hence is a positive cocircuit of M. Thus $G=F\cap F'\subseteq E$ implies $G\subseteq E\backslash\mathbf{Y}_{F'}$, so G is contained in a facet of $L(M)$.

Conversely, suppose $G\subseteq H$ for some facet H of $L(M)$. Then $\mathbf{Y}_H = E\backslash H$ is the underlying set of a positive cocircuit of M. Since $\hat{M}$ is a lift of M, Y_H is a positive cocircuit of $\hat{M}$. Therefore $(E\cup\{p\})\backslash\mathbf{Y}_H = H\cup\{p\}$ is a facet of $L(\hat{M})$. Then $H\cup\{p\}\neq F$, $(H\cup\{p\})\cap F\supseteq G$, and $\rho((H\cup\{p\})\cap F)\leqq\rho(\hat{M})-2$, so $(H\cup\{p\})\cap F=G$ and $F'=H\cup\{p\}$. □

We need the following lemma:

LEMMA 4.3. *Let $M=(E,\mathcal{O})$ be an oriented matroid. Let A, $B\subseteq E$ and suppose there exists $Y\in\mathcal{K}(\mathcal{O}^{\perp}(M))$ such that $A\backslash B\subseteq Y^+$, $B\backslash A\subseteq Y^-$, and $A\cap B\cap\mathbf{Y}=\varnothing$. Then in any extension N of M,* $\operatorname{conv}_N A\cap\operatorname{conv}_N B=\operatorname{conv}_N (A\cap B)$.

Proof. Let N be an extension of M on $E\cup\{p\}$ such that $p\in\operatorname{conv}_N A\cap\operatorname{conv}_N B$. Then there exist X_1, $X_2\in\mathcal{O}(N)$ such that $X_1^-=\{p\}$, $X_1^+\subseteq A$, $X_2^-=\{p\}$ and $X_2^+\subseteq B$. For Y as in the hypothesis, there exists $W\in\mathcal{K}(\mathcal{O}^{\perp}(N))$ such that $W\backslash p = Y$.

Suppose $p\in\mathbf{W}$. Then, since W and X_1 must be orthogonal, $p\in W^+$. But this contradicts the orthogonality of W and X_2. Thus $p\notin\mathbf{W}$.

Therefore $\mathbf{X}_1\cap\mathbf{W}\subseteq X_1^+\cap W^+$, which implies by orthogonality that $\mathbf{X}_1\cap\mathbf{W}=\varnothing$. Thus $X_1^+\subseteq A\backslash(A\backslash B)=A\cap B$, and hence $p\in\operatorname{conv}_N (A\cap B)$. □

We are now ready to prove:

THEOREM 4.4. *Let $\hat{M}=(E\cup\{p\},\hat{\mathcal{O}})$ be the lift of $M=(E,\mathcal{O})$ defined previously. Let $\tau=\{F|F$ is a facet of $L(\hat{M})$ and $p\notin F\}$. Then τ is a matroid triangulation of $L(M)$.*

Proof. We may assume, without loss of generality, that each $e\in E$ is a vertex of $L(M)$.

To show (3.2.1), we observe that since e is a vertex of $L(M)$ for every $e \in E$, $E \backslash e$ is the underlying set of a positive element in the cocircuit span of M and hence of $\hat{M}$. Then $(E \cup \{p\}) \backslash (E \backslash e) = \{e, p\}$ is a face of rank 2 of $L(\hat{M})$, so e and p must be vertices of $L(\hat{M})$. Since they are distinct vertices, there exists a facet F of $L(\hat{M})$ such that $e \in F$ and $p \notin F$. Thus there exists an element of τ containing e, so $\bigcup_{T \in \tau} T = E$.

For property (3.2.2), suppose there exists an extension M' of M on $E \cup \{q\}$ such that $q \in \operatorname{conv}_{M'} F_1 \cap \operatorname{conv}_{M'} F_2$ for simplices F_1 and F_2 of τ. Then $F_1, F_2 \in \tau$ implies there exist $Y_1, Y_2 \in \mathcal{K}(\hat{\mathcal{O}}^{\perp})$ such that $p \in \mathbf{Y}_i = Y_i^+ = E \cup \{p\} \backslash F_i$. By elimination on p between Y_1 and $-Y_2$ in the cocircuit span, there exists $Z \in \mathcal{K}(\hat{\mathcal{O}}^{\perp})$ such that $\mathbf{Z} \subseteq \mathbf{Y}_1 \cup \mathbf{Y}_2 \backslash \{p\}$, $F_1 \backslash F_2 \subseteq Z^-$ and $F_2 \backslash F_1 \subseteq Z^+$ (see Mandel [12, 2.I.1.1]). Thus, by Lemma 4.3, $q \in \operatorname{conv}_N F_1 \cap \operatorname{conv}_N F_2$ implies $q \in \operatorname{conv}(F_1 \cap F_2)$.

Finally, to show (3.2.3), suppose D is a $(\rho - 1)$-simplex of τ, so $D \subseteq F$ for some ρ-simplex F of τ. D is a $(\rho(\hat{M}) - 2)$-face of $L(\hat{M})$, so we know that there exist precisely two facets F_1 and F_2 of $L(\hat{M})$ such that $D \subseteq F_i$, $i = 1, 2$. F is one of these facets; let G be the other. By Lemma 4.2, D is contained in a facet of $L(M)$ if and only if G contains p. Thus D is contained in a facet of $L(M)$ only if D is contained in precisely one ρ-simplex of τ. □

Figure 4 shows the polytopes corresponding to $L(\hat{M})$ and the resulting triangulations of $L(M)$ for two different orderings of the vertices of a hexagon.

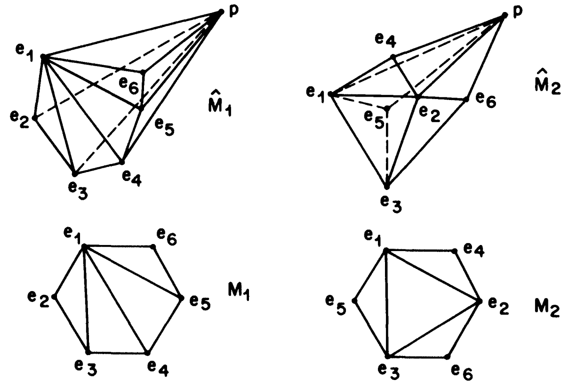


FIG. 4. *The polytopes corresponding to lifts for two different orderings of the vertices of a hexagon, and the resulting triangulations of the hexagon.*

Now consider the Vámos matroid V, described in Bland and Las Vergnas [3]. $L(V)$ is polytopal, being isomorphic to the face lattice of the polytope shown in Fig. 5, but V is not representable, since the set $\{1, 2, 3, 4\}$ is not the underlying set of any circuit of V, while $\{5, 6, 7, 8\}$ is the underlying set of a circuit. Let M be a representable matroid for which $L(M) \cong L(V)$. If we construct the lift $\hat{V}$ by forming $V \oplus \mathbb{F}_{\{p\}}$, making the principal extensions of the form $[p^+, e^+]$ in the order in which the vertices are numbered in Fig. 5, and deleting the original elements, we find that $\{1, 2, 3, 4\}$ is a facet of $L(\hat{V})$ not containing p and hence is a ρ-simplex of the matroid triangulation of τ obtained from $\hat{V}$. However, in P, the dimension of the affine hull of $\{1, 2, 3, 4\}$ is only two, so $\{1, 2, 3, 4\}$ can not be the vertex set of a 3-simplex in P. Therefore $\{1, 2, 3, 4\}$ can not be in any triangulation Δ of P. If we form τ' by deleting from τ the matroid simplex $\{1, 2, 3, 4\}$, there will be simplices of P obtained from the ρ-simplices of τ' which will not satisfy the property that $T_1 \cap T_2$ is a face of both T_1 and T_2, so τ' does not correspond to a triangulation of P either. Thus matroid triangulations

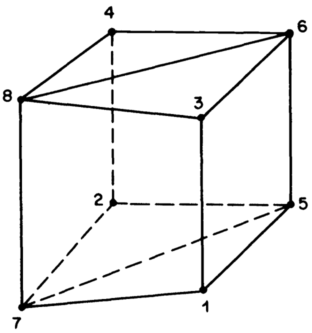


FIG. 5. *The polytope of the pre-Vámos matroid.*

of polytopal lattices of matroids do not necessarily correspond to triangulations of all polytopes having the same face lattices. The reason for this is that the triangulation (in particular, the notion of simplex) depends on the underlying matroid and not just on its face lattice.

Note that in the case where M is representable and $P(M)$ is the polytope whose matroid is M, the triangulation of $P(M)$ obtained from $\hat{M}$ can also be obtained directly without using the matroid. To do this, form the pyramid K on $P(M)$ with apex p. For each vertex v_i of $P(M)$ insert a new point v_i' on the line segment $\{\lambda v_i+(1-\lambda)p|0<\lambda<1\}$ so that v_i' and p lie strictly on the same side of every hyperplane spanned by subsets of $\{v_1', \cdots, v_{i-1}'\}$. After choosing the points $v_1', \cdots, v_n'$ the facets of the convex hull of $\{p, v_1', \cdots, v_n'\}$ which do not contain p are the simplices given by $\hat{M}$ for the triangulation of $P(M)$.

One possible extension of the results thus far would be to consider the more general notion of matroid subdivision which corresponds to subdividing a polytope into smaller polytopes which are not necessarily simplices. These can be defined by asking that the members of τ be (not necessarily minimal) spanning sets which satisfy (3.2.1)–(3.2.3) and

(3.2.4) $\forall A, B \in \tau, A \cap B$, is a face of both $L(M\backslash(E\backslash A))$ and $L(M\backslash(E\backslash B))$.

Such subdivisions would be obtained via general lifts as in Theorem 4.4.

Any lift of M is dual to an extension of $M^{\perp}$. It seems likely that the lift used here is dual to the principal extension determined by some base of $M^{\perp}$, in fact, the first base of $M^{\perp}$ given by the ordering $(e_n, e_{n-1}, \cdots, e_1)$, with each α being negative.

5. An equivalent triangulation. The lift triangulation can be shown to be equivalent to the triangulation Δ_P which is a generalization to acyclic oriented matroids of a polytopal triangulation suggested to us by Provan [15]. Let M be an acyclic oriented matroid of rank ρ, and let $v_1, v_2, \cdots, v_n$ be an ordering of the vertices of $L(M)$ such that $\{v_1, \cdots, v_\rho\}$ is an independent set and hence a ρ-simplex of M. For any independent set denote by $(v_{i_1}, \cdots, v_{i_k})$ the lattice formed by all subsets of $\{v_{i_1}, \cdots, v_{i_k}\}$. In what follows we will often identify a simplex with its lattice of subsets. Δ_P is constructed in steps as follows: let T^ρ be the set consisting of the ρ-simplex $\{v_1, \cdots, v_\rho\}$. For $i=\rho+1, \cdots, n$, define T^i by $S \in T^i$ if and only if

(*) $S \in T^{i-1}$; or

(**) $S=(v_{i_1}, \cdots, v_{i_{\rho-1}}, v_i)$ where $(v_{i_1}, \cdots, v_{i_{\rho-1}})$ is a face of rank $\rho-1$ of some ρ-simplex of T^{i-1} and the hyperplane $H(v_{i_1}, \cdots, v_{i_{\rho-1}})$ spanned by $\{v_{i_1}, \cdots, v_{i_{\rho-1}}\}$ strictly separates v_i from $\{v_1, \cdots, v_{i-1}\}$ in that $\{v_1, \cdots, v_{i-1}\}$ is contained in one closed half-space associated with $H(v_{i_1}, \cdots, v_{i_{\rho-1}})$ and v_i is in the opposite open half-space. (If H is a hyperplane of M and $Y \in \mathcal{O}^{\perp}(M)$ is such that $\mathbf{Y} = E\backslash H$, the closed half-spaces of M associated with H are $Y^+ \cup H$ and $Y^- \cup H$, and the opposite open half-spaces are Y^- and Y^+, respectively.) The triangulation Δ_P of M is the set T^n of simplices. Note that for every $i \geqq \rho$, S is a ρ-simplex of Δ_P contained in $\{v_1, \cdots, v_i\}$ if and only if S is in T^i.

THEOREM 5.1. *Let $M = (E, \mathcal{O})$ be an acyclic oriented matroid of rank ρ such that each $v \in E$ is a vertex of $L(M)$. Let $v_1, \cdots, v_n$ be an ordering of the elements of E such that $\{v_1, \cdots, v_\rho\}$ is an independent set. Then the triangulation τ of $L(M)$ obtained by constructing $\hat{M}$ is precisely the triangulation Δ_P.*

Proof. Let M_i be the oriented matroid formed in constructing $\hat{M}$ by letting $M_0 = M \oplus \mathbb{F}_{\{p\}}$ and then making the first i principal extensions of the form $[p^+, v_j^+]$, $j = 1, \cdots, i$. Let c_i be the localization of M_i with respect to M_{i-1}. Denote by τ^i those ρ-simplices of τ contained in $\{v_1, \cdots, v_i\}$. Note that if $i \leqq j$, $\{S \in \tau^i | p \notin S\} \subseteq \{S \in \tau^j | p \notin S\}$. It suffices to show that $\tau^i = T^i$ for every $i = \rho, \cdots, n$.

By the construction of $\hat{M}$, we know that F is a facet of $L(\hat{M})$ which is contained in $\{v_1, \cdots, v_1\}$ (and hence is an element of τ^i) if and only if F is a facet of $L(M_i\backslash E)$. (Again we identify the elements of $M_i\backslash E$ with $v_1, \cdots, v_i$.) Clearly $\tau^\rho = \{(v_1, \cdots, v_\rho)\} = T^\rho$.

Suppose $i > \rho$ and assume for every $j < i$, $\tau^j = T^j$. Let F be a facet of $L(M_i\backslash E)$ not containing p. If $v_i \notin F$, F is a facet of $L(M_{i-1}\backslash E)$ not containing p and hence $F \in \tau^{i-1} = T^{i-1}$.

Now $v_i \in F$ if and only if F spans a hyperplane of M_i such that one of the corresponding cocircuits, call it Y_F, has $\mathbf{Y}_F \cap (E(M_i)\backslash E) \subseteq Y_F^+$, $v_i \notin \mathbf{Y}_F$, and $p \in Y_F^+$. This is the case if and only if either a) Y_F is a cocircuit of M_{i-1} such that $c_i(E(M_{i-1}\backslash \mathbf{Y}_F)) = \varnothing$, or b) Y_F is the conformal union of cocircuits Y_1 and Y_2 of M_{i-1} such that $\rho_{M_{i-1}}(E(M_{i-1})\backslash \mathbf{Y}_F) = \rho_{M_{i-1}} - 2$, $c_i(E(M_{i-1})\backslash \mathbf{Y}_1) = Y_1$, $c_i(E(M_{i-1})\backslash \mathbf{Y}_2) = -Y_2$, and $p \in \mathbf{Y}_1 \cup \mathbf{Y}_2$. Condition a) cannot hold since $p \in \mathbf{Y}_F$ and M_i is the extension of M_{i-1} determined by $[p^+, v_i^+]$, and b) holds if and only if $E(M_{i-1})\backslash(\mathbf{Y}_F \cup E) = G$ is a face of $L(M_{i-1}\backslash E)$ of rank $\rho(M_{i-1}\backslash E) - 2 = \rho(\hat{M}) - 2$ and $p \notin \mathbf{Y}_2$. But this means that in M_i, $Y_2 + v_i'^-$ is a cocircuit and the hyperplane $H = E(M_i)\backslash(\mathbf{Y}_2 \cup \{v_i'\})$, spanned by $G \cup \{p\}$, strictly separates v_i' from $\{v_1', \cdots, v_{i-1}'\}$. Thus in $(M_i\backslash E)/p$, G separates v_i from $\{v_1, \cdots, v_{i-1}\}$. Therefore F satisfies property (**) defining elements of T^i, and for every $i = \rho, \cdots, n$, $F \in \tau^i$ if and only if $F \in T^i$. Thus the two triangulations are equivalent. □

One may note that the triangulation Δ_P, as stated, requires that in the ordering $v_1, \cdots, v_n$ of the vertices of $L(M)$, $\{v_1, \cdots, v_\rho\}$ must be an independent set, while in the construction of $\hat{M}$ no such requirement is necessary. Therefore it may seem that the set of triangulations of $L(M)$ which arise from $\hat{M}$ for different orderings of the vertices of P properly contains the set of triangulations of the form Δ_P. However, this is not the case, for one can show (Munson [14]) that if $M = (E, \mathcal{O})$ is an acyclic oriented matroid with $E = \{v_1, \cdots, v_n\}$ such that $v \in E$ implies v is a point of $L(M)$, and $\{v_1, v_2, v_{i_3}, \cdots, v_{i_\rho}\}$ is the lexicographically minimal base for M, then the triangulation τ of M obtained by constructing $\hat{M}$ with the ordering $v_1, \cdots, v_n$ of the elements of E is precisely τ', the triangulation obtained by constructing $\hat{M}'$ by means of the ordering $v_1, v_2, v_{i_3}, \cdots, v_{i_\rho}, v_3, \cdots, v_{i_3-1}, v_{i_3+1}, \cdots, v_{i_\rho-1}, v_{i_\rho+1}, \cdots, v_n$. This follows from the definition of τ and the fact that if M is as above and M_i is the partial lift of M formed by doing the sequence of principal extensions $[p^+, v_1^+], \cdots, [p^+, v_i^+]$, starting with $M_0 = M \oplus \mathbb{F}_{\{p\}}$, then $M_i\backslash E = M(\{\overline{v_1, \cdots, v_i}\})$, the lift of $M(\{v_1, \cdots, v_i\})$ formed with the ordering $v_1, \cdots, v_i$. This last statement is an easy corollary of the fact that if e, f and g are three distinct elements of E such that $\rho(\{e, f\}) = 2$, then the matroid M_1 formed by extending M principally on $[e^+, f^+]$ and then deleting g is the same as the matroid M_2 formed by extending $M\backslash g$ on $[e^+, f^+]$. We outline a proof of this result along lines suggested by A. Mandel (private communication). First one can show that if M_1 and M_2 are any oriented matroids on the same set with $\mathcal{K}(\mathcal{O}^\perp(M_1)) \subseteq \mathcal{K}(\mathcal{O}^\perp(M_2))$, then $\rho(M_1) \leqq \rho(M_2)$; if, further, $\rho(M_1) = \rho(M_2)$, then $M_1 = M_2$. For M_1 and M_2 as above, we have $\rho(M_1) = \rho(M\backslash g) = \rho(M_2)$, and it is straight-forward to check that $\mathcal{K}(\mathcal{O}^\perp(M_1)) \subseteq \mathcal{K}(\mathcal{O}^\perp(M_2))$. Thus $M_1 = M_2$.

Acknowledgments. We acknowledge here helpful suggestions from Arnaldo Mandel and anonymous referees.

REFERENCES

[1] L. J. Billera and B. S. Munson, *Polarity and inner products in oriented matroids*, Europ. J. Combinatorics, (1984), to appear.

[2] ———, *Oriented matroids and triangulations of convex polytopes*, Proc. Silver Jubilee Conference on Combinatorics, Vol. 1, W. R. Pulleyblank ed., Academic Press, New York, 1984.

[3] R. G. Bland and M. Las Vergnas, *Orientability of matroids*, J. Combin. Theory, Ser. B, 23 (1977), pp. 33–57.

[4] R. Cordovil, M. Las Vergnas and A. Mandel, *Eüler's relation, Möbius functions, and matroid identities*, Geometriae Dedicata, 12 (1982), pp. 147–162.

[5] J. Edmonds and A. Mandel, *Topology of Oriented Matroids*, Abstract 758-05-9, Notices Amer. Math. Soc., 25 (1978), p. A-510.

[6] J. Folkman and J. Lawrence, *Oriented Matroids*, J. Combin. Theory, Ser. B, 25 (1978), pp. 199–236.

[7] K. Fukuda, *Oriented matroid programming*, Ph.D. Thesis, Univ. Waterloo, Waterloo, Ontario, Canada, 1982.

[8] B. Grünbaum, *Convex Polytopes*, Wiley-Interscience, New York, 1967.

[9] M. Las Vergnas, *Convexity in oriented matroids*, J. Combin. Theory, Ser. B, 29 (1980), pp. 231–243.

[10] ————, *Extensions ponctuelles d'une geometrie combinatoire orientée*, in Problèmes combinatories et théorie des graphes, Actes du Coll. Int. C.N.R.S., n. 260, Orsay 1976, Paris, 1978, pp. 263–268.

[11] A. Mandel, *Decision process for representability of matroids and oriented matroids*, Research Rep. CORR 78-40, Dept. Combinatorics and Optimization. University of Waterloo, Waterloo, Ontario, Canada, November, 1978.

[12] ———, *Topology of oriented matroids*, Ph.D. Thesis, Univ. Waterloo, Waterloo, Ontario, Canada, 1982.

[13] J. H. Mason, *Matroids as the study of geometrical configurations*, Higher Combinatorics, M. Aigner, ed., Reidel, Dordrecht, Holland, 1977, pp. 133–176.

[14] B. S. Munson, *Face lattices of oriented matroids*, Ph.D. Thesis, Cornell University, Ithaca, NY, 1981.

[15] J. S. Provan, private communication.

[16] R. T. Rockafellar, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[17] D. J. A. Welsh, *Matroid Theory*, Academic Press, London, New York, 1976.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

008

GRAPH COLORING USING EIGENVALUE DECOMPOSITION*

BENGT ASPVALL†‡ AND JOHN R. GILBERT†§

Abstract. Determining whether the vertices of a graph can be colored using k different colors so that no two adjacent vertices receive the same color is a well-known NP-complete problem. Graph coloring is also of practical interest (for example, in estimating sparse Jacobians and in scheduling), and many heuristic algorithms have been developed. We present a heuristic algorithm based on the eigenvalue decomposition of the adjacency matrix of a graph. Eigenvectors point out "bipartite-looking" subgraphs that are used to refine the coloring to a valid coloring. The algorithm optimally colors complete k-partite graphs and certain other classes of graphs with regular structure.

AMS(MOS) classifications. 05C15, 15A18, 15A48, 68E10

1. Introduction.

Discrete mathematics and combinatorial algorithms have made considerable contributions to numerical methods in recent years. Many of these contributions have come from graph theory; for example, graphs can be used to model sparse Gaussian elimination (Rose (1972), George (1977)) and graph coloring gives efficient ways to estimate sparse Jacobian matrices (Coleman and Moré (1983)).

In this paper we will turn this around and apply some numerical analysis to a problem in graph theory, namely coloring graphs. We will present two coloring heuristics based on the eigenvalues and eigenvectors of a graph's adjacency matrix. For the present, we do not claim that these heuristics are effective or efficient enough to compete with the various purely combinatorial coloring heuristics that exist. They do, however, offer a new view of the area where numerical and discrete computation overlap. Hence we believe that they are worth investigating further.

These heuristics are novel in that they use global information about the graph rather than local information; indeed, a small enough change in the graph does not change the behavior of the heuristics. We will discuss this point further in § 5. It turns out that the heuristics perform best on graphs with very regular structures; § 4 goes into more detail.

The organization of the rest of this paper is as follows. The next section reviews some necessary background in graph theory and linear algebra. Section 3 presents the basic ideas of the paper and uses them in an algorithm to find an approximately correct two-coloring of a graph. Section 4 presents an algorithm to find a correct coloring that may use more colors than necessary; it also describes some classes of graphs for which this algorithm finds a minimum coloring. The last section considers questions of stability: what happens if the graph changes slightly, and how accurate must the numerical calculations be to get the correct discrete answer? In this section we also discuss open problems and directions for further work.

2. Background.

We begin with some standard definitions.

A *graph* $G=(V, E)$ consists of a set V of *vertices* and a set E of *edges*. An edge is an unordered pair $\{v, w\}$ of distinct vertices. If $\{v, w\}$ is an edge, vertices v and w

* Received by the editors March 10, 1983, and in revised form June 22, 1983. A preliminary version of this report was presented at the SIAM Conference on Applied Linear Algebra in Raleigh, North Carolina, on April 26, 1982.

† Computer Science Department, Cornell University, Ithaca, New York 14853.

‡ The work of this author was supported in part by the National Science Foundation under grant MCS-81-05556. Author's current address: Chr. Michelsen Institute, Bergen, Norway.

§ The work of this author was supported in part by the National Science Foundation under grant MCS-82-02948.

are *adjacent.* Edge $\{v, w\}$ is *incident* on vertices v and w, which are its *endpoints.* The number of edges incident on a vertex is its *degree.* If all the vertices in a graph have the same degree, d, the graph is *regular* of degree d.

A *path* of length k between vertices v and w is a sequence of vertices $v = v_0, v_1, \cdots, v_k = w$ such that $\{v_{i-1}, v_i\}$ is an edge for $1 \leqq i \leqq k$ and all the vertices $v_1, \cdots, v_k$ are distinct. If every pair of vertices in G is joined by a path, G is *connected.*

A *coloring* of a graph is an assignment of a color to each vertex. It is a *correct coloring* if every edge has different colors on its endpoints. We say that an edge *violates* or *satisfies* the coloring condition according to whether its endpoints have the same or different colors. A *minimum coloring* is a correct coloring with as few colors as possible; the number of colors in a minimum coloring of a graph G is its *chromatic number,* written $\chi(G)$.

If a minimum coloring can be found in polynomial time for every graph, then $P = NP$. We will be interested in polynomial time algorithms that find approximations to a minimum coloring. There are two kinds of approximations. An *approximate coloring* is a coloring that may not be correct; the fewer edges that violate the coloring condition, the better the approximation. In Fig. 1, for example, eleven edges satisfy the coloring condition and four edges violate it. Any graph with m edges can be two-colored so that more than $m/2$ edges satisfy the coloring condition (Erdös and Kleitman (1968)) and such a coloring can be found in polynomial time. If the graph is regular, the coloring can be chosen so that half the vertices are of each color.

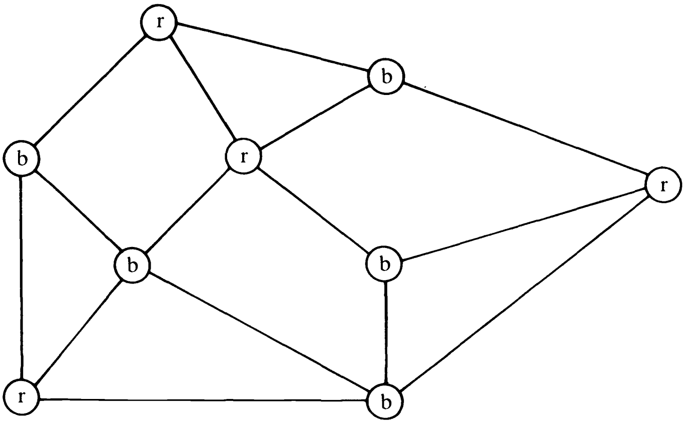


FIG. 1. *An approximate two-coloring.*

The other kind of approximation is an *approximately minimum correct coloring,* which is a correct coloring; the fewer colors it uses, the better the approximation. In Fig. 2, for example, we have used four colors to color a 3-colorable graph. Many heuristics of this sort have been proposed (Welsh and Powell (1967); Matula, Marble, and Isaacson (1972); Johnson (1974); Brélaz (1979)). Garey and Johnson (1976) showed that if a polynomial algorithm exists that always uses less than $2-\varepsilon$ times the minimum number of colors, for any positive ε, then $P = NP$. Wigderson (1982) gave a polynomial algorithm that colors any n-vertex k-colorable graph with no more than

$$k^2 n^{(k-2)/(k-1)}$$

colors.

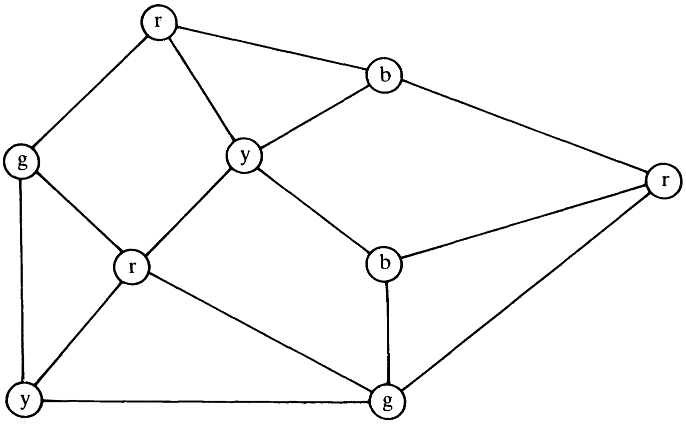


FIG. 2. *A correct but nonminimum coloring.*

Another way to say that a graph has a correct k-coloring is that its vertices can be partitioned into k sets $V_1, \cdots, V_k$ such that no edge has both endpoints in the same set. Then the graph is called k-partite, the sets are called *parts*, and the graph is sometimes written $(V_1, V_2, \cdots, V_k, E)$. A *complete k-partite graph* is a k-partite graph in which every pair of vertices in different parts is joined by an edge.

If we number the vertices of graph G from 1 to n, the *adjacency matrix* of G is the n by n matrix $A = A(G)$ whose entry a_{ij} is 1 if vertices i and j are adjacent, 0 otherwise. Finding a correct k-coloring of G is equivalent to finding a permutation matrix P such that PAP^{T} has a set of k square blocks of zeros that cover the diagonal.

The adjacency matrix is real, symmetric, and nonnegative. Therefore it has n real *eigenvectors* $\mathbf{u}_1, \cdots, \mathbf{u}_n$. (We shall use italic uppercase for matrices, italic lowercase for scalars, and boldface lowercase for vectors. The components of the vector $\mathbf{x}$ are $x_1, \cdots, x_n$.) Furthermore, the eigenvectors can be chosen to be *orthonormal*, that is, so that the inner product $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j$ is zero if $i \neq j$ ($\mathbf{u}_i$ and $\mathbf{u}_j$ are *orthogonal*), and so that $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i = 1$ ($\mathbf{u}_i$ has *unit length*). The eigenvectors can also be chosen so the first component of each is nonnegative. With each eigenvector $\mathbf{u}_i$ is associated an *eigenvalue* λ_i such that $A\mathbf{u}_i = \lambda_i \mathbf{u}_i$. An eigenvalue is *simple* if it occurs only once. Since A is real and symmetric, its eigenvalues are real. The eigenvalues are the roots of the *characteristic equation* $\det(A - \lambda I) = 0$, which is a polynomial of degree n in λ.

The *spectral radius* $\rho(A)$ of A is $\max_i |\lambda_i|$, the largest magnitude of an eigenvalue. We shall number the eigenvalues and eigenvectors so that $\lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_1$. The sum of the diagonal elements of A is its *trace*, and is equal to the sum of the eigenvalues. Since the diagonal elements of an adjacency matrix are all zero, the sum of the eigenvalues is zero. Therefore $\lambda_1 \geqq 0$ and $\lambda_n \leqq 0$.

A matrix A is *reducible* if its rows and columns can be permuted symmetrically to place a block of zeros in the lower left-hand corner, that is, if there is a permutation matrix P such that

$$PAP^{\mathrm{T}} = \begin{pmatrix} C & D \\ 0 & E \end{pmatrix}.$$

The adjacency matrix of G is irreducible if and only if G is a connected graph with at least two vertices.

There is a rich theory of nonnegative matrices; Varga (1962, Chapter 2) gives a good exposition. The results we need are due to Perron, Frobenius, and Gerschgorin, and we summarize them here along with the results mentioned above.

THEOREM 1. *Let G be a connected graph with $n > 1$ vertices and let A be G's adjacency matrix. Then*

1. *The eigenvalues $\lambda_1 \geqq \cdots \geqq \lambda_n$ of A are real.*
2. *The eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_n$ can be chosen to be orthonormal.*
3. $\sum_i \lambda_i = 0$.
4. *$\lambda_1 \geqq |\lambda_k|$ for all k; that is, λ_1 is the spectral radius of A.*
5. *$\lambda_1 > \lambda_2$; that is, the largest positive eigenvalue is simple.*
6. *$\mathbf{u}_1$ can be chosen to have positive components (we write $\mathbf{u}_1 > \mathbf{0}$).*
7. *λ_1 increases when any entry a_{ij} of A increases.*
8. *Either $\lambda_1 = \sum_j a_{ij}$ for all i or $\min_i \sum_j a_{ij} < \lambda_1 < \max_i \sum_j a_{ij}$.*

All of these except (3) *hold for any real, nonnegative, symmetric, irreducible matrix.*

Remark. Inequality (8) says that the spectral radius (and the largest positive eigenvalue) of G's adjacency matrix is bounded by the maximum and minimum degree of G's vertices.

We can represent A in terms of its eigenvalues and eigenvectors as

$$A = \sum_k \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}},$$

where the outer product $\mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$ is an n by n matrix of rank one. If the sum is in decreasing order of $|\lambda_k|$, then the mth partial sum $A^{(m)}$ minimizes $\|A - A^{(m)}\|_F$ over all rank-m matrices, where $\|B\|_F = (\sum_{i,j} b_{ij}^2)^{1/2}$ is the Frobenius norm of the matrix B. In this case $\|A - A^{(m)}\|_F$ is equal to $(\sum \lambda^2)^{1/2}$, where the sum is over the $n - m$ eigenvalues of smallest magnitude. Stewart (1973) gives details.

The *spectrum* of a graph is the multiset $\{\lambda_n, \cdots, \lambda_1\}$ of eigenvalues of its adjacency matrix. Cvetcović, Doob, and Sachs (1980) survey some known relationships between a graph's spectrum and chromatic number. Perhaps the most elegant is the following inequality, in which the lower bound is due to Hoffman (1970) and the upper bound to Wilf (1967).

THEOREM 2. *Let G be a graph with $n > 1$ vertices and let λ_1 and λ_n be its most positive and most negative eigenvalues. Then its chromatic number $\chi(G)$ satisfies*

$$\frac{\lambda_1}{-\lambda_n} + 1 \leqq \chi(G) \leqq \lambda_1 + 1.$$

The proof of this theorem is not constructive, and does not seem to lead to an efficient algorithm to color G with $\lambda_1 + 1$ colors.

Barnes and Hoffman (Barnes (1982), Barnes and Hoffman (1982)) have used the eigenvalues and eigenvectors of a graph's adjacency matrix to partition the vertices into sets that have few edges between them. This is in a sense the dual of the coloring problem. Gould and other geographers (Straffin (1980)) have used the eigenvectors corresponding to positive eigenvectors to measure the "accessibility" of cities in trade networks; again, this is in a sense dual to the coloring problem.

Our investigation of spectral coloring heuristics was suggested by an algorithm due to Moler and Morrison (1983) that divides the letters of a cipher into vowels and consonants based on the singular values and singular vectors of the matrix of digram frequencies. They observed that, in English and several other languages, pairs of adjacent letters (digrams) are more likely to consist of a vowel and a consonant than two vowelhted directed graph corresponding to the matrix of digram frequencies is approximately bipartite.

3. Two-colorings. The basic idea behind our coloring algorithm is best described in terms of bipartite graphs. In this section, we show how to color any bipartite graph correctly by examining the eigenvalue decomposition of its adjacency matrix. (Of course, bipartite graphs can be colored correctly in linear time using a simple depth-first search.) We then show how to use the ideas for approximate two-colorings.

Let $G=(V, E)$ be a bipartite graph with parts V_1 and V_2. By a suitable numbering of the vertices, the adjacency matrix $A(G)$ can be written in the form

$$A=\begin{pmatrix} 0 & B \\ B^{\mathrm{T}} & 0 \end{pmatrix}.$$

Let $\mathbf{u}$ be an eigenvector of A with eigenvalue λ. If we write $\mathbf{u}=\binom{\mathbf{x}}{\mathbf{y}}$, we have

$$\begin{pmatrix} 0 & B \\ B^{\mathrm{T}} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}=\lambda\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}.$$

We claim that $\binom{\mathbf{x}}{-\mathbf{y}}$ is also an eigenvector of A and its corresponding eigenvalue is $-\lambda$. Indeed, we have

$$\begin{pmatrix} 0 & B \\ B^{\mathrm{T}} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ -\mathbf{y} \end{pmatrix}=\lambda\begin{pmatrix} -\mathbf{x} \\ \mathbf{y} \end{pmatrix}=-\lambda\begin{pmatrix} \mathbf{x} \\ -\mathbf{y} \end{pmatrix}.$$

Thus the spectrum of a bipartite graph is symmetric around zero. In fact, Sachs (Cvetković, Doob, and Sachs (1980, Thms. 3.4, 3.11)) proved a stronger result.

THEOREM 3. *The graph G is bipartite if and only if its eigenvalue spectrum is symmetric about the origin, and this happens if and only if $\lambda_1=-\lambda_n$.*

From the Perron-Frobenius theorem, we know $\mathbf{u}_1=\binom{\mathbf{x}}{\mathbf{y}}>0$. Thus the signs of the components of $\mathbf{u}_n=\binom{\mathbf{x}}{-\mathbf{y}}$ correctly partition the vertices of G into the sets V_1 and V_2. We have the following algorithm for coloring bipartite graphs.

ALGORITHM 1 (*two-coloring*). Color the vertices according to the signs of the components of the eigenvector $\mathbf{u}_n$ corresponding to the most negative eigenvalue λ_n.

If the graph G is not bipartite, we can still partition its vertices by the signs of the components of $\mathbf{u}_n$. In the remainder of this section we give some intuition about why this might help find good colorings for arbitrary graphs. In the following section we will present an algorithm based on this idea and analyze its behavior on some classes of graphs.

Recall that the adjacency matrix A can be written in the form

$$A=\sum_k \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}},$$

and that if the sum is in decreasing order of $|\lambda_k|$ then the m-th partial sum $A^{(m)}$ is the best rank-m approximation to A in the Frobenius norm. Now suppose that λ_n is the second largest eigenvalue in magnitude (i.e., $\lambda_1>-\lambda_n\geqq\lambda_2$), and let us take a look at the best rank-1 and rank-2 approximations to the adjacency matrix A. The matrix $\mathbf{u}_1\mathbf{u}_1^{\mathrm{T}}$ has all positive elements, so $A^{(1)}=\lambda_1\mathbf{u}_1\mathbf{u}_1^{\mathrm{T}}$ has all positive elements.

Since $\mathbf{u}_n$ is orthogonal to the positive vector $\mathbf{u}_1$, it must have both positive and negative components. Let $\binom{+}{-}$ denote a vector whose first components are positive and whose remaining ones are negative. (We will arbitrarily consider zero to be positive.) Let V_1 be the set of vertices corresponding to positive components and let V_2 correspond to negative components. By a suitable numbering of the vertices, we can write $\mathbf{u}_n=\binom{+}{-}$, so

$$\lambda_n\mathbf{u}_n\mathbf{u}_n^{\mathrm{T}}=\lambda_n\begin{pmatrix} + & - \\ - & + \end{pmatrix}=\begin{pmatrix} - & + \\ + & - \end{pmatrix}.$$

When adding the second update $\lambda_n \mathbf{u}_n \mathbf{u}_n^{\mathrm{T}}$ to $A^{(1)}$, obtaining the best rank-2 approximation $A^{(2)}$, we have

$$A^{(2)} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}} + \lambda_n \mathbf{u}_n \mathbf{u}_n^{\mathrm{T}} = \begin{pmatrix} + & + \\ + & + \end{pmatrix} + \begin{pmatrix} - & + \\ + & - \end{pmatrix}. \tag{3.1}$$

The elements in the off-diagonal blocks in the two rank-1 matrices have the same signs and reinforce each other when added. In the diagonal blocks the elements added are of opposite signs and cancel each other. Thus, roughly speaking, we expect more edges between V_1 and V_2 than within V_1 and V_2, so we may view $A^{(2)}$ as the adjacency matrix for a "bipartite-looking" graph.

How good is this approximate two-coloring? Recall that every graph has an approximate two-coloring in which more than half the edges satisfy the coloring condition. The coloring above may not be this good. The following fairly weak result says that, for a class of graphs in which the most negative eigenvector partitions the vertices strongly enough, the approximate two-coloring cannot be too bad.

THEOREM 4. *Let G be connected and regular of degree $n/2$ and let its adjacency matrix A have eigenvalues $\lambda_n \leqq \cdots \leqq \lambda_1$ and eigenvectors $\mathbf{u}_n, \cdots, \mathbf{u}_1$. Let u_{jk} be the jth component of eigenvector $\mathbf{u}_k$. If $|\lambda_n| \geqq \lambda_2$ and there is a constant $0 < \eta < \sqrt{3}$ such that $|u_{in}/u_{jn}| < \eta$ for all i and j, then the number of edges that satisfy the coloring condition is more than $(3-\eta^2)/(2+2\eta^2)$ times the total number of edges.*

Proof. By Theorem 1(8) the principal eigenvalue is $\lambda_1 = n/2$, and then $\mathbf{u}_1 = (1/\sqrt{n}, \cdots, 1/\sqrt{n})^{\mathrm{T}}$. The best rank-1 approximation to A is $\lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}} = \frac{1}{2} J$, where J is the matrix of all ones. Let

$$S = (s_{ij}) = 2(A - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}}).$$

Then s_{ij} is 1 if vertices i and j are adjacent, -1 if not. Since λ_n is the eigenvalue of A of second largest magnitude, the best rank-1 approximation to S is $2\lambda_n \mathbf{u}_n \mathbf{u}_n^{\mathrm{T}}$. Let $\mathbf{s} = \sqrt{-2\lambda_n}\mathbf{u}_n$. Then $2\lambda_n \mathbf{u}_n \mathbf{u}_n^{\mathrm{T}} = -\mathbf{s}\mathbf{s}^{\mathrm{T}}$, so $\mathbf{s}\mathbf{s}^{\mathrm{T}}$ is the best rank-1 approximation to $-S$. That is, $\sum_{ij} (s_i s_j + s_{ij})^2$ is minimum over all choices of $\mathbf{s}$.

Now

$$\sum_{ij} (s_i s_j + s_{ij})^2 = \sum_{ij} s_{ij}^2 + \sum_{ij} s_i^2 s_j^2 + 2 \sum_{ij} s_i s_j s_{ij}. \tag{3.2}$$

The first sum on the right-hand side above is n^2. The second is $\|\mathbf{s}\|^4$, which is $4\lambda_n^2$. Thus the third sum,

$$t = \sum_{ij} s_i s_j s_{ij},$$

is minimum over all choices of $\mathbf{s}$ with $\|\mathbf{s}\| = \sqrt{-2\lambda_n}$. In fact, we can evaluate t: the sum on the left in equation (3.2) is $4\|A - A^{(2)}\|_F^2$, which is $4\sum_{1 \leqq i \leqq n} \lambda_i^2 - 4\lambda_1^2 - 4\lambda_n^2$. The latter sum is the square of the Frobenius norm of A, which is $n^2/2$. Since G is regular, λ_1 is $n/2$. Plugging this all into (3.2) gives $t = -4\lambda_n^2$. We will use only the fact that t is negative.

We divide the sum t into four parts. Let

$$a = |\{(i, j) : s_i s_j < 0,\ s_{ij} = 1\}|,$$

$$b = |\{(i, j) : s_i s_j > 0,\ s_{ij} = 1\}|,$$

$$c = |\{(i, j) : s_i s_j < 0,\ s_{ij} = -1\}|,$$

$$d = |\{(i, j) : s_i s_j > 0,\ s_{ij} = -1\}|.$$

Then a is twice the number of edges that satisfy the coloring condition, b is twice the number of edges that violate the coloring condition, c is twice the number of nonedges that satisfy the coloring condition, and d is twice the number of nonedges that violate the coloring condition. (Each edge and nonedge is counted once as (i, j) and once as (j, i), except that d counts each nonedge (i, i) just once.) Our goal is a lower bound on a. Since the graph is regular of degree $n/2$, we have

$$a+b=c+d=\frac{n^2}{2}. \tag{3.3}$$

Recall that the sum t above is negative. This sum has $b+c$ positive terms and $a+d$ negative terms. The ratio of the magnitudes of any two terms is less than η^2, so

$$b+c<\eta^2(a+d). \tag{3.4}$$

Now $\sum_i s_i=0$ because $\mathbf{s}$ is parallel to $\mathbf{u}_n$ and hence orthogonal to $\mathbf{u}_1$. Therefore $\sum_{ij} s_i s_j=0$. This sum has $b+d$ positive terms and $a+c$ negative terms. Thus

$$b+d<\eta^2(a+c). \tag{3.5}$$

Adding inequalities (3.4) and (3.5) and substituting $a+b$ for $c+d$ (by (3.3)) in the result yields $3b+a<\eta^2(3a+b)$. Rearranging terms gives

$$a>\frac{3-\eta^2}{2+2\eta^2}(a+b). \tag{3.6}$$

Since a is twice the number of edges that satisfy the coloring condition and $a+b$ is twice the number of edges, this completes the proof.

4. A heuristic coloring algorithm. We can use the ideas of the last section in an algorithm to find an approximately minimum correct coloring of an arbitrary connected graph. The sign pattern of one eigenvector partitions the vertices into two sets. This gives an approximate two-coloring, which we refine to a valid coloring by partitioning the vertices according to additional eigenvectors.

ALGORITHM 2 (*correct coloring*). Begin with all vertices the same color. Repeatedly select an eigenvector and use the signs of its components (with zero considered positive) to refine the coloring, until a correct coloring is obtained.

For example, if the eigenvectors $\mathbf{u}_8$, $\mathbf{u}_7$, and $\mathbf{u}_6$ of an 8-vertex graph have the sign patterns shown in Fig. 3, the vertex partition they induce has 5 parts as shown.

The algorithm does not specify which eigenvectors to use. The discussion of low rank approximations above suggests that we select eigenvectors in increasing order of

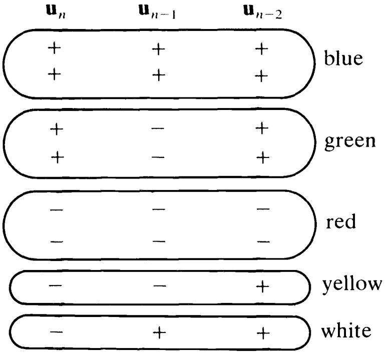


FIG. 3. *Partitioning vertices by sign pattern.*

their eigenvalues, beginning with $\mathbf{u}_n$. It turns out that, if we use enough eigenvectors, we eventually do get a correct coloring; indeed, we eventually color every vertex a different color.

THEOREM 5. *If the algorithm above is continued until it has used all the eigenvectors of $A = A(G)$, then every vertex is assigned a different color.*

Proof. Consider the n by n matrix $U = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$ whose columns are the eigenvectors of A. The color of vertex i is the sign pattern of row i of U. Since the columns of U are orthonormal, U is an orthogonal matrix and the rows $\mathbf{r}_1^{\mathrm{T}}, \cdots, \mathbf{r}_n^{\mathrm{T}}$ of U are also orthogonal; that is, $\mathbf{r}_i^{\mathrm{T}}\mathbf{r}_j = 0$ if $i \neq j$. Eigenvector $\mathbf{u}_1$ has positive components, so the first components of $\mathbf{r}_i^{\mathrm{T}}$ and $\mathbf{r}_j^{\mathrm{T}}$ are both positive. Hence their inner product can be zero only if some other component is positive in one of them and negative in the other. Thus no two rows have the same sign pattern, so no two vertices have the same color. □

The discussion in the last section suggests that eigenvectors with negative eigenvalues should partition the vertices so that many edges satisfy the coloring condition, and eigenvectors with positive eigenvalues should partition the vertices so that few edges satisfy the coloring condition. That is, a negative eigenvalue divides the vertices into approximately independent sets, and a positive eigenvalue divides the vertices into approximately complete subgraphs. We conjecture that Algorithm 2 always finds a correct coloring after considering only the eigenvectors with negative eigenvalues.

Of course, a correct coloring that gives each vertex a different color is not very surprising. We now show that this algorithm finds minimum correct colorings for a class of graphs for which we know of no purely combinatorial polynomial-time coloring algorithm. To simplify the presentation, we first consider tripartite graphs.

Let G be a tripartite graph with parts V_1, V_2, and V_3. Let the parts have r, s, and t vertices respectively. Partition the adjacency matrix of G as

$$A = \begin{pmatrix} 0 & A_{12} & A_{13} \\ A_{12}^{\mathrm{T}} & 0 & A_{23} \\ A_{13}^{\mathrm{T}} & A_{23}^{\mathrm{T}} & 0 \end{pmatrix}.$$

We call G *block regular* if in each block the row sum is constant and the column sum is constant. That is, $A_{ij}\mathbf{1} = b_{ij}\mathbf{1}$ and $A_{ij}^{\mathrm{T}}\mathbf{1} = b_{ji}\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones. In a block regular matrix, the number of edges between a given vertex v in V_i and vertices in V_j depends only on i and j (and this number is b_{ij}).

THEOREM 6. *Let G be a block regular tripartite graph. Then the adjacency matrix A has two eigenvectors with negative eigenvalues whose sign patterns correctly* 3*-color G.*

Remark. This theorem holds under the weaker assumption that there exist n vectors $\mathbf{x}_i > \mathbf{0}$, where $1 \leqq i \leqq n$, such that $A_{ij}\mathbf{x}_j = b_{ij}\mathbf{x}_i$ for $1 \leqq i, j \leqq n$.

Proof. The theorem follows from the following two lemmas. □

Let

$$B = \begin{pmatrix} 0 & b_{12} & b_{13} \\ b_{21} & 0 & b_{23} \\ b_{31} & b_{32} & 0 \end{pmatrix},$$

where b_{ij} is defined above. We call B the *block degree* matrix.

LEMMA 1. *Let A be the adjacency matrix of a block regular tripartite graph and let B be the corresponding block degree matrix. Then $(\alpha\mathbf{1}, \beta\mathbf{1}, \gamma\mathbf{1})^{\mathrm{T}}$ is an eigenvector of A with eigenvalue λ if and only if $(\alpha, \beta, \gamma)^{\mathrm{T}}$ is an eigenvector of B with eigenvalue λ.*

Proof. Assume $(\alpha\mathbf{1}, \beta\mathbf{1}, \gamma\mathbf{1})^T$ is an eigenvector of A with eigenvalue λ. We have

$$\begin{pmatrix} 0 & A_{12} & A_{13} \\ A_{12}^T & 0 & A_{23} \\ A_{13}^T & A_{23}^T & 0 \end{pmatrix}\begin{pmatrix} \alpha\mathbf{1} \\ \beta\mathbf{1} \\ \gamma\mathbf{1} \end{pmatrix} = \begin{pmatrix} (\beta b_{12}+\gamma b_{13})\mathbf{1} \\ (\alpha b_{21}+\gamma b_{23})\mathbf{1} \\ (\alpha b_{31}+\beta b_{32})\mathbf{1} \end{pmatrix} = \lambda\begin{pmatrix} \alpha\mathbf{1} \\ \beta\mathbf{1} \\ \gamma\mathbf{1} \end{pmatrix}.$$

Looking at the right equality componentwise, we see that the first r equations are identical; so are the next s equations and the last t equations. Selecting one equation from each group, we have

$$\begin{pmatrix} 0 & b_{12} & b_{13} \\ b_{21} & 0 & b_{23} \\ b_{31} & b_{32} & 0 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \lambda\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}. \qquad \square$$

LEMMA 2. *Let B be the block degree matrix for a block regular tripartite graph G. The sign patterns of two of the eigenvectors (with negative eigenvalues) of B partition the rows of B into singleton sets.*

Proof. Since B is not necessarily symmetric, we cannot apply the Perron–Frobenius theorem directly to show that the eigenvectors partition the rows. However, we can use a similarity transform to get the desired result.

Counting the total number of edges between vertices in V_1 and V_2 in two ways, we see that $rb_{12}=sb_{21}$. Similarly, $rb_{13}=tb_{31}$ and $sb_{23}=tb_{32}$. Let

$$D=\begin{pmatrix} \sqrt{r} & 0 & 0 \\ 0 & \sqrt{s} & 0 \\ 0 & 0 & \sqrt{t} \end{pmatrix}.$$

Then

$$B'=DBD^{-1}=\begin{pmatrix} 0 & \sqrt{r/s}\,b_{12} & \sqrt{r/t}\,b_{13} \\ \sqrt{r/s}\,b_{12} & 0 & \sqrt{s/t}\,b_{23} \\ \sqrt{r/t}\,b_{13} & \sqrt{s/t}\,b_{23} & 0 \end{pmatrix}$$

is a symmetric nonnegative matrix with zero trace. Furthermore $D(\alpha, \beta, \gamma)^T$ is an eigenvector of B' with eigenvalue λ if and only if $(\alpha, \beta, \gamma)^T$ is an eigenvector of B with eigenvalue λ. Since D is diagonal with positive elements, the sign patterns of the eigenvectors of B are the same as the sign patterns of those of B'. Now the same argument as in the proof of Theorem 5 shows that these sign patterns partition the rows of B' (and those of B) into singletons. □

Theorem 6 says that the adjacency matrix A has two eigenvectors (with negative eigenvalues) whose sign patterns correctly partition the vertices, but it does not say which ones they are. We can find the right eigenvectors in polynomial time, unless one of B's negative eigenvalues λ has higher multiplicity as an eigenvalue of A than as an eigenvalue of B. In this case we may not be able to tell which of A's λ-eigenvectors correspond to B's λ-eigenvector(s).

Assuming no eigenvalue of B has higher multiplicity as an eigenvalue of A, however, we do not have to try all pairs of eigenvectors. We know that the negative sum of the corresponding eigenvalues is equal to the spectral radius $\rho(A)$. (It is immediate from Lemma 1 and the proof of Lemma 2 that $\rho(A)=\rho(B)$.) Thus we can find a minimum coloring of a block regular tripartite graph by trying at most half as many colorings as there are negative eigenvalues of its adjacency matrix.

The result generalizes straightforwardly to block regular k-partite graphs, but we may no longer be able to restrict our attention to eigenvectors with negative eigenvalues.

THEOREM 7. *Let G be a block regular k-partite graph. Then there is a set of at most $k-1$ eigenvectors whose sign patterns correctly partition the vertices of the graph. Furthermore, the negative sum of the corresponding eigenvalues is equal to the spectral radius $\rho(A(G))$.*

We now have an algorithm that correctly colors block regular k-partite graphs under the multiplicity assumption above. In the remainder of this section we shall turn our attention to a restricted class of graphs for which, still under the multiplicity assumption, we can show more precisely which eigenvectors partition the vertices correctly.

Let G be a block regular k-partite graph for which b_{ij} depends only on j; that is, the off-diagonal elements in a column of the block degree matrix are all equal. We call such a graph *strongly block regular*. One example is a grid graph on a torus; see Fig. 4.

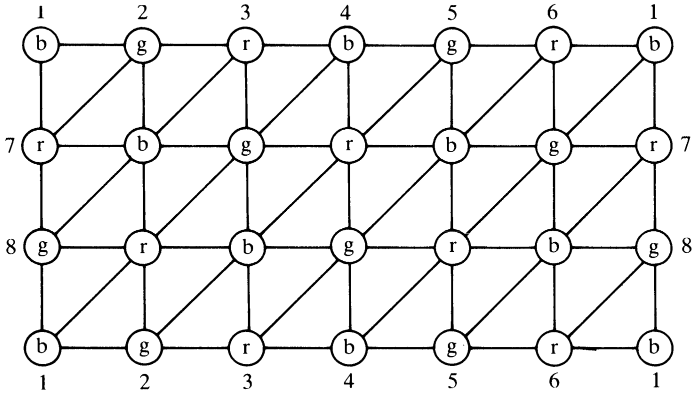


FIG. 4. *A strongly block regular graph* (*vertices with the same number are identical*).

Let the k parts have $n_1, n_2, \cdots, n_k$ vertices, where $n_1 \geqq n_2 \geqq \cdots \geqq n_k$. Let $\alpha = b_{1k}$. Counting the number of edges between different partitions as we did in the proof of Lemma 2, we see that the block degree matrix B can be written as

$$B = \frac{\alpha}{n_k} \begin{pmatrix} 0 & n_2 & \cdots & n_k \\ n_1 & 0 & \cdots & n_k \\ \vdots & \vdots & & \vdots \\ n_1 & n_2 & \cdots & 0 \end{pmatrix}.$$

Except for the factor α / n_k, B looks like the block degree matrix B' for the complete k-partite graph with the same vertex partition. The eigenvectors of B and B' are thus the same, and their eigenvalues are related by the factor α / n_k. We therefore restrict our attention to complete k-partite graphs. We have the following theorem (Smith (1970)).

THEOREM 8. *A graph has exactly one positive eigenvalue if and only if its nonisolated vertices form a complete k-partite graph for some k.*

In fact, for complete k-partite graphs the characteristic equation is known (Cvetković, Doob, and Sachs (1980, § 2.6.8)):

$$p(\lambda) = \lambda^{n-k}\left(1 - \sum_{1 \leqq i \leqq k} \frac{n_i}{\lambda + n_i}\right) \prod_{1 \leqq j \leqq k} (\lambda + n_j).$$

From this equation, it follows that $p(-n_i)\times p(-n_{i+1})\leqq 0$ with equality if and only if $n_i=n_{i+1}$. Thus the nonzero eigenvalues of a complete k-partite graph satisfy

$$-n_1\leqq\lambda_n\leqq -n_2\leqq\lambda_{n-1}\leqq\cdots\leqq\lambda_{n-k+2}\leqq -n_k<0<\lambda_1.$$

(Note that for strongly block regular graphs that are not complete k-partite graphs, the remaining $n-k-1$ eigenvalues of the adjacency matrix A might not be zero; in fact, they may be interleaved with the eigenvalues that are common with B.)

What can we say about the corresponding eigenvectors? Assume that $\lambda\neq -n_i$, for all $1\leqq i\leqq k$. Let $\mathbf{u}$ be an eigenvector of the block degree matrix B' of a complete k-partite graph. Then $(B'-\lambda I)\mathbf{u}=\mathbf{0}$. By subtracting the first equation of $(B'-\lambda I)\mathbf{u}=\mathbf{0}$ from the jth equation, it follows that $(n_1+\lambda)u_1=(n_j+\lambda)u_j$. Thus

$$\mathbf{u}=u_1(\lambda+n_1)\begin{pmatrix}1/(\lambda+n_1)\\1/(\lambda+n_2)\\\vdots\\1/(\lambda+n_k)\end{pmatrix}.$$

So for complete k-partite graphs the eigenvectors corresponding to negative eigenvalues have no zero components, and their blocks have the following sign patterns:

$$\mathbf{u}_n=(+,-,-,\cdots,-)^{\mathrm{T}},\qquad \mathbf{u}_{n-1}=(+,+,-,\cdots,-)^{\mathrm{T}},\cdots,$$
$$\mathbf{u}_{n-k+2}=(+,+,+,\cdots,-)^{\mathrm{T}}.$$

(If $n_i=n_{i+1}$ for some $1\leqq i<k$, then $\lambda_{n-i+1}=n_i$. In this case, there exists an eigenvector with zero components corresponding to vertices in certain block(s) in the partition.) From Theorem 7, we know that the sign patterns of $k-1$ of the eigenvectors correctly partition the vertices of any block regular graph. Therefore, for strongly block regular graphs, we now know exactly how the eigenvectors partition the vertices.

5. Conclusions. We have presented a new heuristic for coloring graphs. The approach is unusual in that it uses continuous mathematics for solving a combinatorial problem, and in this section we will discuss some of the implications. How accurately must the numerical computations be performed? Is our algorithm sensitive to perturbations in the input? That is, if the graph changes slightly, how does this affect the coloring? Finally, we conclude with some open problems.

How accurately must the numerical computations be performed in finite precision arithmetic? That is, can we determine whether two eigenvalues λ and μ are equal, or whether a component of an eigenvector is positive, in time polynomial in the size of the graph? The eigenvalues are the zeroes of the characteristic polynomial det $(A-\lambda I)$, which is a polynomial of degree n with coefficients bounded by $n!$ in magnitude. The eigenvalues are thus algebraic numbers of degree n. It follows from Theorem 1 of Mignotte (1982) that $\lambda\neq\mu$ implies $|\lambda-\mu|\geqq\exp(-O(n^3\log n))$. To test whether two eigenvalues are equal, we therefore need only to examine a polynomial number of bits. A similar argument holds for the components; that is, if $u_i\neq 0$, then u_i has at most a polynomial number of leading zeros. Therefore the number of bits of precision required is polynomial in n. In theory, we can obtain this precision in polynomial time by using any algorithm that is at least linearly convergent, that is, any algorithm for which each iteration produces at least some constant number of bits. For example, the QR algorithm is quadratically convergent. In practice, one might use the power method to obtain a few eigenvectors. For more details on computing the eigenvectors see Stewart (1973, Chapter 7), Parlett (1980).

Our heuristic is based on global information about the graph in the sense that each eigenvector contains information about the entire adjacency matrix A. Changing the graph slightly will thus change all or almost all eigenvectors. The important point is that we expect the changes to be small. We present some experimental and some theoretical evidence for this.

First, we showed in § 4 that we obtain correct colorings of strongly block regular graphs. The strongly block regular graph in Fig. 4 differs from a planar grid graph by only $O(\sqrt{n})$ edges. Our preliminary experiments show that various sorts of planar grid graphs are typically colored by Algorithm 2 in the minimum number of colors. Our experiments also indicate that Algorithm 1 gives, for many graphs, much better approximate two-colorings than Theorem 4 would lead us to believe.

Second, the eigenvalue decomposition of a symmetric matrix is stable. That is, if a cluster of eigenvalues is well separated from the other eigenvalues, the subspace spanned by the eigenvectors corresponding to the cluster of eigenvalues is stable with respect to perturbation of A. (We refer the interested reader to Davis and Kahan (1970), Stewart (1971), Stewart (1973, Chapter 6) for a detailed discussion.) Thus, if the distinct eigenvalues of A are sufficiently well separated, then the subspace spanned by all the eigenvectors corresponding to an eigenvalue λ is insensitive to a perturbation. (If λ is a multiple eigenvalue, the eigenvectors corresponding to λ are not unique, but the subspace they span is unique.) Therefore, if a particular set of eigenvectors of $A(G)$ partitions the vertices of G correctly, we expect that the vertices of a slightly perturbed G' will be correctly or almost correctly partitioned by a corresponding set of eigenvectors of $A(G')$.

As mentioned above, we conjecture that Algorithm 2 always colors a graph correctly (though not necessarily minimally) after considering only the eigenvectors with negative eigenvalues. If eigenvectors with negative eigenvalues partition the graph into pieces having few edges within them, then eigenvectors with positive eigenvalues can be viewed as partitioning the graph into pieces having few edges between them. (Notice what happens to (3.1) when λ_n is replaced by $\lambda_2 > 0$.) Can this idea be used to find small separators in graphs, or perhaps to find large cliques?

In summary, we believe we have demonstrated that a numerical approach can sometimes give algorithms for purely combinatorial problems. Our main hope is to stimulate further research in the broad area of intersection between continuous and discrete mathematics.

REFERENCES

EARL R. BARNES, *An algorithm for partitioning the nodes of a graph*, this Journal, 3 (1982), pp. 541–550.

EARL R. BARNES AND ALAN J. HOFFMAN, *Partitioning, spectra, and linear programming*, IBM Research Report RC 9511 (#42058), 1982.

DANIEL BRÉLAZ, *New methods to color the vertices of a graph*, Comm. ACM, 22 (1979), pp. 251–256.

THOMAS F. COLEMAN AND JORGE J. MORÉ, *Estimation of sparse Jacobian matrices and graph coloring problems*, SIAM J. Numer. Anal., 20 (1983), pp. 187–209.

DRAGOŠ M. CVETKOVIĆ, MICHAEL DOOB AND HORST SACHS, *Spectra of Graphs: Theory and Application*, Academic Press, New York, 1980.

CHANDLER DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a permutation*, III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

PAUL ERDÖS AND DANIEL J. KLEITMAN, *On coloring graphs to maximize the proportion of multicolored k-edges*, J. Combin. Theory, 5 (1968), pp. 164–169.

M. R. GAREY AND D. S. JOHNSON, *The complexity of near-optimal graph coloring*, J. Assoc. Comput. Mach., 23 (1976), pp. 43–49.

J. ALAN GEORGE, *Solution of linear systems of equations: Direct methods for finite element problems*, Sparse Matrix Techniques: Copenhagen 1976, V. A. Barker, ed., Lecture Notes in Mathematics, 572, Springer-Verlag, New York, 1977, pp. 52–101.

ALAN J. HOFFMAN, *On eigenvalues and colorings of graphs*, in Graph Theory and Its Applications, ed., Bernard Harris, Academic Press, New York, 1970, pp. 79–91.

DAVID S. JOHNSON, *Worst case behavior of graph coloring algorithms*, Proc. the 5th Southeastern Conference on Combinatorics, Graph Theory, and Computing, Utilitas Mathematica, 1974, pp. 513–527.

DAVID W. MATULA, GEORGE MARBLE AND JOEL D. ISAACSON, *Graph coloring algorithms*, in Read (1972), pp. 109–122.

MAURICE MIGNOTTE, *Identification of algebraic numbers*, J. Algorithms, 3 (1982), pp. 197–204.

CLEVE MOLER AND DONALD MORRISON, *Singular value analysis of cryptograms*, Amer. Math. Monthly, 90 (1983), pp. 78–87.

BERESFORD N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

RONALD C. READ, ed., *Graph Theory and Computing*, Academic Press, New York, 1972.

DONALD J. ROSE, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Read (1972), pp. 183–217.

JOHN H. SMITH, *Some properties of the spectrum of a graph*, in Combinatorial Structures and Their Applications, Richard Guy et al., Gordon and Breach, 1970, pp. 403–406.

G. W. STEWART, *Error bounds for approximate invariant subspaces of closed linear operators*, SIAM J. Numer. Anal., 8 (1971), pp. 796–808.

———, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

PHILIP D. STRAFFIN, JR., *Linear algebra in geography: Eigenvectors of networks*, Math. Magazine, 53 (1980), pp. 269–276.

RICHARD S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

D. J. A. WELSH AND M. B. POWELL, *An upper bound for the chromatic number of a graph and its application to timetabling problems*, Comput. J., 10 (1967), pp. 85–86.

AVI WIGDERSON, *A new approximate graph coloring algorithm.* Proc. Fourteenth Annual ACM Symposium on Theory of Computing, 1982, pp. 325–329.

H. S. WILF, *The eigenvalues of a graph and its chromatic number*, J. London Math. Soc., 42 (1967), pp. 330–332.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

A FIXED POINT APPROACH TO UNDISCOUNTED MARKOV RENEWAL PROGRAMS*

A. FEDERGRÜN† AND P. J. SCHWEITZER‡

Abstract. This paper establishes a simple existence proof for a solution to the optimality equations arising in finite undiscounted Markov Renewal Programs, by applying Brouwer's fixed point theorem to the so-called reduced value-iteration operator. Because of its simplicity, our approach lends itself to *new* existence results for more general models.

Key words. Markov renewal programs, value iteration operator, fixed point approach, communicating systems, Brouwer fixed point theorem

AMS(MOS) subject classification. 90C45

1. Introduction and summary.

This paper establishes a simple existence proof for a solution to the *pair* of optimality equations:

$$g_i = \max_{k \in K(i)} \left\{ \sum_{j=1}^{N} P_{ij}^k g_j \right\}, \qquad i = 1, \cdots, N, \tag{1}$$

$$v_i = \max_{k \in B(i)} \left\{ q_i^k - g_i T_i^k + \sum_{j=1}^{N} P_{ij}^k v_j \right\}, \qquad i = 1, \cdots, N \tag{2}$$

where

$$B(i) \equiv \left\{ k \in K(i) \,\middle|\, g_i = \sum_{j=1}^{N} P_{ij}^k g_j \right\}, \qquad i = 1, \cdots, N.$$

Here $K(i)$, $i = 1, \cdots, N$, represents a finite set of alternatives; in addition

$$T_i^k > 0, \qquad P_{ij}^k \geqq 0 \quad \text{and} \quad \sum_{j=1}^{N} P_{ij}^k = 1, \qquad i, j = 1, \cdots, N \quad \text{and} \quad k \in K(i).$$

This pair of optimality equations arises in *undiscounted* Markov Renewal Programs (MRPs) (cf. Jewell [16]) where $\Omega = \{1, \cdots, N\}$ represents the state space and $K(i)$, $i \in \Omega$, the finite set of alternatives in state i. The system is observed at instants when a transition of state occurs. The times between two consecutive transitions of state are random variables whose distributions depend both on the current state and the action chosen. When the system is observed in state i, and alternative $k \in K(i)$ is chosen, a one-step expected reward q_i^k is obtained, state $j \in \Omega$ is the next state observed with probability P_{ij}^k, and $T_i^k > 0$ represents the expected holding time until the next observation of state. (When $T_i^k = 1$ for all i, k, the model reduces to the special case of a *discrete-time* MDP [15].)

In the discounted model, the existence of a (unique) solution to the optimality equations follows by showing that the value iteration operator is a contraction mapping (cf. Denardo [5]). In the undiscounted model, this no longer holds. Only in the *unichain case*, where every stationary policy has a single subchain (closed, irreducible set of states), can a solution to the optimality equations be exhibited as the fixed point of an N-step contraction operator, albeit in a transformed though equivalent MRP (cf. Federgrün, Schweitzer and Tijms [12] and method F below).

* Received by the editors June 18, 1981, and in final revised form September 6, 1983.

† Graduate School of Business, Columbia University, New York, New York 10027.
‡ Graduate School of Management, University of Rochester, Rochester, New York 14627.

The purpose of this paper is to establish for the general *multichain* case (where some stationary policies may have multiple subchains) a similarly simple existence proof, as an application of an elementary fixed point theorem. This is obtained in two stages. We first consider models with a single *communicating* system (where every pair of recurrent states has access to each other via some stationary policy; cf. Bather [1]). Here a solution to the optimality equations is shown to follow by applying Brouwer's fixed point theorem to the *reduced* value-iteration operator (cf. (14) below). Then the result is extended to general MRPs by decomposing the state space into a hierarchical structure of communicating systems.

Other approaches have been suggested for this finite MRP-model but we believe that the fixed point approach, as opposed to others, has an immediate simplicity and lends itself easily to extensions to more general models (e.g., Federgrün, Schweitzer and Tijms [13] use this approach to derive an existence proof in the denumerable state space model under conditions which considerably weaken all existing conditions). Shapiro [24] was the first to illustrate the simplicity of using the Brouwer fixed point mapping theorem in the discounted model (see § 3). For the sake of completeness, we briefly enumerate the existence proofs in the finite undiscounted MRP-model. (Some of the methods below have been presented for discrete-time systems only; extensions to general MRPs are straightforward, however, thanks to a data-transformation, cf. § 2, transforming arbitrary MRPs into equivalent discrete-time MDPs.)

A. *The policy iteration algorithm* (cf. Howard [15], Jewell [16]). This algorithm generates a sequence of distinct policies. The associated sequence of relative value and gain rate vectors can be shown to converge to a solution of (1) and (2) in a finite number of steps. This algorithm may be interpreted as solving the optimality equations by the Newton–Raphson method. (For the discounted model this was shown by Pollatschek and Avi-Itzhak [19] and Puterman and Brumelle [20]. The same proof works in the undiscounted model with a single communicating set of states.)

B. *An optimality principle for Markovian decision processes* (cf. Schweitzer and Gavish [23]). The following optimality principle is established: if a policy is optimal in one state, it is also optimal for all states reachable from this state, using this policy. The optimality principle is used constructively to demonstrate the existence of a policy that is *optimal in every state*, and then to derive a solution to the coupled functional equations (1) and (2).

C. *The algebraic approach.* Bewley and Kohlberg [3], in the more general area of two-person zero-sum stochastic games, used Tarski's principle to show that the value vector in the discounted game has a Laurent series expansion in the Mth root of the interest rate, for some $M \geqq 1$. In the case of an MRP, M can be shown to be equal to one. The Laurent series expansion can then be used to show that the terms of this expansion satisfy a sequence of nested optimality equations, the first two of which are given by (1) and (2). Similar ideas were employed in Kohlberg [17].

D. *The limiting behavior of the total maximal discounted return vector as the interest rate tends to zero* (cf. Blackwell [4], Miller and Veinott [18], and Denardo [6]). For ordinary MRPs, a term by term Laurent series expansion can be obtained with mathematically simpler techniques. A solution to (1) and (2) then follows as under C.

E. *Nonstationary discounted value-iteration* (cf. Bather [1], Hordijk and Tijms [14]). As in our approach, one first considers models with a single *communicating* system. For such systems, a sequence of discount factors $\{\beta_n\}_{n=1}^{\infty}$ is chosen converging to 1 at an appropriate rate. The discounted value iteration scheme, where β_n is chosen as the discount factor at the nth iteration, is shown to converge to a solution of the optimality equations in the undiscounted model. The results are extended to general

MRPs by decomposing the state space into a hierarchical structure of communicating systems.

F. *Contraction mapping arguments* (cf. Federgrün et al. [12]). As mentioned above, this approach is restricted to the *unichain case* where the transition probability matrix (tpm) of each policy has a single subchain. A data-transformation is applied to transform the model into an "equivalent" one in which the value-iteration operator is an N-step contraction operator with respect to the quasi-norm:

$$sp[x] = \max_i x_i - \min_i x_i, \qquad x \in E^N. \tag{3}$$

G. *Lyapunov functions* (cf. Federgrün and Schweitzer [11]). This approach consists of exhibiting both a function whose zeros solve the optimality equations, and an iterative scheme which drives this function to zero.

H. *Successive elimination of variables* (cf. [10]). This approach eliminates one variable at a time from the functional equations, thereby leading to smaller and smaller systems of remaining equations.

We conclude this section by pointing out the plan of the paper. In § 2 we give some notation and preliminaries. In § 3 we establish an existence proof for models with a single communicating system. Also we give a (partial) characterization of the solution space of the optimality equations. The existence proof is extended to the general multichain model in § 4.

2. Preliminaries and notation. A *randomized* (stationary) policy f is characterized by a tableau $[f_{ik}]_{i\in\Omega, k\in K(i)}$ where $f_{ij} \geqq 0$ represents the probability with which alternative $k \in K(i)$ is chosen when the system is observed to be in state $i \in \Omega$, $(\sum_{k\in K(i)} f_{ik} = 1$, $i \in \Omega)$. Let S_R represents the set of all randomized policies. *Pure* (stationary) policies prescribe a *single* alternative in every state of the system (i.e., for pure policies, each $f_{ik} = 0$ or 1) and $S_P \subseteq S_R$ denotes the set of all such policies.

With each randomized policy f, we associate the N-component reward vector $q(f)$, and the $N \times N$ transition probability matrix (tpm) $P(f)$:

$$\begin{aligned} q(f)_i &\equiv \sum_{k\in K(i)} f_{ik} q_i^k, \\ P(f)_{ij} &\equiv \sum_{k\in K(i)} f_{ik} P_{ij}^k, \qquad i, j \in \Omega, \end{aligned} \tag{4}$$

$P^n(f)$ denotes the n-fold products of $P(f)$ with itself, i.e., $P^n(f) = P^{n-1}(f)P(f)$, $n \geqq 1$ and $P^0(f) = I$ (identity matrix).

Without loss of generality we assume all $T_i^k = 1$, $i \in \Omega$, $k \in K(i)$ i.e. we reduce the general MRP model to a discrete-time MDP model. This reduction is enabled by a data-transformation introduced in Schweitzer [21] and briefly reviewed below. An arbitrary MRP is transformed into an MDP which is equivalent in the sense that it has the same state and action spaces and the same set of maximal gain policies.

$$\tilde{q}_i^k \equiv \frac{q_i^k}{T_i^k}, \quad \tilde{P}_{ij}^k \equiv \tau \frac{(P_{ij}^k - \delta_{ij})}{T_i^k} + \delta_{ij}, \quad \tilde{T}_i^k \equiv 1, \qquad i, j \in \Omega, \quad k \in K(i) \tag{5}$$

where δ_{ij} represents the Kronecker delta, i.e., $\delta_{ij} = 1$ if $i = j$ and $= 0$ otherwise, and where $\tau > 0$ is chosen such that $\tau < \min_{i,k} \{T_i^k/(1 - P_{ii}^k) | (i, k)$ with $P_{ii}^k < 1\}$ so as to ensure that all $\tilde{P}_{ij}^k \geqq 0$; $i, j \in \Omega$, $k \in K(i)$. The optimality equations in the transformed model therefore read:

$$\tilde{g}_i = \max_{k\in K(i)} \sum_j \tilde{P}_{ij}^k \tilde{g}_j, \qquad i \in \Omega, \tag{6}$$

$$(7) \qquad \tilde{v}_i = \max_{k \in \tilde{B}(i)} \left\{ \tilde{q}_i^k - \tilde{g}_i + \sum_j \tilde{P}_{ij}^k \tilde{v}_j \right\}, \qquad i \in \Omega,$$

where

$$\tilde{B}(i) \equiv \left\{ k \in K(i) \,\middle|\, \tilde{g}_i = \sum_j \tilde{P}_{ij}^k \tilde{g}_j \right\}.$$

In addition, the equivalence between the original and the transformed model is reflected by the following correspondence between the solution spaces of their optimality equations:

(8a) if $\{g, v\}$ is a solution to (1), (2), then $\{g, \tau^{-1}v\}$ is a solution to (6), (7) and each $\tilde{B}(i) = B(i)$;

(8b) if $\{\tilde{g}, \tilde{v}\}$ is a solution to (6), (7), then $\{\tilde{g}, \tau\tilde{v}\}$ is a solution to (1), (2), and each $\tilde{B}(i) = B(i)$.

In view of (8b), $T_i^k = 1$, $i \in \Omega$, $k \in K(i)$, or a discrete-time MDP may be assumed without loss of generality.

In addition, § 4 shows that the general multichain model can be reduced to an MDP satisfying condition A below. First, state $i \in \Omega$ is said to *reach* state $j \in \Omega$ (under a class of policies $S \subset S_P$) if there exists a policy $f \in S_P$ with $P^n(f)_{ij} > 0$ for some $n \geqq 0$. A pair of states $i, j \in \Omega$ is said to *communicate* if i reaches j and j reaches i. A set $B \subset \Omega$ is *communicating* (*under a class of policies* $S \subset S_P$) if each pair of states in B communicates (under this class). Communication defines an equivalence relation on Ω and the equivalence classes will be referred to as the *communicating classes.*

Condition A. The set $R \equiv \{i \in \Omega | i$ is recurrent for some $P(f)$, $f \in S_R\}$ is communicating.

Condition A is equivalent to the existence of a randomized policy ψ^* which has R as its single subchain (cf. Bather [1]). It holds, for example, if every choice of pure stationary policy has a tpm with a single subchain. Moreover, the data-transformation (5) preserves condition A since it leaves the chain structure of any policy unaltered, cf. (5) or [9]. In other words, condition A holds in the transformed model if and only if it holds in the original model. In § 3, we will show that under condition A, a solution (g, v) to (6), (7) exists with $g = \gamma\mathbf{1}$, where $\mathbf{1}$ denotes the N-vector all of whose components are unity. The optimality equations thus reduce to:

$$(9) \qquad v_i = \max_{k \in K(i)} \left\{ q_i^k - \gamma + \sum_j P_{ij}^k v_j \right\}, \qquad i \in \Omega.$$

Equation (9) resembles the standard functional equations for single-chained MDPs; yet, even under condition A, it may be that several policies have multiple chains!

We conclude this section with some notation:

Let $\Pi(f)$ represent the Cesaro limit of the sequence $\{P^n(f)\}_{n=1}^{\infty}$. For each $f \in S_R$, we define the gain rate vector $g(f)$ by $g(f) = \Pi(f)q(f)$ so that $g(f)_i$ represents the long run average expected return per unit time when the initial state is i, and policy f is used. Finally we define the maximal gain rate vector g^* by

$$(10) \qquad g_i^* = \sup_{f \in S_R} g(f)_i, \qquad i \in \Omega,$$

and call a policy f *maximal gain* if $g(f) = g^*$, i.e., if it achieves all N suprema in (10) simultaneously. It follows from Denardo and Fox [7] that any (pure) policy which prescribes in each state $i \in \Omega$ an alternative $k \in B(i)$ achieving the maximum to the right of (2), is a maximal gain policy.

Suppose that condition A holds, and fix a randomized policy ψ^* which has R as its only subchain. (The existence of such a policy is easily verified, see also Bather [1].) Let $m_{ij}(i \in \Omega, j \in R)$ denote the mean number of transitions (in the transformed model) needed to go from i to j under $P(\psi^*)$:

$$m_{ij} = 1 + \sum_{t \neq j} P(\psi^*)_{it} m_{tj}, \qquad i \in \Omega, \quad j \in R \tag{11}$$

and define

$$m^* \equiv \max \{m_{ij} | i \in \Omega, j \in R\}. \tag{12}$$

Similarly, under condition A, let $\theta_i (i \in \Omega \backslash R)$ denote the maximum expected number of transitions one can stay outside of R, under any stationary policy, when starting in state i. Since the restriction of each $P(f), f \in S_R$, to the set $\Omega \backslash R$ is a transient matrix, the vector $[\theta_i, i \in \Omega \backslash R]$ may be considered as the total reward vector of a transient MDP (cf. Veinott [25]) and hence satisfies the functional equation:

$$\theta_i = \max_{k \in K(i)} \left\{ 1 + \sum_{j \in \Omega \backslash R} P_{ij}^k \theta_j \right\}, \qquad i \in \Omega \backslash R.$$

3. The average return optimality equation for models with a single communicating system. We first define the following two (value-iteration) operators from E^N into itself:

$$Tx_i \equiv \max_{k \in K(i)} \left\{ q_i^k + \sum_j P_{ij}^k x_j \right\}, \qquad i \in \Omega, \tag{13}$$

$$Qx_i \equiv Tx_i - Tx_r, \qquad i \in \Omega \quad \text{(reduced value-iteration operator)}, \tag{14}$$

where r is an arbitrary fixed state in R. Let

$$q_{\min} \equiv \min_{i,k} q_i^k, \quad q_{\max} \equiv \max_{i,k} q_i^k, \quad \Delta q \equiv q_{\max} - q_{\min} \geqq 0.$$

Our analysis is based on the construction of a compact convex subset of E^N which is closed for the Q-operator. Since this operator is continuous on E^N, Brouwer's fixed point theorem (cf. [8]) can be invoked to establish the existence of a fixed point. Our choice of this compact convex subset of E^N was partly inspired by Bather [1].

Define the following five subsets of E^N:

$$D_1 \equiv \{x \in E^N | x_i - x_j \geqq -\Delta q m_{ij};\ i \in \Omega, j \in R, i \neq j\}, \tag{15}$$

$$D_2 \equiv \{x \in E^N | Tx \leqq x + q_{\max} \mathbf{1}\}, \tag{16}$$

$$D_3 \equiv \{x \in E^N | x_i - x_j \leqq m^* \Delta q \theta_i;\ i \in \Omega \backslash R, j \in R\}, \tag{17}$$

$$D_4 \equiv \{x \in E^N | x_r = 0\}, \tag{18}$$

$$D \equiv D_1 \cap D_2 \cap D_3 \cap D_4.$$

THEOREM 1. *Let condition* A *hold. Then*

(a) *Q maps D into itself.*

(b) *D is a nonempty convex compact subset of E^N; hence Q has a fixed point on D by the Brouwer theorem.*

Proof. (a) We first show that the T-operator maps $D_1 \cap D_2 \cap D_3$ into itself. Part (a) then follows since $y \in D_1 \cap D_2 \cap D_3$ implies $y - (y_r)\mathbf{1} \in D$.

(i) *If* $x \in D_1 \cap D_2 \cap D_3$, *show* $Tx \in D_1$ *as follows*:
fix $j \in R$ and $i \neq j$. Note that

$$Tx_i \geqq q(\psi^*)_i + P_{ij}(\psi^*)x_j + \sum_{t \neq j} P(\psi^*)_{it}x_t.$$

Insert $x_t \geqq x_j - \Delta q m_{tj}$ for $t \neq j$ (in view of $x \in D_1$):

$$Tx_i \geqq q_{\min} + x_j - \Delta q \sum_{t \neq j} P(\psi^*)_{it} m_{tj}.$$

Next, insert $x_j \geqq Tx_j - q_{\max}$ (in view of $x \in D_2$) and use (11) to conclude:

$$Tx_i - Tx_j \geqq q_{\min} - q_{\max} - \Delta q \sum_{t \neq j} P(\psi^*)_{it} m_{tj} = -(\Delta q) m_{ij}.$$

(ii) *If* $x \in D_1 \cap D_2 \cap D_3$, $Tx \in D_2$ follows immediately from (16) and the monotonicity of the T-operator:

$$T^2 x = T(Tx) \leqq T(x + q_{\max}\mathbf{1}) = Tx + q_{\max}\mathbf{1}.$$

(iii) *If* $x \in D_1 \cap D_2 \cap D_3$, *show* $Tx \in D_3$ *as follows*:
fix $i \in \Omega \backslash R$ and $j \in R$. Note that

$$Tx_i = \max_{k \in K(i)} \left\{ q_i^k + \sum_{t \in R} P_{it}^k x_t + \sum_{t \in \Omega \backslash R} P_{it}^k x_t \right\}. \tag{19}$$

Let $s \in R$ satisfy $x_s = \max_{l \in R} x_l$.

In (19), insert for $t \in R$, $x_t \leqq x_s$ and insert for $t \in \Omega \backslash R$, in view of $x \in D_3$, $x_t \leqq m^* \Delta q \theta_t + x_s$, to conclude:

$$\begin{aligned} Tx_i &\leqq q_{\max} + x_s + \Delta q m^* \max_{k \in K(i)} \left[\sum_{j \in \Omega \backslash R} P_{ij}^k \theta_j \right] \\ &= q_{\max} + x_s + \Delta q m^* (\theta_i - 1). \end{aligned} \tag{20}$$

Since R is closed for $P(\psi^*)$, we have

$$Tx_j \geqq q(\psi^*)_j + \sum_t P(\psi^*)_{jt} x_t \geqq q_{\min} + \sum_{t \in R} P(\psi^*)_{jt} x_t q_{\min} + \sum_{t \in R} P(\psi^*)_{jt} x_t.$$

Since $x \in D_1$, insert $x_t \geqq x_s - \Delta q m_{ts}$ for $t \in R \backslash \{s\}$:

$$\begin{aligned} Tx_j &\geqq q_{\min} + x_s - \Delta q \sum_{t \in R \backslash \{s\}} P(\psi^*)_{jt} m_{ts} \\ &= q_{\min} + x_s - \Delta q (m_{js} - 1) \geqq q_{\min} + x_s - \Delta q (m^* - 1). \end{aligned}$$

Finally, subtract this inequality from (20) to conclude $Tx \in D_3$.

(b) $\mathbf{0} \in D$, hence $D \neq \emptyset$. D_1, D_3, D_4 are convex polyhedra; D_2 is convex in view of the convexity of the T-operator, and hence their intersection D is convex as well. Next fix $l \neq r$. Choose $i = l$ and $j = r$ in (17), or $j = l$ and $i = r$ in (15) to conclude that

$$x_l \leqq m^* \Delta q \theta_l, \quad l \in \Omega \backslash R, \quad x_l \leqq \Delta q m_{rl} \quad \text{for } l \in R.$$

This, together with $x_l \geqq -\Delta q m_{lr}$ for all $l \in \Omega$ (cf. (15) with $i = l$ and $j = r$) proves the boundedness of D. Since D is also closed, D is compact. □

COROLLARY 1. *Let condition* A *hold. Let* x^* *be a fixed point of* Q. *Then* $\{v = x^*, \gamma = (Tx^*)_r\}$ *is a solution of* (9).

Theorem 2 below shows that the γ-part of the solution $\{\gamma, v\}$ is uniquely determined by $\gamma = \gamma^*$, where γ^* is the common value of the components of the maximal gain rate vector, cf. Condition A. The v-part of the solution, however, is never uniquely determined since if $\{\gamma, v\}$ solves (9), then so does $\{\gamma, v+c\mathbf{1}\}$ for any scalar c. Under Condition A, even more "degrees of freedom" in the choice of the v-vector may exist (cf. [22]). We conclude this section by showing that the solution space to (9) is included in $D_1 \cap D_2 \cap D_3$ and *hence is bounded in the sp*[ ]*-norm*, defined by (3).

THEOREM 2. *Let Condition* A *hold. Then every solution* $\{\gamma, v\}$ *of* (9) *has* $\gamma\mathbf{1} = g^*$ *and* $v \in D_1 \cap D_2 \cap D_3$. *Moreover,* $v^* \overset{\text{def}}{=} v-(v_r)\mathbf{1} \in D$ *and is a fixed point of* Q.

Proof. Fix a solution $\{\gamma, v\}$ to (9). Note that $v \geqq q(f) - \gamma\mathbf{1} + P(f)v$ for all $f \in S_R$, with equality for some $f \in S_R$. Multiply these inequalities with $\Pi(f) \geqq 0$ to obtain $\gamma\mathbf{1} = \max_{f \in S_R} g(f) = g^*$. Here we have used $\Pi(f)P(f) = \Pi(f)$ and $\Pi(f)\mathbf{1} = \mathbf{1}$. For future use we also note

$$(21) \qquad q_{\min} \leqq \gamma \leqq q_{\max}.$$

Next subtract $v_r = Tv_r - \gamma$ from $v = Tv - \gamma\mathbf{1}$ to obtain $v^* = v - v_r\mathbf{1} = Tv - (Tv_r)\mathbf{1}$, and verify that $Qv^* = T(v - v_r\mathbf{1}) - [T(v - v_r\mathbf{1})_r]\mathbf{1} = Tv - (Tv)_r\mathbf{1} = v^*$. This leaves us to prove that $v \in D_1 \cap D_2 \cap D_3$, with $v^* \in D$ immediately following from (18).

(i) $v \in D_1$. Fix $j \in R$. Note for all i that $v_i = Tv_i - \gamma \geqq q(\psi^*)_i - \gamma + \sum_t P(\psi^*)_{it} v_t \geqq q_{\min} - q_{\max} + \sum_t P(\psi^*)_{it} v_t$, since $\gamma \leqq q_{\max}$. Subtract v_j from this inequality to obtain $v_i - v_j \geqq -\Delta q + \sum_{t \neq j} P(\psi^*)_{it}(v_t - v_j)$ for all i. Add to this inequality $\Delta q m_{ij} = \Delta q + \Delta q \sum_{t \neq j} P(\psi^*)_{it} m_{tj}$ to conclude $[v_i - v_j + \Delta q m_{ij}] \geqq \sum_{t \neq j} P(\psi^*)_{it}[v_t - v_j + \Delta q m_{tj}]$, $i \in \Omega \backslash \{j\}$. Do repeated iterations of this inequality, noting that the restriction of $P(\psi^*)$ to $\Omega\backslash\{j\}$ is transient, so its powers will approach zero and $v_i - v_j + \Delta q m_{ij} \geqq 0$, $i \in \Omega\backslash\{j\}$, so $v \in D_1$.

(ii) $v \in D_2$. $Tv = v + \gamma\mathbf{1} \leqq v + q_{\max}\mathbf{1}$, cf. (21).

(iii) $v \in D_3$ follows from the four properties

$$(22) \qquad (Tv)_i \leqq v_s + \Delta q(\theta_i - 1) + q_{\max}, \qquad i \in \Omega\backslash R,$$

$$(23) \qquad (Tv)_j \geqq v_s + q_{\min} - \Delta q(m^* - 1), \qquad j \in R, \quad m^* \geqq 1, \quad \theta_i \geqq 1, \quad i \in \Omega\backslash R$$

where $s \in R$ satisfies $v_s = \max_{t \in R} v_t$, because subtraction of (23) from (22) yields

$$v_i - v_j = Tv_i - Tv_j \leqq \Delta q(\theta_i + m^* - 1) \leqq \Delta q(\theta_i + (m^* - 1)\theta_i) = \Delta q m^* \theta_i,$$

for all $i \in \Omega\backslash R$, $j \in R$, or $v \in D_3$.

Property (22) is established via $v_i = Tv_i - \gamma = q(f^0)_i + \sum_t P(f^0)_{it} v_t - \gamma$ or $v_i - v_s \leqq q_{\max} - q_{\min} + \sum_{t \in \Omega\backslash R} P(f^0)_{it}[v_t - v_s]$, where f^0 is any pure policy achieving all maxima in (9). Subtracting

$$\Delta q \theta_i \geqq \Delta q + \Delta q \sum_{t \in \Omega\backslash R} P(f^0)_{it} \theta_t,$$

we obtain

$$v_i - v_s - \Delta q \theta_i \leqq \sum_{t \in \Omega\backslash R} P(f^0)_{it}[v_t - v_s - \Delta q \theta_t], \qquad i \in \Omega\backslash R.$$

Repeated iteration and use of $[P(f^0)_{it}]_{i,t \in \Omega\backslash R}$ being transient implies $v_t - v_s - \Delta q \theta_t \leqq 0$, $t \in \Omega\backslash R$ and

$$\begin{aligned} Tv_i &\leqq q_{\max} + \max_{k \in K(i)} \left[\sum_{t \in R} P_{it}^k v_t + \sum_{t \in \Omega\backslash R} P_{it}^k v_t \right] \\ &\leqq q_{\max} + v_s + \max_{k \in K(i)} \sum_{t \in \Omega\backslash R} P_{it}^k \Delta q \theta_t, \end{aligned}$$

which yields (22).

To establish (23), we recall R is closed for $P(\psi^*)$:

$$v_l+q_{\max}\geqq v_l+\gamma=(Tv)_l\geqq q(\psi^*)_l+\sum_{t\in R}P(\psi^*)_{lt}v_t, \qquad l\in R,$$

or

$$v_l-v_s\geqq -\Delta q+\sum_{t\in R\backslash\{s\}}P(\psi^*)_{lt}[v_t-v_s], \qquad l\in R.$$

Add $\Delta q m_{ls}=\Delta q+\Delta q\sum_{t\in R\backslash\{s\}}P(\psi^*)_{lt}m_{ts}$ to obtain

$$[v_l-v_s+\Delta q m_{ls}]\geqq \sum_{t\in R\backslash\{s\}}P(\psi^*)_{lt}[v_t-v_s+\Delta q m_{ts}], \qquad l\in R.$$

Repeated iteration of this inequality for $l\in R\backslash\{s\}$ and use of the fact that $[P(\psi^*)_{lt}]_{l,t\in R\backslash\{s\}}$ is transient, yields $v_l-v_s+\Delta q m_{ls}\geqq 0$, $l\in R$. Then, for $j\in R$:

$$\begin{aligned}Tv_j&\geqq q(\psi^*)_j+P(\psi^*)_{js}v_s+\sum_{l\in R\backslash\{s\}}P(\psi^*)_{jl}v_l\\&\geqq q_{\min}+v_s+\sum_{l\in R\backslash\{s\}}P(\psi^*)_{jl}[-\Delta q m_{ls}]\\&=q_{\min}+v_s-\Delta q(m_{js}-1)\geqq q_{\min}+v_s-\Delta q(m^*-1), \quad \text{or (23).}\end{aligned}$$ □

We note that the right-hand sides in the inequalities (15) and (17) defining the polyhedra D_1 and D_3 are specified as multiples of the numbers $\{\theta_i|i\in\Omega\backslash R\}$ and $\{m_{ij}|i\in\Omega, j\in R\}$. The multiples are the tightest possible which work for an arbitrary choice of ψ^*, as is shown by the following example in which $D_1\cap D_2\cap D_3$ is the solution space of (9).

Example 1.

i	k	P_{i1}^k	P_{i2}^k	P_{i3}^k	q_i^k
1	1	1	0	0	0
	2	0	1	0	−1
2	1	0	1	0	0
	2	1	0	0	−1
3	1	1	0	0	−1

$\Omega=\{1,2,3\}$; $K(i)=\{1,2\}$, $i=1,2$; $K(3)=\{1\}$. $R=\{1,2\}$. $m^*=2$; $\Delta q=1$; $\theta_3=1$. Using policy ψ^* with $\psi^*_{i2}=1$, $i=1,2$:

$$D_1\cap D_2\cap D_3=\{v\,|\,|v_1-v_2|\leqq 1;\ |v_3-v_2|\leqq 1;\ v_1-1\leqq v_3\leqq v_1+2\}$$

and $\{(\gamma=0, v)\,|\,|v_1-v_2|\leqq 1;\ v_3=v_1-1\}$ is the solution set to (9).

4. The general multichain model. In this section we show how the above described fixed point approach can be extended to exhibit the existence of a solution to the *pair* of optimality equations (6), (7) that arise in the general multichain case.

Our analysis is based upon a decomposition of the state space Ω. This decomposition procedure bears some resemblance with the one described in Bather [2]. The decomposition is achieved via an iterative procedure. In the lth iteration the state space is reduced to $\Omega_l\subset\Omega_{l-1}(l\geqq 2, \Omega_1=\Omega)$ and for $i\in\Omega_l$ we define

$$(24) \qquad K_l(i)=\left\{k\in K(i)\,\middle|\,\sum_{j\in\Omega_l}P_{ij}^k=1\right\},$$

the subset of $K(i)$ under which the system stays in Ω_l with probability 1. On the restricted state space Ω_l, one determines $\{C_{(l,p)} | p = 1, \cdots, n(l)\}$ the communicating classes in Ω_l under $X_{i\in\Omega_l}K_l(i)$.

Decomposition procedure.

Step 0. (*Initialization*). Set $\Omega_1 = \Omega$; $l = 1$; $K_l(i) = K(i)$, $i \in \Omega$.

Step 1. (*Iterative step*). Determine $\{C_{(l,p)} | p = 1, \cdots, n(l)\}$ the communicating classes in Ω_l under $X_{i\in\Omega_l}K_l(i)$. Let $\Lambda_l = \bigcup_{p=1}^{n(l)} C_{(l,p)}$ and let Δ_l be the possibility empty set of states in $\Omega_l \backslash \Lambda_l$ for which there is a positive probability of leaving $\Omega_l \backslash \Lambda_l$ under every possible alternative. Let $\Omega_{l+1} = (\Omega_l \backslash \Lambda_l) \backslash \Delta_l$. If $\Omega_{l+1} = \emptyset$, stop.

Step 2. For $i \in \Omega_{l+1}$, define $K_{l+1}(i)$ by (24); increment l by one and return to step 1.

The decomposition procedure converges in $l^* \leqq N$ iterations since Λ_l contains at least one state as long as $\Omega_l \neq \emptyset$. We illustrate this decomposition procedure with the following example:

Example 2.

i	k	P_{i1}^k	P_{i2}^k	P_{i3}^k	P_{i4}^k	P_{i5}^k	P_{i6}^k
1	1	1	0	0	0	0	0
2	1	0	0	1	0	0	0
2	2	0	1	0	0	0	0
3	1	0	1	0	0	0	0
4	1	0	0	1	0	0	0
4	2	0	0	0	1	0	0
5	1	0	.5	.5	0	0	0
5	2	0	0	1	0	0	0
6	1	0	0	0	.8	0	.2
6	2	0	0	0	0	0	1

$$n(1) = 2; \quad \Lambda_1 = \{1, 2, 3\}; \quad C_{(1,1)} = \{1\}; \quad C_{(1,2)} = \{2, 3\}; \quad \Delta_1 = \{5\}.$$

$$\Omega_2 = \{4, 6\}; \quad K_2(4) = \{2\}; \quad K_2(6) = \{1, 2\}; \quad n(2) = 1; \quad \Lambda_2 = C_{(2,1)} = \{4\}; \quad \Delta_2 = \emptyset.$$

$$\Omega_3 = \{6\}; \quad K_3(6) = \{2\}; \quad n(3) = 1; \quad \Lambda_3 = C_{(3,1)} = \{6\}; \quad \Delta_3 = \emptyset.$$

For each $l = 1, \cdots, l^*$, $X_{i\in\Omega_l}K_l(i)$ will be referred to as the set of l-level policies.

Note that for each *level* l, and each set $C_{(l,p)}$ $(p = 1, \cdots, n(l))$ a solution $\{\sigma_{l,p}; u_i^{(l,p)} | i \in C_{(l,p)}\}$ exists to the optimality equation:

$$u_i^{(l,p)} = \max_{k\in K_l(i)} \left\{ q_i^k - \sigma_{l,p} + \sum_{j\in C_{(l,p)}} P_{ij}^k u_j^{(l,p)} \right\}, \qquad i \in C_{(l,p)} \tag{25}$$

with $\sigma_{l,p}$ representing the maximal gain rate among *all* l-level policies on $C_{(l,p)}$.

To verify this statement, use the fact that $C_{(l,p)}$ is a *communicating* set with respect to the l-level policies and invoke corollary 1. We next construct a solution g to (6), the components of which are determined by the following recursive scheme. First, for any subset $A \subseteq \Omega$, let $\bar{A} = \Omega \backslash A$.

ALGORITHM 1 (*constructing a solution g to* (6)).

Step 0. Initialize $l = 1$; for each $p = 1, \cdots, n(1)$ set $g_{1,p} = \sigma_{1,p}$ and $g_i = g_{1,p}$ for all $i \in C_{(1,p)}$.

Step 1. Determine the unique solution to the transient optimality equation (26) (cf. Veinott [25]) in $\{x_i, i \in \Delta_l\}$:

$$x_i = \max_{k \in K(i)} \left[\left(\sum_{j \in \bar{\Omega}_l \cup \Lambda_l} P_{ij}^k g_j \right) + \sum_{j \in \Delta_l} P_{ij}^k x_j \right]; \qquad i \in \Delta_l \tag{26}$$

and let $g_i = x_i$ for all $i \in \Delta_l$. If $l = l^*$, exit; otherwise increment l by one and set $p = 1$.

Step 2. Set

$$g_{l,p} = \max \left\{ \sigma_{l,p}; \max_{i \in C_{(l,p)}; k \in K(i) \setminus K_l(i)} \left[\left(1 - \sum_{t \in C_{(l,p)}} P_{it}^k \right)^{-1} \sum_{j \in \Omega_l} P_{ij}^k g_j \right] \right\}, \tag{27}$$

and set $g_i = g_{l,p}$.

Step 3. If $p < n(l)$, increment p by one and go to step 2; otherwise go to step 1.

With g being determined as a solution to (6), the sets $B(i)$, $i \in \Omega$, can now be determined. This leaves us with the construction of a solution v to (7).

ALGORITHM 2 (*constructing a solution to* (7)).

Step 0. Initialize $l = 1$; for each $p = 1, \cdots, n(1)$, fix a solution $u^{(l,p)}$ to (25) and let $v_i = u_i^{(l,p)}$ for all $i \in C_{(l,p)}$ and $p = 1, \cdots, n(1)$.

Step 1. Determine the unique solution to the transient optimality equation (30) in $\{x_i, i \in \Delta_l\}$, cf. Veinott [25]:

$$x_i = \max_{k \in B(i)} \left\{ q_i^k - g_i + \sum_{j \in \bar{\Omega}_l \cup \Lambda_l} P_{ij}^k v_j + \sum_{j \in \Delta_l} P_{ij}^k x_j \right\}, \qquad i \in \Delta_l \tag{28}$$

and let $v_i = x_i$ for all $i \in \Delta_l$. If $l = l^*$, exit; otherwise increment l by one and set $p = 1$.

Step 2. (i) *If* $g_{l,p} = \sigma_{l,p}$: Fix a solution $u^{(l,p)}$ to (25). Define $\bar{u}^{(l,p)}$ on $C_{(l,p)}$ by setting $\bar{u}_i^{(l,p)} = u_i^{(l,p)} + c$ with c sufficiently large that

$$\bar{u}_i^{(l,p)} = \max_{k \in B(i)} \left\{ q_i^k - g_i + \sum_{j \notin \Omega_l} P_{ij}^k v_j + \sum_{j \in C_{(l,p)}} P_{ij}^k \bar{u}_j^{(l,p)} \right\}, \qquad i \in C_{(l,p)} \tag{29}$$

and set $v_i = u_i^{(l,p)}$ for all $i \in C_{(l,p)}$.

(ii) *If* $g_{l,p} > \sigma_{l,p}$: for all $i \in C_{(l,p)}$ set

$$v_i = \max \left\{ \sum_{j \in C_{(l,p)}} [I - \hat{P}(f)]_{ij}^{-1} \left[q(f)_j - g_j + \sum_{t \notin \Omega_l} P(f)_{jt} v_t \right] \right| \; f \in X_j B(j), \hat{P}(f) \text{ is transient on } C_{(l,p)} \right\} \tag{30}$$

where $\hat{P}(f)$ is the square submatrix of $P(f)$ corresponding to the rows and columns in $C_{(l,p)}$.

Step 3. If $p < n(l)$, increment p by one and return to step 2. Otherwise, return to step 1.

THEOREM 3. *The pair of vectors* $\{g, v\}$ *constructed by Algorithms* 1 *and* 2 *are a solution to* (6), (7). *In particular,* $g = g^*$.

Proof. We first verify that all of the steps in the two algorithms are well defined. Note that for $l \geqq 2$, $K(i) \setminus K_l(i) \neq \varnothing$, for some $i \in \Omega_l$. Otherwise $\Omega_l \cap \Lambda_1 \neq \varnothing$ a contradiction. The maxima in (27) are thus well defined. Also, in case (ii) of step 2, observe that the maximum is taken over a nonempty set, for assuming the contrary would imply that some subset of $C \subseteq C_{(l,p)}$ is closed under $X_j B(j)$, hence $B(j) = K_l(j)$ for all $j \in C$, contradicting $g_{l,p} > \sigma_{l,p}$. To verify that the vector $\bar{u}^{(l,p)}$ in step 2, case (i) is well defined, note that $B(i) \supseteq K_l(i)$. Moreover, if $u^{(l,p)}$ satisfies (25), then so does $\bar{u}^{(l,p)}$, defined by $\bar{u}_i^{(l,p)} = u_i^{(l,p)} + c$ for any constant c. It thus suffices to show that

$$\bar{u}_i^{(l,p)} \geqq \left\{ q_i^k - g_i + \sum_{j \notin \Omega_l} P_{ij}^k v_j + \sum_{j \in C_{(l,p)}} P_{ij}^k u_j^{(l,p)} \right\}$$

for all $k \in B(i) \backslash K_l(i)$ and all $i \in C_{(l,p)}$, provided c is chosen sufficiently large. This, however, is immediate from $\sum_{j \in C_{(l,p)}} P_{ij}^k < 1$ for $k \in B(i) \backslash K_l(i)$, $i \in C_{(l,p)}$.

We next show that g satisfies (6). Use induction on l. For $l = 1$ and all $p = 1, \cdots, n(1)$, we have for all $i \in C_{(1,p)}$:

$$g_i = g_{(1,p)} = \max_{k \in K(i)} \sum_j P_{ij}^k g_j = \max_{k \in K(i)} \sum_{j \in C_{(1,p)}} P_{ij}^k g_j.$$

(The last equality is due to $C_{(1,p)}$ being closed under every policy.) If (6) is satisfied for all $i \in \bar{\Omega}_l \cup \Lambda_l$, then it is satisfied for $i \in \Delta_l$ as well, cf. (26). Thus, assume (6) is satisfied for all $i \notin \Omega_l$, $l \geqq 2$, and show it is satisfied for $i \in \Lambda_l$ as well. Fix $1 \leqq p \leqq n(l)$ and $i \in C_{(p,l)}$. For all $k \in K_l(i)$, $g_i = g_{(l,p)} = \sum_{j \in C_{(l,p)}} P_{ij}^k g_j = g_{(l,p)}$. Hence it suffices to show that $g_{(l,p)} \geqq \sum_j P_{ij}^k g_j = \sum_{j \notin \Omega_l} P_{ij}^k g_j + \sum_{j \in C_{(p,l)}} P_{ij}^k g_j$ or $g_{l,p} \geqq (1 - \sum_{t \in C_{(l,p)}} P_{it}^k)^{-1} \sum_{j \in \Omega_l} P_{ij}^k g_j$ for all $k \in K(i) \backslash K_l(i)$ which follows from (27).

We now verify that $\{g, v\}$ satisfy (7). Use induction on l. First consider the case where $l = 1$. Fix $p = 1, \cdots, n(1)$. Note that $\sigma_{1,p} = g_{1,p}$ and $K_1(i) = K(i) = B(i)$ for all $i \in C_{(1,p)}$ with $C_{(1,p)}$ being closed under every policy. Hence (7) is satisfied for all $i \in C_{(1,p)}$. If (7) is satisfied for all $i \in \bar{\Omega}_l \cup \Lambda_l$, then it is satisfied for $i \in \Delta_l$ as well, cf. (28). Next assume (7) is satisfied for all $i \notin \Omega_l$, $l \geqq 2$, and show it is satisfied for $i \in \Lambda_l$ as well. Fix $p = 1, \cdots, n(l)$. In case (i), it is immediate that v_i, $i \in C_{(l,p)}$ satisfy (7), cf. (29). Next assume that case (ii) prevails for $C_{(l,p)}$ and observe that a policy $f \in X_j B(j)$ exists such that for *all* $i \in C_{(l,p)}$:

$$v_i = q(f)_i - g_i + \sum_{j \in C_{(l,p)} \cup \bar{\Omega}_l} P(f^*)_{ij} v_j.$$

Now assume to the contrary that (7) is violated for some $i \in C_{(l,p)}$. This implies the existence of a policy $f \in X_j B(j)$ for which for all $i \in C_{(l,p)}$:

$$v_i \leqq q(f)_i - g_{l,p} + \sum_{j \notin \Omega_l} P(f)_{ij} v_j + \sum_{j \in C_{(l,p)}} P(f)_{ij} v_j \tag{31}$$

with strict inequality applying to *some* $i \in C_{(l,p)}$. Note that $\hat{P}(f)$ cannot be transient on $C_{(l,p)}$ since otherwise for *some* $i \in C_{(l,p)}$:

$$v_i < \sum_{j \in C_{(l,p)}} [I - \hat{P}(f)]_{ij}^{-1} \left[q(f)_j - g_j + \sum_{r \notin \Omega_l} P(f)_{jr} v_r \right]$$

contradicting (30). Hence $P(f)$ has a subchain $C \subseteq C_{(l,p)}$ with $\{\pi(f)_i | i \in C\}$ representing the (unique) stationary probability distribution of $P(f)$ on C. Since C is closed under $P(f)$, f prescribes an action in $K_l(i)$ for all $i \in C$. For $i \in C$, (31) reduces to

$$v_i \leqq q(f)_i - g_{l,p} + \sum_{j \in C} P(f)_{ij} v_j, \qquad i \in C. \tag{32}$$

Multiply (32) by $\pi(f)_i > 0$ and sum over $i \in C$ to conclude

$$g_{l,p} \leqq \sum_{i \in C} \pi(f)_i q(f)_i \leqq \sigma_{l,p}$$

contradicting (ii). □

Acknowledgment. We would like to thank the referees for many useful comments.

REFERENCES

[1] J. BATHER, *Optimal decision procedures for finite Markov chains, part* II, Adv. Appl. Prob., 5 (1973), pp. 521–540.

[2] ———, *Optimal decision procedures for finite Markov chains, part* III, Adv. Appl. Prob., 5 (1973), pp. 541–552.

[3] T. BEWLEY AND E. KOHLBERG, *The asymptotic theory of stochastic games*, Math. Oper. Res., 1 (1976), pp. 197–208.

[4] D. BLACKWELL, *Discrete dynamic programming*, Ann. Math. Stat., 33 (1962), pp. 719–726.

[5] E. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165–177.

[6] ———, *Markov renewal programs with small interest rates*, Ann. Math. Stat., 42 (1971), pp. 477–496.

[7] E. DENARDO AND B. FOX, *Multichain Markov renewal programs*, SIAM J. Appl. Math., 16 (1968), pp. 468–487.

[8] J. DUGUNDJI, *Topology*, 5th ed., Allyn and Bacon, Boston, 1980.

[9] A. FEDERGRÜN AND P. J. SCHWEITZER, *Discounted and undiscounted value iteration in Markov decision processes: a survey*, in Dynamic Programming and its Applications, M. Puterman, ed., Academic Press, New York, 1978.

[10] ———, *Solving Markov decision problems by successive elimination of variables* (in preparation), 1984.

[11] ———, *Lyapunov functions in Markov renewal programming* (in preparation), 1984.

[12] A. FEDERGRÜN, P. J. SCHWEITZER AND H. C. TIJMS, *Contraction mappings underlying undiscounted Markov decision problems*, J. Math. Anal. Appl., 65 (1978), pp. 711–730.

[13] ———, *Denumerable undiscounted semi-Markov decision processes with unbounded rewards*, Math. Oper. Res., 8 (1983), pp. 298–314.

[14] A. HORDIJK AND H. C. TIJMS, *A modified form of the iterative method of dynamic programming*, Ann. Stat., 3 (1975), pp. 203–208.

[15] R. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.

[16] W. JEWELL, *Markov renewal programming*, Oper. Res., 11 (1963), pp. 938–971.

[17] E. KOHLBERG, *Invariant half-lines of nonexpansive piecewise-linear transformations*, Math. Oper. Res., 5 (1980), pp. 366–372.

[18] B. MILLER AND A. VEINOTT, JR., *Discrete dynamic programming with a small interest rate*, Ann. Math. Stat., 40 (1969), pp. 366–370.

[19] M. POLLATSCHEK AND B. AVI-ITZHAK, *Algorithms for stochastic games with geometrical interpretation*, Management Sci., 15 (1969), pp. 399–415.

[20] M. PUTERMAN AND S. BRUMELLE, *On the convergence of policy iteration in stationary dynamic programming*, Math. Oper. Res., 4 (1979), pp. 60–70.

[21] P. J. SCHWEITZER, *Iterative solution of the functional equations for undiscounted Markov renewal programming*, J. Math. Anal. Appl., 34 (1971), pp. 495–501.

[22] P. J. SCHWEITZER AND A. FEDERGRÜN, *Functional equations of undiscounted Markov renewal programming*, Math. Oper. Res., 3 (1978), pp. 308–322.

[23] P. J. SCHWEITZER AND B. GAVISH, *An optimality principle for Markovian decision processes*, J. Math. Anal. Appl., 54 (1976), pp. 173–184.

[24] J. F. SHAPIRO, *Brouwer's fixed point theorem and finite state space Markovian decision theory*, J. Math. Anal. Appl., 49 (1975), pp. 710–712.

[25] A. VEINOTT, JR., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Stat., 40 (1969), pp. 1635–1660.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

010

ON METAPATHS IN METAGRAPHS*

KAREL CULIK†

Abstract. Data graphs without (DG) or with (DGP) predicates are directed graphs with labeled vertices and edges. They reflect data flow algorithms if their inner vertices are interpreted by functions or predicates and their roots are initialized by some values. A metapath is an execution sequence of directed shrubs reflecting functions or predicates together with their argument positions. Acyclic data graphs generalizing terms and conditional terms are investigated. A data graph is functional if for each initialization all metapaths determine the same resultation. A finite acyclic DG is functional iff each inner vertex of its simplification DG* has exactly one shrub, where DG* is a data homomorphic image such that each data homomorphic image DG** of DG* is isomorphic with DG*. A finite acyclic DGP is functional iff any two shrubs of an inner vertex of its simplification are incompatible, i.e., they cannot be obtained by the same execution sequence. In the general case of nonacyclic data graphs the (serial) permit execution rule is formulated.

AMS(MOS) subject classifications. 05C20, 05C99, 68C05

1. Introduction. A fully parenthesized arithmetic expression $E = (((X-Y)/(Y-X))+(Z*(X-Y)))$ is usually represented in Fig. 1.1a as a *directed acyclic graph* $\langle V, \delta\rangle$, where $V=\{1,2,3,\cdots,8,9\}$ is the set of its *vertices*, and $\delta = \{(1,4), (1,5), (1,6), (2,4),\cdots,(7,9), (8,9)\}$ is the set of its *data edges*, with two *vertex labelings*: $\text{ar}: V \rightarrow N$ where $N=\{0,1,2,\cdots\}$, is called the *arity*, and $\text{nam}: V \rightarrow \text{FunctNam} \cup \text{Var}$ where FunctNam, Var is the set of *function names, variables*, respectively, is called the *naming*. For example, ar (4) = 2, nam (4) = –, nam (1) = X, etc.

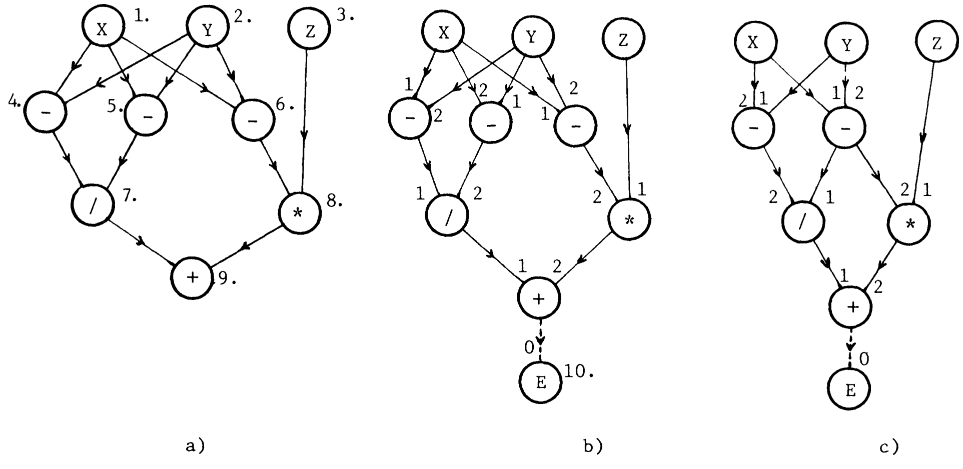


FIG. 1.1

Having Fig. 1.1a one cannot reconstruct the original expression E because one does not know whether $X-Y$ or $Y-X$ should be associated with the vertices 4, 5 and 6. To save the reconstructability an *edge labeling* $\Delta: \delta \rightarrow N$, called the *argument position*, can be added, as shown in Fig. 1.1b, where a new vertex 10 is added

* Received by the editors June 9, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Computer Science, Wayne State University, Detroit, Michigan 48202.

corresponding to the variable E (serving as a name of the whole expression, or of the function represented by it, as we want to avoid usual functional notation). For example, $\Delta(1,4)=\Delta(2,5)=1$, $\Delta(2,4)=\Delta(1,5)=2$, etc.

If $\operatorname{idg}(v)$, $\operatorname{odg}(v)$ denote the *indegree*, *outdegree* of v in $\langle V, \delta\rangle$, let $R=\{v\in V; \operatorname{idg}(v)=0\}$, $L=\{v\in V; \operatorname{odg}(v)=0\}$ denote the sets of all *roots*, *leaves* of $\langle V, \delta\rangle$, respectively. Thus $R=\{1,2,3\}$ and $L=\{10\}$ in Fig. 1.1b while $L=\{9\}$ in Fig. 1.1a.

A graph structure $\mathrm{DG}=\langle V, \delta, \mathrm{ar}, \mathrm{nam}, \Delta\rangle$ is called a *data graph* (without decisions) if the following natural requirements are satisfied:

(1.1) (i) $R\neq 0\neq L, R\cap L=0$;
(ii) $(v,v)\notin\delta$ for each $v\in V$;
(iii) there are no isolated vertices.

(1.2) (i) $v, w\in R,\ v\neq w\Rightarrow \operatorname{nam}(v)\neq\operatorname{nam}(w)$;
(ii) $0\leqq \operatorname{ar}(v)\leqq \operatorname{idg}(v)$ for each $v\in V$;
(iii) $\operatorname{nam}(v)=\operatorname{nam}(w)\Rightarrow \operatorname{ar}(v)=\operatorname{ar}(w)$ for all $v, w\in V$;
(iv) $\operatorname{ar}(v)=0$ for each $v\in R$ and $\operatorname{ar}(v)>0$ for each $v\in V-R$.

(1.3) (i) $\operatorname{ar}(v)>0$ and $(w,v)\in\delta\Rightarrow 1\leqq\Delta(w,v)\leqq\operatorname{ar}(v)$;
(ii) $1\leqq i\leqq \operatorname{ar}(v)\Rightarrow$there exist $w\in V$ and $(w,v)\in\delta$ such that $\Delta(w,v)=i$.

The names of all roots will be viewed as variables while the names of all other vertices will be viewed as function names (or as predicate names later). From each DG one can get an *extended data graph* by adding one new leaf w to each old one v, the edge (v,w) with $\Delta(v,w)=0$, and with $\operatorname{nam}(w)$ being viewed as a variable, as it is shown in Fig. 1.1b which is an extension of Fig. 1.1a.

Each fully parenthesized arithmetic expression, and, in general, each *term* defined with respect to FunctNam and Var is used as an *algorithm* when it is *evaluated* (*executed*) assuming its function names are *interpreted* by an interpretation int: FunctNam → Funct, where Funct is a set of usual functions (int preserves the arity), and its variables are initialized by an *initialization* Init: Var → Val where Val is a set of values.

The well-known (serial) *execution rule* is as follows: take the function name, f, inside an innermost pair of parentheses, apply the function $\operatorname{int}(f)$ to the argument values provided either by the initialization Init of variables, or by results of previous applications, and then replace the whole pair of parentheses by the result value of application of $\operatorname{int}(f)$; one continues doing so until one single value is left (assuming all functions concerned are always defined when needed).

If the execution rule above is applied to the graph representation in Fig. 1.1a of the expression E, one can get the following *execution sequence*: $(4,5,7,6,8,9)$ which is *not a path* according to graph theory, because neither $(4,5)$ nor $(7,6)$ are edges of Fig. 1.1a. Nevertheless the following weaker condition concerning the execution sequence $(4,5,7,6,8,9)$ is satisfied:

(1.4) for each $i>|R|$ there exist two vertices j, k such that $1\leqq j<k<i$ and $(j,i)\in\delta$ and $(k,i)\in\delta$, or, eventually, $j\in R$ or $k\in R$.

In *metamathematics* (proof theory, logic) a very similar condition is used to define *formal proof from axioms* [Grze 74], or from assumptions [Klee 67], $F_1, F_2, \cdots, F_n$ as a finite sequence of *formulae* $(F_{n+1}, F_{n+2}, \cdots, F_{n+p})$ such that for each i, $n+1\leqq i\leqq n+p$, there exist two formulae F_j, F_k, $1\leqq j<k<i$, and F_i is obtained by the *rule of detachment* (modus ponens) from them (e.g. F_k has the form $(F_j\Rightarrow F_i)$ where $\Rightarrow$ is the implication).

Therefore the needed generalized concept of a path can be called a *metapath.* In both cases binary functions (either arithmetic or of detachment) are concerned. They are ternary relations, and therefore usual graph theoretical concepts do not suffice. The usual path is a very special case of a metapath when only unary functions are concerned.

In general we want to represent an arbitrary m-ary function (operation) $f^{(m)}$, where $m \geqq 1$, and, later, also an arbitrary m-ary predicate (relation) $p^{(m)}$, as a whole, as one unit. It can be done using directed trees as in Fig. 1.2a, called (directed) *shrubs* in [Koni 36].

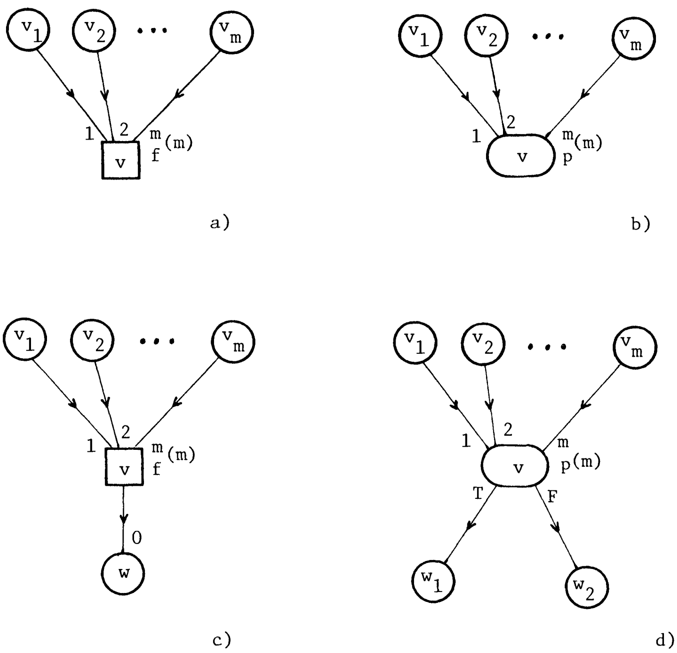


FIG. 1.2

The shrub in Fig. 1.2a (or 1.2b) is called a *v-shrub* and is denoted and defined by $\mathrm{SH}_v = \{(v_1, v), (v_2, v), \cdots, (v_m, v)\}$, where exactly m edges are included (as $\mathrm{ar}\,(v) = m$, and any two of them are labeled by different integers in Δ). Their common vertex, v, is called the *leaf of* SH_v, while all remaining vertices of these edges are called *roots of* SH_v. In particular v_i is called the ith *root of* SH_v, $1 \leqq i \leqq m$, if $\Delta(v_i, v) = i$.

Using the concept of shrub one can amplify and simplify the concept of metapath as follows. The existential requirement in (1.4) will be replaced by the presentation of the shrubs themselves. The original metapath $(4, 5, 7, 6, 8, 9)$ will be replaced by the following sequence of shrubs: $\mathrm{me} = (\mathrm{SH}_4, \mathrm{SH}_5, \mathrm{SH}_7, \mathrm{SH}_6, \mathrm{SH}_8, \mathrm{SH}_9)$, where $\mathrm{SH}_4 = \{(1, 4), (2, 4)\}$, $\mathrm{SH}_5 = \{(1, 5), (2, 5)\}$, $\mathrm{SH}_7 = \{(4, 7), (5, 7)\}$, $\mathrm{SH}_6 = \{(1, 6), (2, 6)\}$, $\mathrm{SH}_8 = \{(6, 8), (3, 8)\}$, $\mathrm{SH}_9 = \{(7, 9), (8, 9)\}$ satisfying

(1.5) if $i \notin R = \{1, 2, 3\}$ then each root of SH_i is either a leaf of SH_j for some j, $1 \leqq j < i$, or it belongs to R itself.

Fig. 1.2c is an extension of Fig. 1.2a where v is a function vertex. If v is a predicate vertex as in Fig. 1.2b, then the corresponding extension must have two different leaves as shown in Fig. 1.2d, and the two edges terminating in them are labeled by two different *truth values*, T and F. In fact the *extended shrubs* in Fig. 1.2c and d are only building blocks of all data graphs. Obviously the new leaves of extended shrubs will be named by variables only.

The main construction which will be needed concerning extended shrubs and extended data graphs is *identification*, usually identification of a root of one shrub with a leaf of the second one (assuming the two graphs are vertex disjoint).

One can say that Fig. 1.1b is *decomposed* into mutually (vertex) disjoint shrubs SH_4^*, SH_5^*, SH_6^*, SH_7^*, SH_8^*, SH_9^*, SH_{10}^* represented in Fig. 1.3a, because each SH_i^* is isomorphic with SH_i, and Fig. 1.1b is obtained from Fig. 1.3a by identification of pairs of vertices inside the dotted lines. For example, by identification of the leaf of SH_4^* with the 1st root of SH_7^* one obtains a data graph represented in Fig. 1.3b. Its leaf can be identified with the 1st root of SH_9^*, etc.

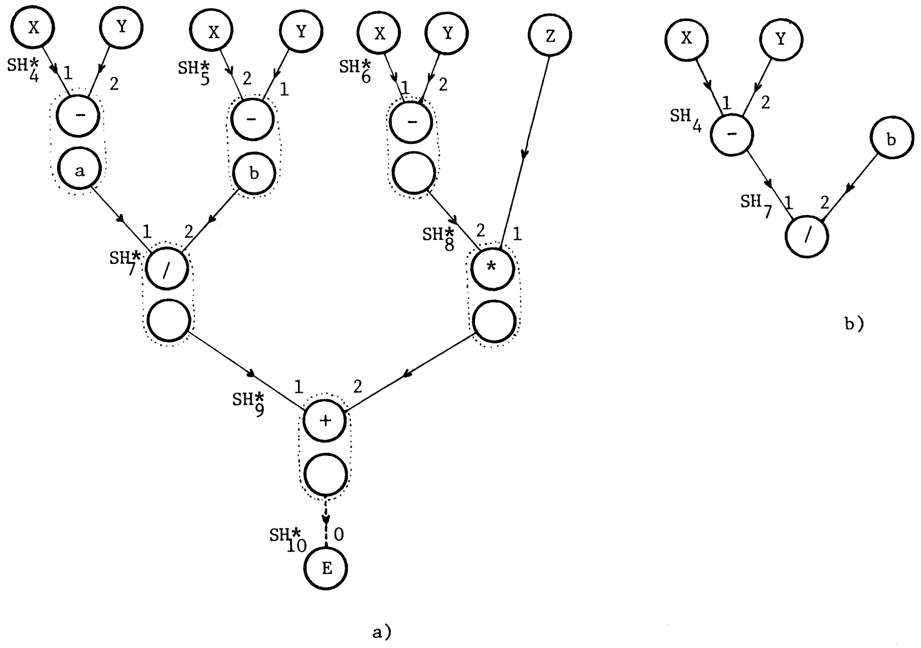


FIG. 1.3

A graph viewed as a result of identification (of vertices or edges) of some simpler graphs is called a *metagraph relative to a given set of graphs*. Thus Fig. 1.1b is a metagraph relative to the set of shrubs SH_i^* from Fig. 1.3a.

The identification of the first root of SH_7^* in Fig. 1.3a with the leaf of SH_4^* corresponds to the *substitution* of $(X-Y)$ for a in (a/b), but the graph representation allows us to omit variables while preserving the underlying relationship and making it more general. An example is Fig. 1.1c, which is a homomorphic image of Fig. 1.1b, and allows us not to repeat the same subexpression $X-Y$ twice. This cannot be expressed using usual terms. Fig. 1.1c can also be executed and the corresponding execution sequence is shorter than before as the subexpression $X-Y$ will be evaluated only once.

A *composition of two vertex disjoint graphs* or multigraphs (either directed or undirected) by identification of their vertices or edges can be defined in several different ways, but, in all generality, it cannot be viewed as a binary operation, because in addition for two given graphs it must also be prescribed which vertices or edges should be identified. Therefore we are saying that the identification of vertices of two graphs is a *construction* but not necessarily an *operation* (accepting the term "construction" being superior to the term of "operation" (function)).

Section 2 presents basic definitions of rooted and closed metapaths in a data graph $\mathrm{DG} = \langle V, \delta, \mathrm{ar}, \mathrm{nam}, \Delta\rangle$ defined by (1.1)–(1.3). The concepts of a metapath and usual path are compared, and some basic properties of data graphs are studied.

Section 3 contains the definition of data homomorphism and studies some of its properties which will be needed further.

Section 4 concerns acyclic finite data graphs viewed as permit algorithms with a (serial) execution rule formulated independently of the concept of term. If they are interpreted (and initialized in all admissible ways) they compute (evaluate, define) *m/n-ary functions* when $m = |R|$ and $m = |L|$, which are systems of n m-ary functions with identical domains, or so called *functions of general form* [Mann 74].

A (serial) *permit execution rule*, which generalizes that one for terms, is introduced; the *functionality* of data graph is defined by the requirement, all execution sequences (admissible by the execution rule) determine the same result for the same initialization (and given interpretation).

The sequence $(6, 8, 5, 4, 7, 9)$ is also an execution sequence of Fig. 1.1a, and computes the same result as the previous one, because each term, when evaluated in all possible ways, determines a function. This is not the case in general, but the functionality of finite and acyclic data graphs (without predicates) is fully characterized using the concept of data homomorphism.

Section 5 concerns acyclic finite data graphs with predicates using the same (serial) execution rule as in § 4. If they are interpreted (and initialized in all admissible ways), they may compute (evaluate, define) *decisions* which are systems of $k \geqq 2$ m/n-ary functions, where $1 \leqq i \leqq k$, with mutually disjoint domains. Their union can (but need not) be a m/n-ary function if $n_i = n$ for $i = 1, 2, \cdots, k$. The functionality of these more general permit algorithms is fully characterized.

Section 6 concerns nonacyclic finite data graphs with or without predicates using almost the same (serial) execution rule as before. The functionality is defined but not fully characterized, although it is the most interesting and general case.

Section 7 presents some applications of the theory of data graphs to computer science, in particular, to the design of data flow machines and languages, to functional/applicative languages, and to the theory of computation in general.

2. Metapaths of data graphs. The defining requirements (1.1)–(1.3) of a data graph are not very restrictive, as is shown in the following:

LEMMA 2.1. *Each directed graph $\langle V, \delta\rangle$ which satisfies* (1.1) *can be provided with the following definitions of labelings* ar, nam *and* Δ *such that* $\mathrm{DG} = \langle V, \delta, \mathrm{ar}, \mathrm{nam}, \Delta\rangle$ *is a data graph*:

a) *if* $v \in R$ *let* $\mathrm{ar}(v) = 0$ *and if* $n \in V - R$ *let* $\mathrm{ar}(v) = \mathrm{idg}(v)$ *for each* $v \in V$;

b) $\mathrm{nam}(v) = v$ *for each* $v \in V$;

c) $\Delta(w, v)$ *is chosen arbitrarily such that* $1 \leqq \Delta(w, v) \leqq \mathrm{ar}(v)$, *and* $\Delta(w, v) \neq \Delta(u, v)$ *whenever* $w \neq v$ *for all* $v, w, u \in V$.

Proof. By a) follows (1.2)(ii) and (iv), by b) follows (1.2)(i) and (iii), and (1.3) is implied by c).

A sequence, finite or infinite, $\mathrm{me} = (\mathrm{SH}_{v_1}, \mathrm{SH}_{v_2}, \cdots, \mathrm{SH}_{v_n})$ of the v_i-shrubs of a data graph DG, with roots in R and leaves in L, is called a *rooted metapath in* DG, if

(2.1) (i) for each i, $1 \leqq i \leqq n$, each root v of SH_{v_i} satisfies either $v \in R$, or there exists SH_{v_j}, $1 \leqq j < i$, such that $v = v_j$.

It is called a *rooted and leafed metapath in* DG if, in addition,

(2.1) (ii) for each i, $1 \leqq i \leqq n$, either there exists SH_{v_j} where $i < j \leqq n$ such that v_i is a root of SH_{v_j}, or $v_i \in L$.

With each rooted and leafed metapath $\mathrm{me} = (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_n})$ in DG are associated uniquely its *set* $L(\mathrm{me})$, and its *multiset* ML (me) of leaves $v_i \in L$. Obviously $|L(\mathrm{me})| \geqq 1$ and $|\mathrm{ML}\,(\mathrm{me})| \geqq 1$.

A binary relation δ and a data graph $\mathrm{DG} = \langle V, \delta, \mathrm{ar}, \mathrm{nam}, \Delta \rangle$ are called *acyclic* if

(2.2) $(v_1, v_2) \in \delta, (v_2, v_3) \in \delta, \cdots, (v_{k-1}, v_k) \in \delta$ where $k \geqq 2 \Rightarrow (v_k, v_1) \notin \delta$.

LEMMA 2.2. *If* DG *is finite and acyclic, then each shrub* SH_{v_0} of DG *belongs to a rooted and leafed metapath* me *in* DG *such that* $|\mathrm{ML}\,(\mathrm{me})| = 1$. *If* DG *is either infinite or not acyclic, then* SH_{v_0} *need not belong to any rooted path in* DG.

Proof. If each root w of SH_{v_0} satisfies $w \in R$, then $\mathrm{me} = (\mathrm{SH}_{v_0})$ is a rooted metapath in DG according to (2.1)(i), otherwise there exists a root v_{-1} of SH_{v_0} such that $v_{-1} \notin R$. Then according to (1.2)(iv) $\mathrm{ar}\,(v_{-1}) > 0$, and according to (1.3) there exists a $\mathrm{SH}_{v_{-1}}$, and we constructed $(\mathrm{SH}_{v_{-1}}, \mathrm{SH}_{v_0})$. If we have constructed $(\mathrm{SH}_{v_{-i+1}}, \mathrm{SH}_{v_{-i+2}}, \cdots, \mathrm{SH}_{v_{-1}}, \mathrm{SH}_{v_0})$ where $i \geqq 1$, then either it is a rooted metapath in DG, or there exists a root v_{-i} of a $\mathrm{SH}_{v_{-j}}$, $i > j \geqq 0$, such that $v_{-i} \notin R$, and as before we can prolong the assumed sequence to a longer one $(\mathrm{SH}_{v_{-i}}, \mathrm{SH}_{v_{-i+1}}, \cdots, \mathrm{SH}_{v_0})$. In virtue of acyclicity $v_{-i} \neq v_{-j}$ for each $j = 0, -1, \cdots, -i+1$ for each $i = 1, 2, \cdots$, so it would be infinite, but in virtue of finiteness this construction must terminate with a rooted metapath $\mathrm{me} = (\mathrm{SH}_{v_{-i}}, \mathrm{SH}_{v_{-i+1}}, \cdots, \mathrm{SH}_{v_0})$.

Now if $v_0 \in L$, then me is also a leafed metapath, with $|\mathrm{ML}\,(\mathrm{me})| = 1$, because all other leaves w of a $\mathrm{SH}_{v_{-j}}$ satisfy (2.1)(ii). If $v_0 \notin L$, then $\mathrm{odg}\,(v_0) > 0$, and therefore there exists a $v_1 \in V$ such that $(v_0, v_1) \in \delta$. By (1.3) there exists SH_{v_1} such that $(v_0, v_1) \in \mathrm{SH}_{v_1}$, and either $(\mathrm{SH}_{v_{-i}}, \cdots, \mathrm{SH}_{v_{-1}}, \mathrm{SH}_{v_0}, \mathrm{SH}_{v_1})$ is a rooted metapath in DG, or there is a root w of SH_{v_1} such that $w \notin R$. Then we can apply the previous construction to SH_{v_1} to get another rooted metapath $(\mathrm{SH}_{w_{-j}}, \cdots, \mathrm{SH}_{w_0}, \mathrm{SH}_{v_1})$. After merging them into $(\mathrm{SH}_{v_{-i}}, \cdots, \mathrm{SH}_{v_0}, \mathrm{SH}_{w_{-j}}, \cdots, \mathrm{SH}_{w_0}, \mathrm{SH}_{v_1})$, we check either $v_1 \in L$. If yes a rooted and leafed metapath is found, and if not $(v_1 \notin L)$ then $\mathrm{odg}\,(v_1) > 0$, and we can repeat the construction. Again from the acyclicity and finiteness it follows that this construction of prolongation of a rooted metapath must terminate with finding a rooted and leafed metapath me with $|\mathrm{ML}\,(\mathrm{me})| = 1$ containing SH_v, which is what we wanted to prove.

Figure 2.1a is an example of an infinite DG in which $\mathrm{SH}_{v_0} = \{(v_{-1}, v_0)\}$ and does not belong to any rooted metapath, and Fig. 2.1b represents a finite but not acyclic data graph in which $\mathrm{SH}_{w_0} = \{(w_1, w_0), (w_2, w_0)\}$ and does not belong to any rooted metapath.

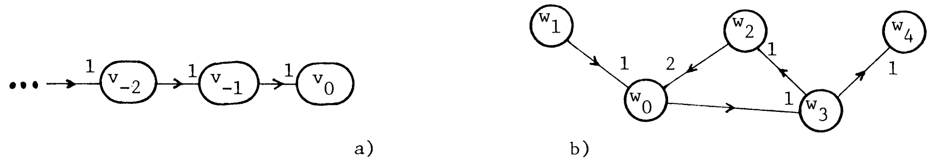


FIG. 2.1

A rooted and leafed metapath $\mathrm{me} = (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_i}, \cdots, \mathrm{SH}_{v_n})$ in DG is called *thin* if

(2.3) either $n = 1$, or the sequence $(\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_{i-1}}, \mathrm{SH}_{v_{i+1}}, \cdots, \mathrm{SH}_{v_n})$ obtained from me by deleting SH_{v_i} is not a rooted and leafed metapath in DG for each $i = 1, 2, \cdots, n$.

LEMMA 2.3. *If* me $= (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_n})$ *is a rooted and leafed metapath in a* DG, *then*

a) *there exist indices* $1 \leqq i_1 < i_2 < \cdots < i_m \leqq n$ *such that* $(\mathrm{SH}_{v_{i_1}}, \mathrm{SH}_{v_{i_2}}, \cdots, \mathrm{SH}_{v_{i_m}})$ *is a thin rooted and leafed metapath in* DG, *and*

b) $(\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_i}, \mathrm{SH}_{v_i}, \cdots, \mathrm{SH}_{v_n})$ *is a rooted and leafed metapath in* DG *for each* $i = 1, 2, \cdots, n$.

The proof follows from the definitions (2.1) and (2.3) immediately.

A finite sequence me $= (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_n})$, $n \geqq 2$, of v_i-shrubs in DG is called a *closed p-metapath in* DG, where $1 \leqq p \leqq n-1$, if

(2.4) (i) for each i, $1 \leqq i \leqq n$, each root w of SH_{v_i} satisfies either $w \in R$, or $w = v_j$ where $i - p \leqq j \leqq i - 1 \pmod n$;

(ii) for each i, $1 \leqq i \leqq n$, there is a root w of SH_{v_i} such that $w \in R$.

Let L (me), ML (me) be the set, multiset of all leaves $v_i \in L$ when SH_{v_i} belongs to me. It may be $|\mathrm{L}(\mathrm{me})| = |\mathrm{ML}(\mathrm{me})| = 0$.

If $p < n-1$, then each closed p-metapath is also a $(p+1)$-metapath, and a *closed metapath* means a closed $(n-1)$-path when $n \geqq 2$ is its length.

LEMMA 2.4. *If a data graph* DG *is not acyclic, then there exists a closed usual path* pa $= (v_1, v_2, \cdots, v_k)$, $k \geqq 2$, *in* DG, *but on the other hand there need not exist any closed p-metapath,* $1 \leqq p < k$, me, *which contains* pa (*in a* DG).

Proof. The first part is an immediate consequence of the definitions (1.1) and (2.2), while the second part follows by the example in Fig. 2.2a for $k = 2$, and in b) for a general k, where $\{v_1, v_2, \cdots, v_k\}$ induces a complete directed and irreflexive graph, and there are k paths (t_i, u_i, v_i), mutually disjoint, and one path (v_k, w). Further $\Delta(v_i, v_j) \leqq k-1$ and $\Delta(u_i, v_i) = k$ for $i = 1, 2, \cdots, k$.

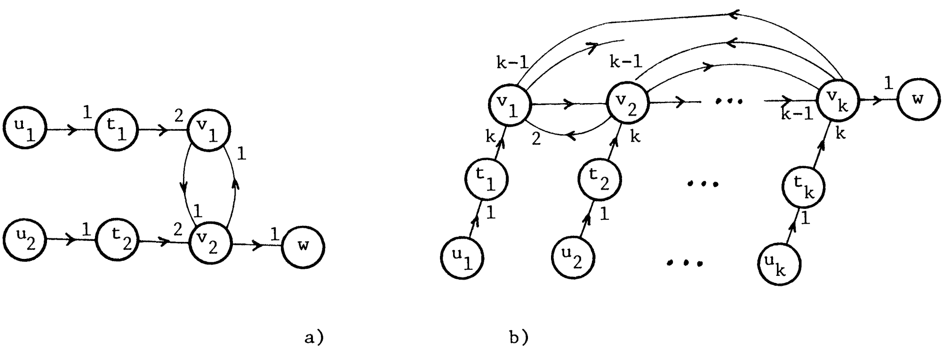


FIG. 2.2

If DG is a data graph and $v \in V - R$, then let sDG_v be the subgraph of DG which is induced by the subsets of vertices and edges belonging to any usual path pa $= (v_1, v_2, \cdots, v_n)$ in DG such that $v_1 \in R$ and $v_n = v$. Obviously v is the only leaf of sDG_v, sDG_v is connected, and sDG_v is a data graph itself if all labelings are considered.

Similarly, if me $= (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_n})$ is a rooted metapath in a DG, then let $\mathrm{sDG}_{\mathrm{me}}$ be that subgraph of DG which is induced by the following set of edges: $\bigcup_{i=1}^{n} \mathrm{SH}_{v_i}$.

LEMMA 2.5. *If* DG *is a data graph which represents a fully parenthesized arithmetic expression, or, in general, a term defined relatively to* FunctNam *and* Var, *then* $|L| = 1$, DG *is connected, finite and acyclic, the only vertices* v *which satisfy* odg $(v) > 1$ *satisfy* $v \in R$, *and it satisfies*

(2.5) $v \in V - R \Rightarrow$ *there exists exactly one v-shrub in* DG.

The proof follows from the fact that such a DG can be obtained from a directed tree with one single leaf by identifying some of its roots.

3. Data homomorphism. A (*data*) *homomorphism* of a $\mathrm{DG} = \{V, \delta, \mathrm{ar}, \mathrm{nam}, \Delta\}$ onto a $\mathrm{DG}^* = \langle V^*, \delta^*, \mathrm{ar}^*, \mathrm{nam}^*, \Delta^*\rangle$ is a mapping $\varphi : V \to V^*$ such that

(3.1) (i) $(w, v) \in \delta \Rightarrow (\varphi(w), \varphi(v)) \in \delta^*$;

(ii) $(w^*, v^*) \in \delta^* \Rightarrow$ there exist $w, v \in V$ such that $(w, v) \in \delta$ and $\varphi(w) = w^*$, $\varphi(v) = v^*$;

(iii) $\mathrm{odg}\,(v) = 0 \Rightarrow \mathrm{odg}^*\,(\varphi(v)) = 0$ for each $v \in V$;

(iv) v and w are not connected in $\mathrm{DG} \Rightarrow \varphi(v)$ and $\varphi(w)$ are not connected in DG*.

(3.2) All three labelings are preserved, that is,

(i) $\mathrm{ar}\,(v) = \mathrm{ar}^*\,(\varphi(v))$ for each $v \in V$;

(ii) $\mathrm{nam}\,(v) = \mathrm{nam}^*\,(\varphi(v))$ for each $v \in V$;

(iii) $\Delta(w, v) = \Delta^*(\varphi(w), \varphi(v))$ for each $(w, v) \in \delta$.

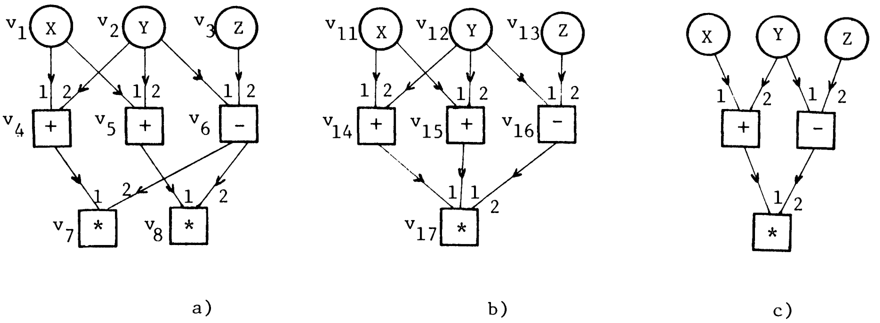
LEMMA 3.1. *If φ is a homomorphism mapping* DG *onto* DG*, *then*

a) $|R| = |R^*|$, $|L| \geqq |L^*|$ *and the inequality may occur*;

and further

b) *if* $\mathrm{SH}_v = \{(v_1, v), \cdots, (v_k, v)\}$, $k \geqq 1$ *is a v-shrub in* DG *then* $\{(\varphi(v_1), \varphi(v)), \cdots, (\varphi(v_k), \varphi(v))\}$ *is a $\varphi(v)$-shrub* $\mathrm{SH}^*_{\varphi(v)}$ *in* DG* *isomorphic with* SH_v.

Proof. a) According to (1.2) different roots have different names and the arity 0. By (3.2)(i) roots must be mapped on roots again, and by (3.2)(ii) $|R| = |R^*|$ follows directly. According to (3.1)(iii) leaves must be mapped onto leaves, and therefore $|L| \geqq |L^*|$ follows, and the inequality may occur as it is shown in Fig. 3.1. The Fig. 3.1a can be homomorphically mapped onto Fig. 3.1b and, obviously, the leaves v_7 and v_8 must be mapped on the same leaf v_{17}.



a) b) c)

FIG. 3.1

b) Follows from (3.2) immediately.

A DG is called *simple* (similarly as in group theory) if each (data) homomorphic image DG* of DG is data isomorphic with DG. A DG* which is a simple and data homomorphic image of DG is called a *simplification of* DG. Figure 3.1c is a simplification of both Fig. 3.1a and Fig. 3.1b.

THEOREM 3.2. *A finite and acyclic* DG *is simple iff there do not exist two different vertices v_1 and v_2 such that* sDG_{v_1} *and* sDG_{v_2} *are data isomorphic.*

Proof. Let DG $=\langle V, \delta, \text{ar}, \text{nam}, \Delta\rangle$ and let there exist $v_1, v_2 \in V-R$, $v_1 \neq v_2$, such that there exists a data isomorphism $\varphi: V_1 \to V_2$ when $\text{sDG}_{v_1}=\langle V_1, \delta_1\rangle$ and $\text{sDG}_{v_2}=\langle V_2, \delta_2\rangle$. Let us define $\langle V^*, \delta^*\rangle$ as follows: $V^*=V-(V_1-V_2)$ and $\delta^*=(\delta-(\delta_1-\delta_2)\cup\{(v_2, v);$ there is a $(v_1, v)\in\delta\}$, and let $\Delta^*(v_2, v)=\Delta(v_1, v)$ for all new edges (v_2, v), while all labelings of original vertices and edges are preserved.

Let φ^* be a trivial extension of φ to V; that is, $\varphi^*(v)=v$ for each $v\in V-V_1$ (and $\varphi^*(v)=\varphi(v)$ for $v\in V_1$). Then φ^* maps V onto V^*, and is a data homomorphism of DG onto DG* because (3.1) and (3.2) are satisfied. As $\varphi^*(v_1)=\varphi^*(v_2)=v_2$, DG is not simple.

Now let us assume DG is not simple. Then there exist a DG* and a data homomorphism $\psi: V\to V^*$ of DG onto DG* which is not a data isomorphism, such that there exist v_1, $v_2\in V$, $v_1\neq v_2$, and $\psi(v_1)=\psi(v_2)=v^*\in V^*$. Now $\text{sDG}^*_{v^*}$ is a data homomorphic image of sDG_{v_1}, and also of sDG_{v_2} under the homomorphism $\psi_1=\psi|_{V_1}$ and $\psi_2=\psi|_{V_2}$, respectively, assuming $\text{sDG}_{v_i}=\langle V_i, \delta_i\rangle$, $i=1, 2$.

Finally, in virtue of acyclicity of DG and DG*, the vertex v^* can be chosen in such a way that for each $w^*\in V^*$, $w^*\neq v^*$, which belongs to a path from an $u^*\in R^*$ to v^* there exists exactly one $w\in V$ such that $\psi(w)=w^*$. It means that ψ_1 and ψ_2 are isomorphisms, from which it follows that sDG_{v_1} and sDG_{v_2} are isomorphic.

LEMMA 3.3. *If φ is a data homomorphism of a* DG *onto a* DG* *and* $\text{me}=(\text{SH}_{v_1},\cdots,\text{SH}_{v_n})$ *is a rooted metapath in* DG *where* $\text{SH}_{v_i}=\{(w_1, v_1),\cdots,(w_k, v_i)\}$ *and* w_j *is the jth root of* SH_{v_i}, *then*

a) $(\varphi(\text{SH}_{v_1}),\cdots,\varphi(\text{SH}_{v_n}))$ *is a rooted metapath* me* *in* DG*, *which can be denoted by* $\varphi(\text{me})=\text{me}^*$; *and*

b) $\text{sDG}^*_{\text{me}^*}$ *is a data homomorphic image of* sDG_{me}.

Proof. a) According to Lemma 3.1b $\varphi(\text{SH}_{v_i})$ is a $\varphi(v_i)$-shrub in DG* for $i=1, 2,\cdots, n$, and it remains to show that (2.1)(i) is satisfied. Considering an i, $1\leqq i\leqq n$, and a root v^* of $\varphi(\text{SH}_{v_i})$ such that $v^*\notin R^*$ we want to show that there exists $\varphi(\text{SH}_{v_j})$, $1\leqq j<i$, and $v^*=\varphi(v_j)$. From (3.2)(i) it follows that each $v\in V$ such that $\varphi(v)=v^*$ must satisfy $v\notin R$, and therefore by (2.1)(i) (being satisfied by me) the existence follows of SH_{v_h}, $1\leqq h<i$, such that $v=v_h$, and therefore one can choose $j=h$.

b) If $\text{sDG}_{\text{me}}=\langle V_1, \delta_1\rangle$ and $\text{sDG}^*_{\text{me}^*}=\langle V_2, \delta_2\rangle$, then $\varphi|_{V_1}$ is the data homomorphism required, because from the definition of these subgraphs it follows that $\delta_1=\bigcup_{i=1}^n \text{SH}_{v_i}$ and $\delta_2=\bigcup_{i=1}^n \varphi(\text{SH}_{v_i})$ and that $\varphi|_{V_1}$ maps V_1 onto V_2, and satisfies (3.1)(i) and (ii). Both sDG_{me} and $\text{sDG}^*_{\text{me}^*}$ are connected, and all remaining requirements are satisfied as φ is a data homomorphism.

Observation 3.4. Fig. 3.2 shows that the labeling nam of a DG is an essential component of its definition (1.1)–(1.3), and also the requirement (3.2)(ii) is essential for the data homomorphism. Figure 3.2a is simple but Fig. 3.2e is not, although they

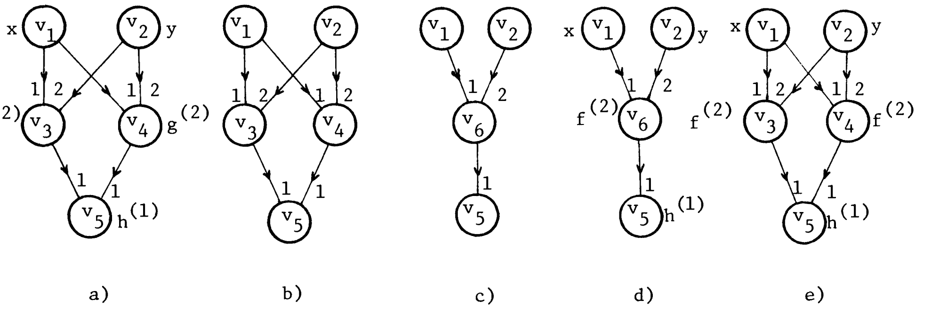


FIG. 3.2

differ only in their labelings of vertices. After abstracting from naming one gets Fig. 3.2b from either a or e which is not simple with respect to a weaker homomorphism. Figure 3.2c is the corresponding simplification of Fig. 3.2b, and Fig. 3.2d is a simplification of Fig. 3.2e.

4. Acyclic data graphs without predicates. The (*serial*) *permit execution rule* for a data graph (without predicates) DG, which is interpreted by int and initialized by Init, determines (1) an *execution sequence* of the form of a rooted and leafed metapath me $= (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_n})$, and (2) the corresponding *computation* of the form of a *value sequence* $\mathrm{val}_{\mathrm{me}} = (\mathrm{val}\,(v_1), \cdots, \mathrm{val}\,(v_n))$ (assuming all functions are always defined when needed) as follows: if Perm (E_i) is a set of *permitting edges*, and Ready (V_i) is the set of vertices (and the corresponding functions) which are ready to be executed (applied), then

(4.1) (i) at the beginning $\mathrm{Perm}\,(E_1) = \{(v, w) \in \delta;\ v \in R\}$;
(ii) $\mathrm{Ready}\,(V_1) = \{v \in V$; there exists a v-shrub such that $\mathrm{SH}_v \subseteq \mathrm{Perm}\,(E_1)\}$;
(iii) one selects $v_1 \in \mathrm{Ready}\,(V_1)$ and one its SH_{v_1} such that $\mathrm{SH}_{v_1} \subseteq \mathrm{Perm}\,(E_1)$, then one applies int (nam (v_1)) to (Init (nam (w_1)), $\cdots$, Init (nam (w_k))) when $\mathrm{SH}_{v_1} = \{(w_1, v_1), \cdots, (w_k, v_1)\}$, and one denotes the result value of the application by val (v_1).

If the $(i-1)$st step is reached $i-1 \geqq 1$, thus Perm (E_{i-1}), Ready (V_{i-1}), $(\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_{i-1}})$ and (val $(v_1), \cdots,$ val (v_{i-1})) have been determined, then

(4.2) (i) $\mathrm{Perm}\,(E_i) = (\mathrm{Perm}\,(E_{i-1}) - \mathrm{SH}_{v_{i-1}}) \cup \{(v_{i-1}, w) \in \delta;\ w \in V\}$;
(ii) $\mathrm{Ready}\,(V_i) = \{v \in V$; there exists a v-shrub SH_v such that $\mathrm{SH}_v \subseteq \mathrm{Perm}\,(E_i)\}$, and *either* Ready $V_i \neq \varnothing$ and
(iii) one selects $v_i \in \mathrm{Ready}\,(V_i)$ and its v_i-shrub such that $\mathrm{SH}_{v_i} \subseteq \mathrm{Perm}\,(E_i)$; then one applies int (nam (v_1)) to (val $(w_1), \cdots,$ val (w_k)) when $\mathrm{SH}_{v_i} = \{(w_1, v_i), \cdots, (w_k, v_i)\}$ and val (w_j) = Init (nam (w_j)) if $w_j \in R$; and one denotes the result value of the application by val (v_i); after that one repeats (ii);
or Ready $(V_i) = \varnothing$ and
(iv) one terminates the construction of me and $\mathrm{val}_{\mathrm{me}}$, and defines the *resultation* Result (v) = val (v) for each $v \in L$.

In (4.1)(iii) and (4.2)(iii) one is selecting one single vertex (the seriality of the execution rule), but, in general, one can select any subset of Ready (V_i). Then a *parallel permit execution rule* is concerned (see [Culi 78]). The selection of one or more vertices, of the set, Ready (V_i), is arbitrary, and represents some sort of *indeterminism*, which is the reason that for some interpretation and for the same initialization two different resultations can be determined. Therefore DG need not define a function at all.

An execution sequence that is a me $= (\mathrm{SH}_{v_1}, \cdots, \mathrm{SH}_{v_n}, \cdots)$, either finite or infinite, in a DG (interpreted and initialized) which is obtained according to (4.1) and (4.2) is called *completed* if

(4.3) each function application required in (4.1)(iii) or (4.2)(iii) can be actually performed, and its result value is obtained.

If DG is a data graph and int is its interpretation, then let the *algorithmic domain of* $\mathrm{DG}_{\mathrm{int}}$ be the following set of initializations Init of DG;

(4.4) (i) ADomain $(\mathrm{DG}_{\mathrm{int}})$ = {Init; each me obtained according to (4.1) and (4.2) is completed}, and let the *domain of* $\mathrm{DG}_{\mathrm{int}}$ be its subset as follows

(4.4) (ii) Domain $(\mathrm{DG}_{\mathrm{int}}) = \{\mathrm{Init};$ each finite me obtained according to (4.1) and (4.2) is completed$\}$.

A data graph DG is called *functional* if

(4.5) for each interpretation int of DG and for each initialization $\mathrm{Init} \in$ ADomain $(\mathrm{DG}_{\mathrm{int}})$ any two completed execution sequences determine the same result.

If DG is functional then the *function* $F_{\mathrm{DG,int}}$: Domain $(\mathrm{DG}_{\mathrm{int}}) \rightarrow$ Range $(\mathrm{DG}_{\mathrm{int}}) = \{\mathrm{Result};$ Result determined by $\mathrm{Init} \in \mathrm{Domain}\,(\mathrm{DG}_{\mathrm{int}})\}$ is uniquely determined. Obviously if $p = |R|$ and $q = |L|$, then $F_{\mathrm{DG}_{\mathrm{int}}}$ represents a p/q-function, or, in other words, a system of q p-ary functions with the same Domain $(\mathrm{DG}_{\mathrm{int}})$.

LEMMA 4.1. *If* DG *satisfies* (2.5), *that is, each* $v \in V - R$ *has exactly one v-shrub, then* DG *is functional.*

Proof. Let us assume DG is not functional and derive from it that DG does not satisfy (2.5). From the assumption follows the existence of an interpretation, int, an initialization of DG, and of two completed metapaths me and me* in DG by which two different resultations are determined. It means there exists a leaf $v_{i_1} \in L$ such that $\mathrm{val}_{\mathrm{me}}(v_{i_1}) \neq \mathrm{val}^*_{\mathrm{me}^*}(v_{i_1})$. These two different values are results of applications of the same function $\mathrm{int}\,(\mathrm{nam}\,(v_{i_1}))$, and therefore $\mathrm{int}\,(\mathrm{nam}\,(v_{i_1}))$ had to be applied to two different argument sets. According to (1.2)(iv) $\mathrm{ar}\,(v_{i_1}) > 0$, and according to (1.2)(ii) $\mathrm{idg}\,(v_{i_1}) > 0$. If there are two different v_{i_1}-shrubs in me and me*, it is a contradiction with (2.5), and the proof is finished. If there exists only one v_{i_1}-shrub $\mathrm{SH}_{v_{i_1}}$, then there must exist a root v_{i_2} of $\mathrm{SH}_{v_{i_1}}$ such $\mathrm{val}_{\mathrm{me}}(v_{i_2}) \neq \mathrm{val}^*_{\mathrm{me}^*}(v_{i_2})$ and $i_2 < i_1$, and, one can repeat the dilemma above and construct $v_{i_1}, v_{i_2}, \cdots, v_{i_j}$, such that $1 \leqq i_j < \cdots < i_2 < i_1 \leqq n$ such that $\mathrm{val}_{\mathrm{me}}(v_{i_j}) \neq \mathrm{val}^*_{\mathrm{me}^*}(v_{i_j})$ but each root w of the single SH_{v_j} satisfies $w \in R$. This is possible only when there are two different v_{i_j}-shrubs which violates (2.5), and the proof is completed.

LEMMA 4.2. *If* DG *is finite, acyclic and simple, but does not satisfy* (2.5), *then* DG *is not functional.*

Proof. As DG does not satisfy (2.5), there exists a v_0 in DG such that there are two different v_0-shrubs SH_{v_0} and $\mathrm{SH}^*_{v_0}$. Therefore there exist their roots $v_1 \neq v_1^*$ such that $(v_1, v_0) \in \mathrm{SH}_{v_0}$, $(v_1^*, v_0) \in \mathrm{SH}^*_{v_0}$ and $\Delta(v_1, v_0) = \Delta(v_1^*, v_0)$. Then among all such vertices v_0 one can choose one such that sDG_{v_1} and $\mathrm{sDG}_{v_1^*}$ either satisfy (2.5) or $v_1 \in R$ and $v_1^* \in R$. Therefore SH_{v_1} and $\mathrm{SH}_{v_1^*}$ are determined uniquely (if $v_1, v_1^* \notin R$).

According to Lemma 2.2 there exists a rooted metapath $\mathrm{me} = (\mathrm{SH}_{w_1}, \cdots, \mathrm{SH}_{w_n})$, $\mathrm{me}^* = (\mathrm{SH}_{w_1^*}, \cdots, \mathrm{SH}_{w_m^*})$ in DG which is admissible and contains SH_{v_0}, $\mathrm{SH}^*_{v_0}$, respectively, and therefore it must contain also SH_{v_1}, $\mathrm{SH}_{v_1^*}$, respectively, which means $v_1 = w_i$, $v_1^* = w_j^*$ where $1 \leqq i \leqq n$, $1 \leqq j \leqq m$.

As sDG_{v_1} satisfies (2.5) and SH_{v_1} is a v_1-shrub in sDG_{v_1} each shrub in DG_{v_1} belongs to me also, and let $\mathrm{sme} = (\mathrm{SH}_{w_{i_1}}, \cdots, \mathrm{SH}_{w_{i_k}})$, where $1 \leqq i_1 < \cdots < i_k = i$, be a subsequence of me which is a rooted metapath in sDG_{v_1}. Similarly for $\mathrm{sDG}_{v_1^*}$ let $\mathrm{sme}^* = (\mathrm{SH}_{w^*_{j_1}}, \cdots, \mathrm{SH}_{w^*_{j_h}})$ where $1 \leqq j_1 < \cdots < j_h = j$, be a subsequence of me* of all shrubs which belong to $\mathrm{sDG}_{v_i^*}$.

Let $\mathrm{Init}\,(\mathrm{nam}\,(v)) = \mathrm{nam}\,(v)$ for each $v \in R$ and if $v \in V - R$, let int be a modified Herbrand interpretation defined as follows: $\mathrm{int}\,(\mathrm{nam}\,(v), \mathrm{SH}_v) = \mathrm{SH}_v$ if $\mathrm{SH}_v = \{(u_1, v), \cdots, (u_p, v)\}$ and $u_i \in R$ for $i = 1, \cdots, p$, while for $u_i \in V - R$: if $\mathrm{int}\,(\mathrm{nam}\,(u_i), \mathrm{SH}_{u_i})$ has been defined with edges in δ_i and with one single leaf u_i, then $\mathrm{int}\,(\mathrm{nam}\,(v), \mathrm{SH}_v) = \mathrm{SH}_v \cup \bigcup_{i=1}^{p} \delta_i$. By the given set of edges a data subgraph sDG (δ')

is determined uniquely (with all the labelings of vertices and edges), and one defines that the value int (nam (v), SH$_v$) = δ' and int (nam (w), SH$_w$) = δ'' are equal if sDG(δ') and sDG(δ'') are data isomorphic, writing DG (δ') ≡ sDG (δ'').

Now one can execute DG once with the execution sequence me and the values will be denoted by val (w_i), and for the second time with the execution sequence me*, when the values will be denoted by val* (w_j). First we want to show val (v_0) ≠ val* (v_0).

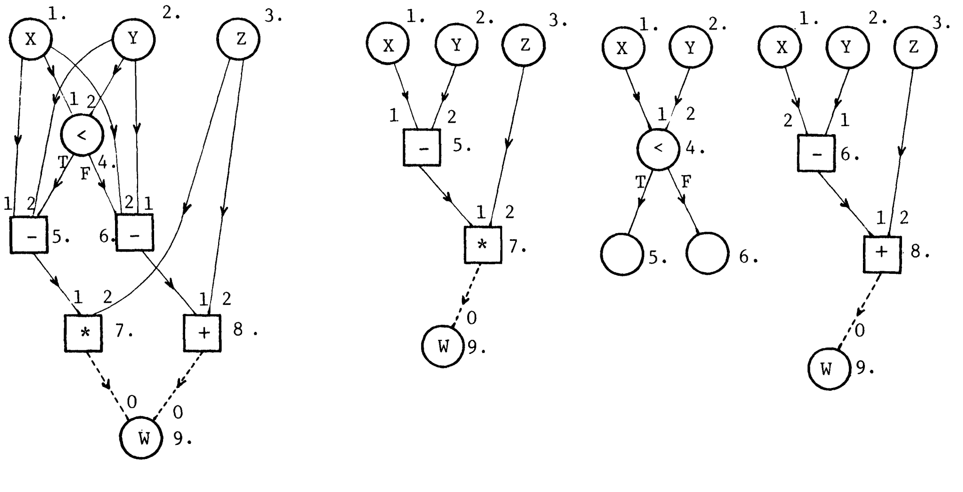
If v_1, $v_1^* \in R$, then int (nam (v_0), SH$_{v_0}$) = δ' where $(v_1, v_0) \in \delta'$, and int (nam (v_0), SH$^*_{v_0}$) = δ'' where $(v_1^*, v_0) \in \delta''$. As nam ($v_1$) ≠ nam ($v_1^*$) (according to (1.2)(i)), val (v_0) = sDG (δ') and val* (v_0) = sDG (δ'') are not data isomorphic, and therefore val (v_0) ≠ val* (v_0).

If either $v_1 \in R$ and $v_1^* \in V-R$, or $v_1 \in V-R$ and $v_1^* \in R$ then again val (v_0) ≠ val* (v_0), because in any data isomorphism between sDG (δ') and sDG (δ''), the shrub SH$_{v_0}$ must correspond to SH$^*_{v_0}$. Therefore v_1 also corresponds to v_1^*, but v_1 is a root of sDG (δ') while v_1^* is not a root of sDG (δ'').

It remains the case $v_1 \in V-R$ and $v_1^* \in V-R$ when SH$_{v_1}$ and SH$_{v_1^*}$ exist. Now we can execute DG with the execution sequence sme with val (w_i) = sDG$_{v_1}$ and with the execution sequence sme* with val* (w_j^*) = sDG$_{v_1^*}$. As DG is simple, sDG$_{v_1}$ and sDG$_{v_1^*}$ are not data isomorphic (in virtue of Theorem 3.2), and therefore val (v_1) ≠ val* (v_1^*), from which it follows immediately that val (v_0) ≠ val* (v_0) again.

Thus we have shown val (v_0) ≠ val* (v_0), and it remains to find a leaf $w \in L$ of DG such that val (w) ≠ val* (w). Obviously each path in DG of maximal length and containing v_0 has such a leaf as its end vertex. From the properties of Herbrand interpretation val (w) ≠ val* (w) follows immediately. Thus DG is not functional, which is what we wanted to prove.

5. Acyclic data graphs with predicates. A *conditional term* (which is a generalization of the usual term [Mann 74]), if $X < Y$ then $((Y-X)^*Z)$ else $((X-Y)+Z)$, can be represented by an acyclic data graph with predicates in Fig. 5.1a where *predicate shrubs* (shrubs with roots labeled by a predicate name from PredNam) are admitted, and new sorts of edges, called *signal* (*control*) *edges*, which leave predicate vertices, and are labeled by *truth values* T, F, are added.



a) b)

FIG. 5.1

The two usual terms $((Y-X)^*Z)$ and $((X-Y)+Z)$ are represented in Fig. 5.1b by the corresponding extended data graphs, which determine, according to permit execution rule (4.1) and (4.2), two functions. The extended predicate shrub in Fig. 5.1b causes that only certain partial functions with mutually disjoint domains are needed, and the identification of their extended leaves represents the union of the two partial functions.

In Fig. 5.1a two types of vertices are differentiated: *signal-free*, as 4, 7, 8, when no signal edge is terminating in it, and *signal-sensitive*, as 5, 6 when at least one signal edge is terminating in it. If v is a signal-sensitive vertex, then a v-shrub will, in addition, contain exactly one signal edge terminating in v. For example, $\mathrm{SH}_5=\{(1,5),(2,5),(4,5)\}$ and $\mathrm{SH}_6=\{(1,6),(2,6),(4,6)\}$ are such shrubs.

The leaf 9 in Fig. 5.1a is named by a variable (and not by a function or a predicate). Nevertheless the concept of its shrub is applicable anyway: $\mathrm{SH}_9=\{(7,9)\}$ and $\mathrm{SH}_9^*=\{(8,9)\}$ are two different 9-shrubs. Therefore the condition (2.5) is too strong to play the crucial role for the functionality.

A data graph with predicates $\mathrm{DGP}=\langle V, W, \delta, \sigma, \mathrm{ar}, \mathrm{nam}, \Delta\rangle$ is a natural generalization of a data graph without predicates $\mathrm{DG}=\langle V, \delta, \mathrm{ar}', \mathrm{nam}', \Delta'\rangle$ satisfying (1.1)–(1.3). Assuming R, L is the set of all roots, leaves, respectively, the following requirements are assumed:

(5.1) (i) $V\cap W=\varnothing$; $L\subseteq V$;
(ii) $\varnothing\neq\delta\subseteq V\times(V\cup W)$; $\sigma\subseteq W\times(V\cup W)$;
(iii) $R\cap V\neq\varnothing\neq L$; $R\cap L\neq\varnothing$; $V-(R\cup L)\neq\varnothing$;
(iv) $(t,t)\notin\delta\cup\sigma$ for each $t\in V\cup W$;
(v) there are no isolated vertices.

(5.2) (i) $\mathrm{ar}(t)=0$ for each $t\in R\cup L$ and $\mathrm{ar}(t)>0$ for each $t\in(V\cup W)-(R\cup L)$;
(ii) $0\leqq\mathrm{ar}(t)\leqq\mathrm{idg}(t)$ for each $t\in V\cup W$;
(iii) $\mathrm{nam}(t)=\mathrm{nam}(u)\Rightarrow\mathrm{ar}(t)=\mathrm{ar}(u)$ for all $t,u\in V\cup W$;
(iv) $t,u\in R$ and $t\neq u\Rightarrow\mathrm{nam}(t)\neq\mathrm{nam}(u)$;
(v) $v\in V-(R\cup L)$, $w\in W-R\Rightarrow\mathrm{nam}(v)\neq\mathrm{nam}(w)$.

(5.3) (i) $t\in(V\cup W)-(R\cup L)$ and $(u,t)\in\delta\Rightarrow 1\leqq\Delta(u,t)\leqq\mathrm{ar}(t)$;
(ii) $t\in L$ and $(u,t)\in\delta\Rightarrow\Delta(u,t)=0$;
(iii) $1\leqq i\leqq\mathrm{ar}(t)\Rightarrow$there exist $u\in V$ and $(u,t)\in\delta$ such that $\Delta(u,t)=i$;
(iv) $t\in W$ and $(t,u)\in\Rightarrow$either $\Delta(t,u)=T$ or $\Delta(t,u)=F$;
(v) $t\in W\Rightarrow$there exist u_1, $u_2\in V\cup W$ and (t,u_1), $(t,u_2)\in\sigma$ such that $\Delta(t,u_1)\neq\Delta(t,u_2)$.

Data graphs with predicates defined by (5.1)–(5.3) are actual generalizations of conditional terms, because they need not be acyclic, but even within the class of acyclic DGPs a real generalization exists. Fig. 5.2a represents an acyclic DGP which cannot be expressed by any conditional term. The generalized term, if $X<Y$ then if $((Y-X)^*Z)<0$ then FLOOR $((Y-X)^*Z)$ else CEILING $((Y-X)^*Z)$ else $((X-Y)+Z)$, is represented by an acyclic DGP in Fig. 5.2b. It is the best we can expect because, obviously, the same function is defined by all of them, but if executed, the term $((Y-X)^*Z)$ will be evaluated three times as some vertices are split in several parts.

On the other hand Fig. 5.2c represents a useless data graph with predicates because the single 20-shrub $\mathrm{SH}_{20}=\{(7,20),(8,20)\}$ will never be permitted and executed, as

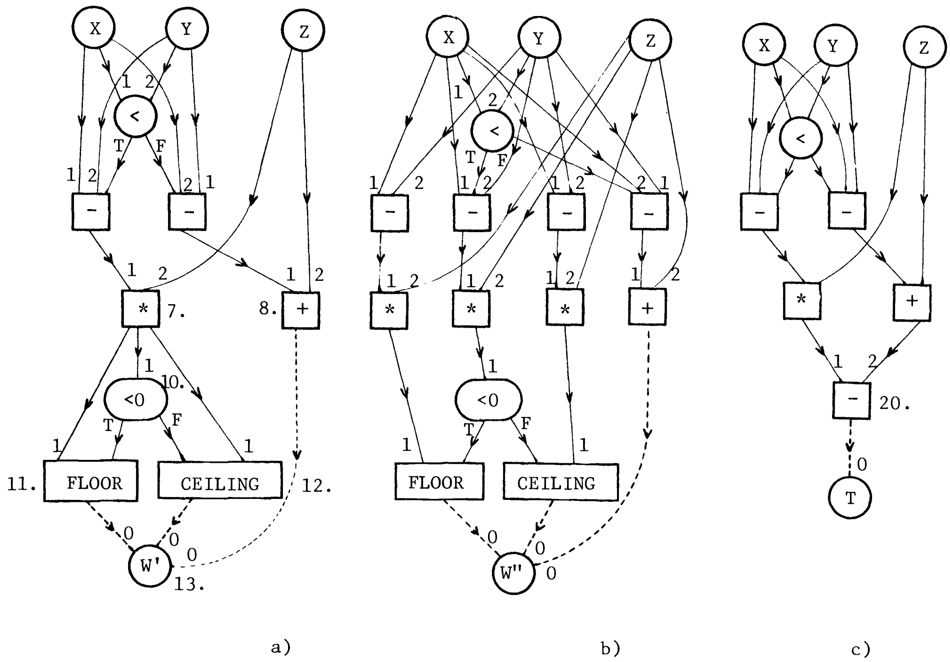


FIG. 5.2

the two edges (7, 20) and (8, 20) will never be permitted simultaneously in the same rooted metapath.

Two v-shrubs SH_v and SH_v^* in an acyclic DGP, where $v\in(V\cup W)-R$, are called *incompatible* if

(5.4) there exists i, $1\leqq i\leqq \mathrm{ar}(v)$ such that for two edges $(w_i, v)\in\mathrm{SH}_v$ and $(w_i^*, v)\in\mathrm{SH}_v^*$ $(\Delta(w_i, v)=\Delta(w_i, v)=i)$ there exist two usual paths $\mathrm{pa}=(t_1, t_2, \cdots, t_{n-1}, t_n)$ and $\mathrm{pa}^*=(t_1^*, t_2^*, \cdots, t_{m-1}^*, t_m^*)$ such that $t_1=t_1^*$, $(t_1, t_2)\in\sigma$, $(t_1^*, t_2^*)\in\sigma$ and $\Delta(t_1, t_2)\neq\Delta(t_1^*, t_2^*)$.

The two 9-shrubs in Fig. 5.1a are incompatible.

LEMMA 5.1. *If a* DGP *represents a conditional term then after splitting each* $t\in R$ *into* $\mathrm{odg}(t)$ *new roots a finite tree arises from* DGP. *Further,* $\mathrm{odg}(v)=1$ *for each* $v\in V-R$, *and* $\mathrm{odg}(w)=2$ *for each* $w\in W-R$. DGP *does not satisfy* (2.5) *but satisfies*
(5.5) $t\in(V\cup W)-R$ *and* $\mathrm{SH}_t\neq\mathrm{SH}_t^*\Rightarrow\mathrm{SH}_t$ *and* SH_t^* *are incompatible.*

Proof. The first assertion follows from the corresponding definition immediately, and the satisfaction of (5.5) follows from the previous part, if one realizes that according to (5.3)(ii) only leaves may not satisfy (2.5).

The concept and definition of rooted and leafed metapath in a DGP is formally the same as presented in (2.1), and also the concept and definition of data homomorphism concerning DGPs is presented in (3.1) and (3.2). The same lemmata as in §§ 2 and 3 can be proved for DGPs.

The (serial) permit execution rule (4.1) and (4.2) can be applied to DGPs after the adding of the corresponding definitions concerning interpretation and execution of predicates. It is assumed $\{\mathrm{nam}(v);\ v\in V-(R\cup L)\}=\mathrm{FunctNam}$, $\{\mathrm{nam}(w);\ w\in W-R\}=\mathrm{PredNam}$, $\{\mathrm{nam}(t);\ t\in R\cup L\}=\mathrm{Var}$, and $\mathrm{int}:\mathrm{PredNam}\to\mathrm{Pred}$ is an interpretation of predicate names by actual predicates.

The execution rules (4.1) and (4.2) must be augmented by the following way of finding the permitting edges

(4.2*) (i*) if $v_{i-1} \in V - L$ then $\mathrm{Perm}\,(E_i) = (\mathrm{Perm}\,(E_{i-1}) - \mathrm{SH}_{v_{i-1}}) \cup \{(v_i - 1, t) \in \delta;$ $t \in V \cup W\}$;

(i**) if $v_{i-1} \in W - L$ then $\mathrm{Perm}\,(E_i) = (\mathrm{Perm}\,(E_{i-1}) - \mathrm{SH}_{v_{i-1}}) \cup \{(v_{i-1}, t) \in \sigma;$ $\Delta(v_{i-1}, t) = \mathrm{val}\,(v_{i-1})$ and $t \in V \cup W\}$;

(i***) if $v_{i-1} \in L$ then $\mathrm{Perm}\,(E_i) = (\mathrm{Perm}\,(E_{i-1}) - \mathrm{SH}_{v_{i-1}})$;

here $\mathrm{val}\,(v_i)$, $v_i \in W$, is a truth value defined according to

(4.1*) (iii*) one selects $v_i \in \mathrm{Ready}\ V_i$ and one $\mathrm{SH}_{v_i} = \{(t_0, v_i), (t_1, v_i), \cdots, (t_k, v_i)\}$ such that $\mathrm{SH}_{v_i} \subseteq \mathrm{Perm}\,(E_i)$; then one applies $\mathrm{int}\,(\mathrm{nam}\,(v_i))$ to $(\mathrm{val}\,(\mathrm{nam}\,(t_1)), \cdots, \mathrm{val}\,(\mathrm{nam}\,(t_k)))$, and denotes the result value (which is either T or F) by $\mathrm{val}\,(v_i)$.

The definition of functionality (4.5) remains formally the same for DGPs also. If acyclic and finite DGPs are concerned, the role of requirement (2.5) is replaced by the weaker requirement (5.5), and the following assertions are proved in very similar ways as previous Lemmata 4.1 and 4.2:

LEMMA 5.2. *If a* DGP *satisfies* (5.5), *then it is functional.*

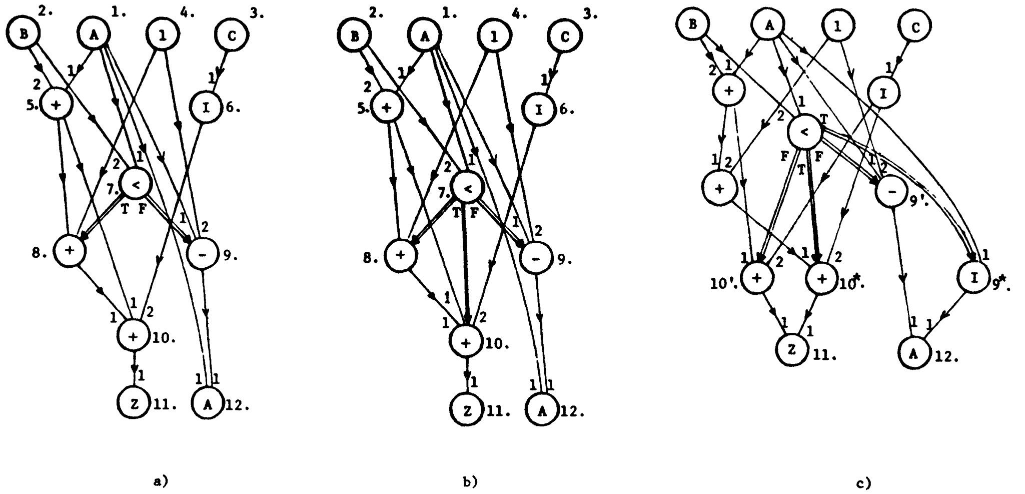
LEMMA 5.3. *If a* DGP *is finite, acyclic, simple but does not satisfy* (5.5), *then it is not functional.*

Fig. 5.3a presents a possible data graph with predicates derived from a control graph of a program segment

Get List(A, B, C); $X = A + B$; $Y = C$; if $A < B$ then $X = X + 1$ else $A = A - 1$;

$Z = X + Y$; Put List(A),

where the double arrows are original control edges, viewed now as signal edges, and I is the unary identity function. According to the execution rule (4.1) and (4.2), the vertex 10 is permitted before and independently of the execution of 7. Therefore 10 should be forbidden to become permitted before the execution of 7, because if the signal edge (7, 8) belonged to permitting edges then Fig. 5.3a would be not functional



a) b) c)

FIG. 5.3

and might define a wrong value. To forbid 10 means to make it signal-sensitive, as is the case in Fig. 5.3b, but it does not help very much, because 10 is still permitted before 8 is executed and therefore a wrong result may be obtained.

The actual solution is in Fig. 5.3c, where the vertex 10 is split into two vertices 10′ and 10* each of which is control-sensitive, but incompatible signal edges lead to them from 7. Similarly, the vertex 9 is also split into two vertices 9′ and 9*. In fact the two shrubs of 11 and of 12 in Fig. 5.3c are incompatible according to (5.4), while the shrubs of 10 and 12 in Fig. 5.3a or 5.3b are not, and Figs. 5.3a and 5.3b are not functional and do not satisfy (5.5).

There seems to be no need to introduce auxiliary vertices (called *selector* and *distributor* in [DaKe 82], [Denn 74]), which do not directly correspond to functions and predicates.

Data graphs with predicates, when interpreted and initialized, allow one to determine decisions, which are unions of domain disjoint functions. By each execution sequence (rooted and leafed metapaths) of a finite and acyclic DGP, a set of leaves is determined, and one can partition the set of all execution sequences into classes according to mutual disjoint sets of leaves $L_1, L_2, \cdots, L_h$, which are determined by them. Where $1 \leqq h \leqq 2^k$, h is called the *decision degree*, and $k \geqq 0$ is the number of predicate vertices in DGP. The decisions are a crucial component of repetitions, but they are not investigated further.

6. Nonacyclic data graphs. From the point of view of execution, and also of graph theory, data graphs which are not acyclic are of the greatest importance and interest; they are obviously also the most complex. Closed metapaths which occur in nonacyclic data graphs correspond to repetitions from the execution point of view.

The repetition is a new and independent feature which requires some modifications concerning the execution rule and the data graph with predicate itself.

Each DGP may be provided with a subset $\text{rep} \subseteq \delta \cup \sigma$, called a set of *repeating* edges, such that

(6.1) (i) $(u, t) \in \text{rep} \Rightarrow \text{ar}(t) > 1$ and there exists $(v, t) \in \delta \cup \sigma$ such that $\Delta(u, t) \neq \Delta(v, t)$ and $(v, t) \notin \text{rep}$;
(ii) $(u, t) \in \text{rep} \Rightarrow$ there exists a closed metapath in DGP containing the vertex t.

The (*serial*) *permit execution rule for a* $\text{DGP} = \{V, W, \delta, \sigma, \text{rep}, \text{ar}, \text{nam}, \Delta\rangle$, which is interpreted by int and initialized by Init, is defined as follows:

(6.2) (i) at the beginning $\text{Perm}(E_1) = \{(u, t) \in \delta \cup \sigma; u \in R\}$;
(ii) then $\text{Ready}(V_1) = \{t \in V \cup W$; there exists a t-shrub SH_t such that $\text{SH}_t \subseteq \text{Perm}(E_1)\}$;
(iii) one selects a $t_1 \in \text{Ready}(V_1)$ and one of its SH_{t_1} such that $\text{SH}_{t_1} \subseteq \text{Perm}(E_1)$, where either t_1 is a signal-free and $\text{SH}_{t_1} = \{(u_1, t_1), \cdots, (u_k, t_1)\}$, or t_1 is a signal-sensitive and $\text{SH}_{t_1} = \{(u_0, t_1), (u_1, t_1), \cdots, (u_k, t_1)\}$, $(u_0, t_1) \in \sigma$ while $(u_j, t_1) \in \delta$ for $j = 1, 2, \cdots, k$; then one applies $\text{int}(\text{nam}(t_1))$ to $\text{Init}(\text{nam}(u_1)), \cdots, \text{Init}(\text{nam}(u_k))$, and denotes the result of application by $\text{val}(t_1)$ which is either a value from val, or T or F (if $\text{int}(\text{nam}(t_1))$ is a predicate).

(6.3) assuming $\text{Perm}(E_{i-1})$, $\text{Ready}(V_{i-1})$, $(\text{SH}_{t_1}, \cdots, \text{SH}_{t_{i-1}})$ and $(\text{val}(t_1), \cdots, \text{val}(t_{i-1}))$ has been already determined, and $i - 1 \geqq 1$,
(i) one defines $\text{Perm}(E_i) = (\text{Perm}(E_{i-1}) - (\text{SH}_{t_{i-1}} - \text{rep})) \cup E$ where
(a) $E = \{(t_{i-1}, u) \in \delta; u \in V \cup W\}$ if $t_{i-1} \in V - L$;

(b) $E=\{(t_{i-1},u)\in\sigma;\ \Delta(t_{i-1},u)=\mathrm{val}\,(t_{i-1})$ and $u\in V\cup W\}$ if $t_{i-1}\in W-L$;
(c) $E=\varnothing$ if $t_{i-1}\in L$;

(ii) then Ready $(V_i)=\{t\in V\cup W$; there exists a t-shrub SH_t such that $\mathrm{SH}_t\subseteq \mathrm{Perm}\,(E_i)\}$; and *either* Ready $(V_i)\neq\varnothing$, and

(iii) one selects $t_i\in\mathrm{Ready}\,(V_i)$ and one its t_i-shrub SH_{t_i} such that $\mathrm{SH}_{t_i}\subseteq\mathrm{Perm}\,(E_i)$; then one applies int (nam (t_i)) to (val $(u_i),\cdots,$ val (u_k)) where $\mathrm{SH}_{t_i}=\{(u_0,t_i),\ (u_1,t_i),\cdots,(u_k,t_i)\}$ and $(u_0,t_i)\in\sigma$ is omitted when t_i is not signal-sensitive, and val $(u_j)=\mathrm{Init}\,(u_j)$ if $u_j\in R$; finally the result of the application of int (nam (t_i)) is denoted by val (t_i); and after that one repeats (6.3) again;
or Ready $(V_i)=\varnothing$, and

(iv) one terminates the construction of me and $\mathrm{val}_{\mathrm{me}}$, and defines the resultation Result $(t_i)=\mathrm{val}\,(t_i)$ for each $t_i\in L$ with the greatest index i.

Fig. 6.1a represents a nonacyclic DGP provided with rep $=\{(2,4),(3,6)\}$ which satisfies (6.1)(i) and also (ii) because me $=(\mathrm{SH}_4,\mathrm{SH}_6,\mathrm{SH}_8)$ is a closed metapath satisfying (2.4) if $\mathrm{SH}_4=\{(2,4),(6,4)\}$, $\mathrm{SH}_6=\{(4,6),(8,6),(3,6)\}$ and $\mathrm{SH}_8=\{(6,8)\}$, and IDENT is unary identity function.

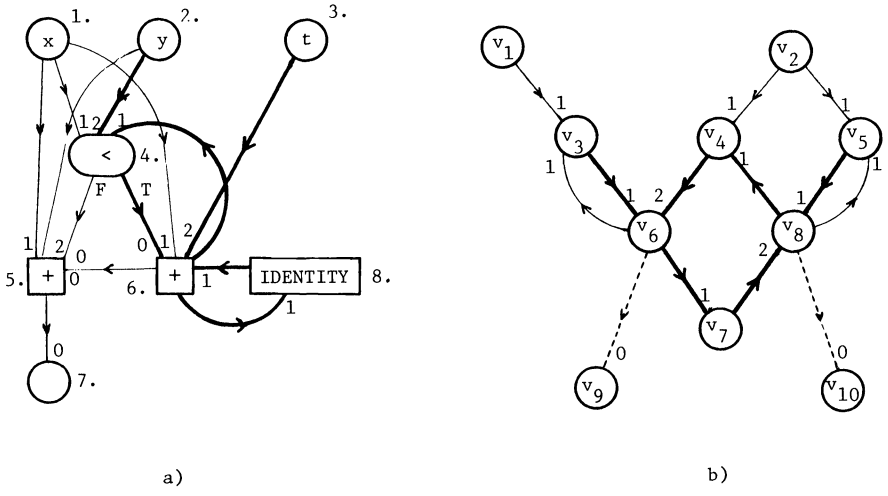


FIG. 6.1

Fig. 6.1a is a functional DGP but the vertex 4 does not satisfy either (2.5) or (5.5), and therefore even a weaker condition than (5.5) should be found to characterize the functionality of DGPs which are not acyclic.

Fig. 6.1b presents a nonacyclic DG without predicates, which never terminates and therefore the functionality condition (4.5) is not applicable directly. Nevertheless its suitable weakening such that the same sequence of values of leaves v_9 and v_{10} should be determined by all execution sequences, is satisfied. Unfortunately again (5.5) is not satisfied.

The problem remains to characterize the functionality of nonacyclic DGPs.

7. Applications in computer science and recursive function theory. Data graphs with or without predicates are motivated by data flow analysis or programs. They are an alternative to program (flowchart) schemata, a special sort of algorithm by which functions can be defined (computed). Data graphs are not expected to be algorithms suitable for problem solving and programming [Denn 74], but are suitable for parallel machines, and for theoretical and optimizing purposes.

Theoretically it seems to be an interesting approach to study various compositions of algorithms instead of compositions of functions. Different types of vertex and edge identifications (concerning metagraphs) can be used [Culi 83]. It seems to be a way to clarify recursive program schemes [Mann 74], and to differentiate a recursion which does not depend on a well-founded universe.

REFERENCES

[BoMu 76] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, North-Holland, Amsterdam, 1976.

[Culi 78] K. Culik, *Almost control-free (indeterministic) parallel computation on permit schemes*, Conference Record of the Fifth Annual ACM Symposium on Principles of Programming Languages, January, 1978,ˈpp. 176–184.

[Culi 83] ———, *Operators of variable-sensitive functions computed by programs*, Proc. Seventeenth Annual Conference on Information Science and Systems, John Hopkins Univ., Baltimore, March 1983, pp. 502–507.

[DaKe 82] A. I. Davis and R. M. Keller, *Data flow program graphs*, IEEE J. Comput. (1982), pp. 26–41.

[Denn 74] J. B. Dennis, *First version of a data flow procedure language*, Proc. International Symposium on Programming, Paris, April 9–11, 1974, Springer, Berlin, 1974, pp. 362–376.

[Grze 74] A. Grzegorczyk, *An Outline of Mathematical Logic*, Reidel, Dordrecht, 1974.

[Klee 67] S. C. Kleene, *Mathematical Logic*, John Wiley, New York, 1967.

[Koni 36] D. König, *Theorie der endlichen und unendlichen Graphen*, Leipzig, 1936.

[Mann 74] Z. Manna, *Mathematical Theory of Computation*, McGraw-Hill, New York, 1974.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

011

ON THE ENCODING OF RELATIONS BY GRAPHS*

YORAM MOSES† AND AMOS NOY‡

Abstract. Encodings of relations by graphs are viewed as 1–1 mappings from relations to directed graphs. A measure called the *size* of such encodings is defined. A subclass of encodings called *translations* is introduced. These are essentially encodings decodable by a first order formula. Lower bounds on their sizes are proven. We present a translation whose size is asymptotically equal to the lower bound, differing from it only in low order terms.

A closely related notion of *perfect* encoding is defined as being a 1–1 and onto mapping, that also has the interesting property that every edge in an encoding graph corresponds directly to a tuple in the original relation, and vice-versa. A perfect encoding is constructed and using it, a classical result about random graphs is converted into a result about random relations. It is believed that many other results can be similarly converted using these notions.

Generalizations of this work to encodings of relations by relations of lower degree are described, and analogous methods are shown to work.

AMS subject classifications. 04A05, 05C99

1. Introduction. It is well known that k-ary relations can be encoded by binary relations. Most readers will have come across this fact, but will not have ever seen or constructed such encodings. It is not very hard to come up with these encodings, although in many cases initial attempts are flawed. It is much harder to come up with efficient ones, i.e., encodings in which the graph that encodes a relation has a small number of nodes. It is natural to ask whether actual encodings can give us interesting information, e.g. show relationships between k-ary relations and graphs, and allow us to translate results in the theory of graphs to the theory of k-ary relations, and vice-versa.

In this paper, we investigate specific encodings of k-ary relations by graphs. We present a sequence of encodings, each of which is interesting in its own right.

An *encoding* of the k-ary relations by graphs is defined to be an injection from the k-ary relations to graphs. This is a very general notion, and we refine it by defining the notion of a *translation*, that is, roughly, an encoding that can be decoded by a first order formula. A notion of the *size* of an encoding is defined, and we prove lower bounds on the sizes of general encodings and of translations.

In § 3 we present specific translations of decreasing sizes possessing interesting combinatorial properties. Our two last encodings have asymptotic sizes that match the lower bound, but a small gap in lower order terms remains.

In § 4 we introduce the notion of a perfect encoding, a small encoding in which every tuple in the encoded relation corresponds to an edge in the graph that encodes it and vice versa. Using the constructions of § 3, we present perfect encodings. We show an example due to E. Shamir that derives a result regarding selection sets in random relations from a classical result on random graphs, using one of our translations.

Section 5 discusses generalizations of this work to the case of encoding k-ary relations by l-ary relations ($l \leqq k$).

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Computer Science Department, Stanford University, Stanford, California 94305. The work of this author was supported in part by ARPA under grant N00039-8-C-0250.

‡ Department of Computer Science, Hebrew University, Givat Ram, Jerusalem, 91904, Israel.

2. Definitions and preliminary results. We call $G=(V, E)$, where V is a finite set and $E \subset V^2$, a graph. Similarly, $R=(A, P)$, with A a finite set and $P \subset A^k$ is called a k-ary relation ($k \geqq 2$). Denote by $\mathscr{G}$ and $\mathscr{R}_k$ the sets of all graphs and all k-ary relations respectively. By $\mathscr{G}(n)$ (resp. $\mathscr{R}_k(n)$) we mean the set of all graphs (resp. k-ary relations) over n elements. The sets of vertices V (resp. A) will in this case have cardinality n. N denotes the natural numbers.

A 1–1 mapping $\sigma: \mathscr{R}_k \to \mathscr{G}$ is called an encoding of the k-ary relations by graphs (abbrev. k-encoding). For a k-encoding σ, we define $\text{size}_\sigma: N \to N$ as follows:

$$\text{size}_\sigma = \min \{\text{polynomial } p(n) \colon \forall n \exists m \cdot m > n \wedge \sigma(\mathscr{R}_k(m)) \subset \mathscr{G}(p(m))\}.$$

By our definition $\{\mathscr{G}(m)\}_{m \in \mathscr{N}}$ and $\{\mathscr{R}_k(n)\}_{n \in \mathscr{N}}$ are towers of sets, i.e. $\mathscr{G}(i) \subset \mathscr{G}(i+1)$ and $\mathscr{R}_k(j) \subset \mathscr{R}_k(j+1)$, $i, j = 0, 1, \cdots$. We identify two graphs if they have the same set of edges, and similarly for relations. Given a set B we write $\sigma(B)$ as shorthand for $\{\sigma(b) \colon b \in B\}$.

LEMMA 1. *Let σ be a k-encoding, $n > 0$. Then* $\text{size}_\sigma(n) \geqq n^{k/2}$.

Proof. By counting, $|\mathscr{R}_k(n)| = 2^{n^k}$, and $|\mathscr{G}(m)| = 2^{m^2}$. For $m = \text{size}_\sigma(n)$, $\sigma: \mathscr{R}_k(n) \to \mathscr{G}(m)$, and since σ is 1–1, we have $2^{m^2} \geqq 2^{n^k}$, $m^2 \geqq n^k$, and $m \geqq n^{k/2}$, i.e. $\text{size}_\sigma(n) \geqq n^{k/2}$. Q.E.D.

Given any pair of enumerations, one for $\mathscr{R}_k$ and one for $\mathscr{G}$, there is a canonical encoding corresponding to that pair—the one that maps the first relation to the first graph etc. However, these encodings may very well be quite random and unnatural. A more natural thing to ask for is that the relation be definable in terms of the graph, or "decodable" from it in a prescribed way. This motivates the following definitions:

A pair of formulas $\delta = \langle \Phi(x), \Psi(x_1, \cdots, x_k) \rangle$, where Φ and Ψ are both first order formulas in a language with equality and a single binary predicate E, is called a k-decoder. We say that $E(x, y)$ is true (or satisfied) in a graph G, denoted $G \vDash E(x, y)$, if the edge (x, y) occurs in G. More elaborate formulas are interpreted accordingly.

Given a k-decoder δ as above, and a graph $G=(V, E)$, let

$$A_{\delta,G} = \{v \in V \colon G \vDash \Phi(v)\},$$

and

$$P_{\delta,G} = \left\{ \langle v_{i_1}, \cdots, v_{i_k} \rangle \colon G \vDash \Psi(v_{i_1}, \cdots, v_{i_k}) \wedge \bigwedge_{1 \leqq j \leqq k} \Phi(v_{i_j}) \right\}.$$

We say that $R_{\delta,G} = (A_{\delta,G}, P_{\delta,G})$ is the k-ary relation decoded by δ from G. δ can therefore be viewed as a mapping $\delta: \mathscr{G} \to \mathscr{R}_k$, where $\delta(G) = R_{\delta,G}$.

A k-encoding τ is called a *k-translation* if there is a fixed k-encoder δ_τ, such that for all $R \in \mathscr{R}_k$, $\delta_\tau(\tau(R)) = R$. δ_τ is therefore an inverse of τ. This requires that if $\tau(R) = G$, for $R = (A, P)$, $G = (V, E)$, then $A \subset V$ and δ_τ can recognize A in V, and decode P, both solely from the structure of G.

Gaifman observed a stricter lower bound for translations than the one shown for general encodings in Lemma 1:

LEMMA 2. *Let τ be a (k-)translation, $n > 0$. Then*

$$\text{size}_\tau(n) \geqq n^{k/2} + \frac{k-2}{4} n^{-k/2+1} \log n + \frac{\log e}{2} n^{-k/2+1} + o(n^{-k/2+1}).$$

Proof. Let $\delta_\tau = (\Phi, \Psi)$ be τ's decoder. There are $\binom{\text{size}_\tau(n)}{n}$ ways to choose n points from $\text{size}_\tau(n)$. δ_τ's Φ chooses them in a single predetermined way. For each encoding graph there are therefore $\binom{\text{size}_\tau(n)}{n}$ graphs isomorphic to it, having the same set of vertices, that do not encode a relation. We now have

$$2^{\text{size}_\tau(n)^2} \geqq \binom{\text{size}_\tau(n)}{n} 2^{n^k}.$$

Substituting $n^{k/2}+\alpha(n)$ for $\text{size}_\tau(n)$, and taking logs of both sides, we get:

$$(n^{k/2}+\alpha(n))^2 \geqq \log_2 \binom{n^{k/2}+\alpha(n)}{n} + n^k,$$

$$2n^{k/2}\alpha(n)+\alpha(n)^2 \geqq \log_2 \binom{n^{k/2}+\alpha(n)}{n} = \log_2 \frac{(n^{k/2}+\alpha(n))\cdots(n^{k/2}+\alpha(n)-n+1)}{n!}$$

$$\geqq \log_2 \frac{(n^{k/2}+\alpha(n)-n)^n}{n!}$$

$$= n \log_2 (n^{k/2}+\alpha(n)-n) - \log_2 n! \geqq n \log_2 (n^{k/2}-n) - \log_2 n!$$

$$\approx n \log_2 n^{k/2} + n \log_2 e - n \log_2 n = \left(\frac{k}{2}-1\right) n \log_2 n + n \log_2 e,$$

so

$$2n^{k/2}\alpha(n)+\alpha(n)^2 \geqq \left(\frac{k}{2}-1\right) n \log_2 n + n \log_2 e,$$

$$\alpha(n)^2+2n^{k/2}\alpha(n) - \left(\frac{k}{2}-1\right) n \log_2 n - n \log_2 e \geqq 0,$$

$$\alpha(n) \geqq -n^{k/2} + \sqrt{n^k + \left(\frac{k}{2}-1\right) n \log_2 n + n \log_2 e}$$

$$= n^{k/2}\left(-1+\sqrt{1+\frac{k-2}{2}\frac{n \log_2 n}{n^k}+\frac{n \log_2 e}{n^k}}\right)$$

$$\approx n^{k/2}\left(-1+1+\frac{k-2}{4} n^{-k+1} \log_2 n + \frac{\log_2 e}{2} n^{-k+1}\right)$$

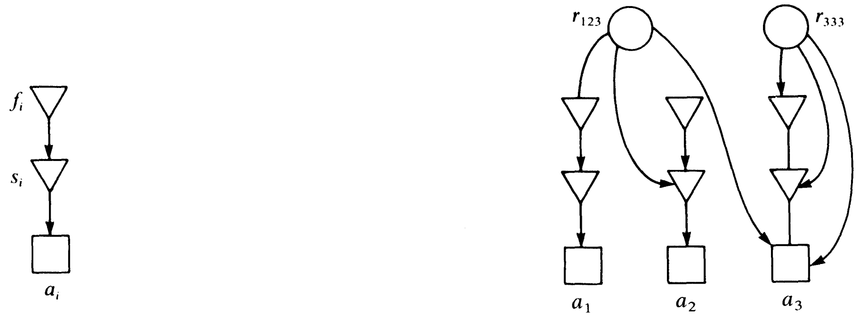
$$= \frac{k-2}{4} n^{-k/2+1} \log_2 n + \frac{\log_2 e}{2} n^{-k/2+1}.$$

Now, by our substitution, $\text{size}_\tau(n)$ is as claimed. Q.E.D.

Notice that Lemma 2 is still based on a counting argument, and not on the expressive power of our first order language. We believe that this lower bound can be improved on. We conjecture that the true lower bound is of the form $n^{k/2}+\Omega(n^{1/2})$, although this is an open question. Gaifman [3] has a characterization of first order properties definable in graphs, and it is an interesting question whether it can be used to strengthen this lower bound.

3. Translations. We now set out to look for upper bounds on $\text{size}_\tau(n)$, by constructing explicit translations. Our examples will be for $k=3$, but will generalize to any k in a natural way. We fix a ternary relation $R=(A, P)$, with $|A|=n$.

τ_1: *The totem pole translation.* For every element $a_i \in A$, we add two auxiliary nodes f_i ("first"$_i$), and s_i ("second"$_i$), and connect the three as in Fig. 1(a). In the general case, we would add $k-1$ auxiliary nodes. We call this construction a totem pole. For each tuple $\langle a_i, a_j, a_k\rangle \in P$ we add a ("relational") node r_{ijk} that points at f_i,



(a) *A totem-pole of height* 3.

(b) *Translation of* $\{\langle a_1, a_2, a_3\rangle, \langle a_3, a_2, a_3\rangle\}$.

FIG. 1. τ_1: *The totem-pole translation.*

s_j and a_k. Each node in the totem pole acts as a place holder, and an edge pointing at it specifies what place the a_i at its bottom has in the tuple.

More formally, define $G = \tau_1(R)$ as $G = (V, E)$, where

$$V = \{f_i, s_i, a_i\colon a_i \in A\} \cup \{r_{ijk}\colon \langle a_i, a_j, a_k\rangle \in P\},$$

$$E = \{(r_{ijk}, f_i), (r_{ijk}, s_j), (r_{ijk}, a_k)\colon \langle a_i, a_j, a_k\rangle \in P\} \cup \{(f_i, s_i), (s_i, a_i)\colon a_i \in A\}.$$

Figure 1(b) is an example of $\tau_1(R)$ for $R = (A, P)$ with $A = \{1, 2, 3\}$, and $P = \{\langle 1, 2, 3\rangle, \langle 3, 3, 3\rangle\}$.

In order to show that τ_1 is a translation, we must supply an adequate decoder: Define $\delta_1 = (\Phi_1, \Psi_1)$, where

$$\Phi_1(x) \equiv \forall y. \neg E(x, y),$$

$$\Psi_1(x, y, z) \equiv \exists r, f_1, s_1, s_2.\ E(r, f_1) \wedge E(f_1, s_1) \wedge E(s_1, x) \wedge E(r, s_2) \wedge E(s_2, y) \wedge E(r, z).$$

It is easy to check that δ_1 is a decoder for τ_1. We need $O(n^k)$ nodes for an encoding with τ_1, because every tuple in the original relation is represented by a node in the encoding graph. Since for every $a \in A$ we add $k-1$ nodes to the graph, we have $\text{size}_{\tau_1}(n) = n^k + kn$.

τ_2: *The bitotem matrix translation.* This translation involves a use of τ_1 that is based on the following idea: Assume $n = h^2$ for an appropriate h. Let us organize the elements of A in an $h \times h$ matrix. Every (k-)tuple over elements of A induces a corresponding tuple over their row coordinates, and similarly over their column coordinates. This correspondence between tuples in the relation and pairs of (row, column) tuples in the matrix is 1–1 and onto, so that every such pair of tuples uniquely determines a tuple in the relation. For notational clarity we assume that the elements of A are $a_{11}, \cdots, a_{1h}, a_{21}, \cdots, a_{hh}$. To represent each row and each column coordinate, we add k new nodes connected as in Fig. 1(a) (a totem pole on each coordinate, rather than on each element). For every tuple in P, construct the induced relational elements over the row totems and over the column totems. Then add an edge from the row relational node to the column one. Figure 2(a) shows how to encode $\langle a_{11}, a_{21}, a_{12}\rangle$. The bottom of each row (resp. column) totem points at all the elements of its row (resp. column). An edge from a row relational node $rr_{i_1i_2i_3}$ to a column relational $cr_{j_1j_2j_3}$ represents the tuple $\langle a_{i_1j_1}, a_{i_2j_2}, a_{i_3j_3}\rangle$. This is clearly well-defined and unambiguous, and all we need to show is that it can be decoded by some decoder δ_2. Define $\delta_2 = (\Phi_2, \Psi_2)$

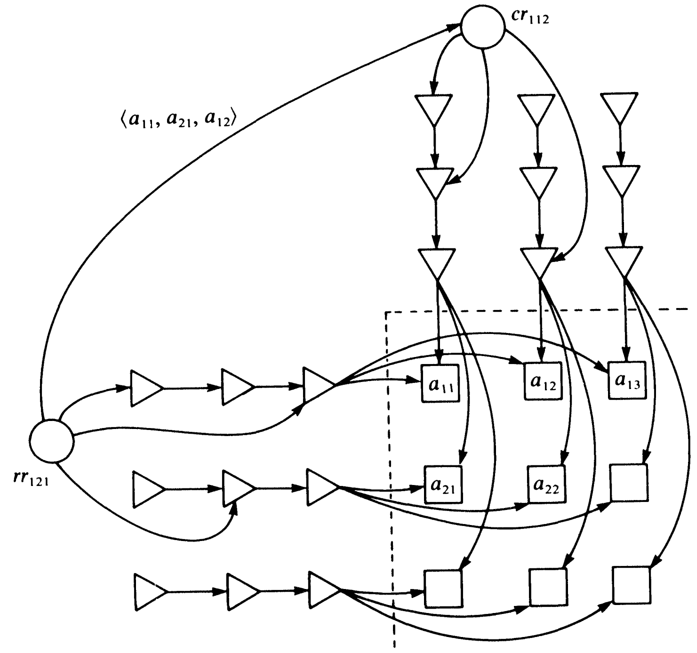


FIG. 2(a). *The bitotem matrix translation.*

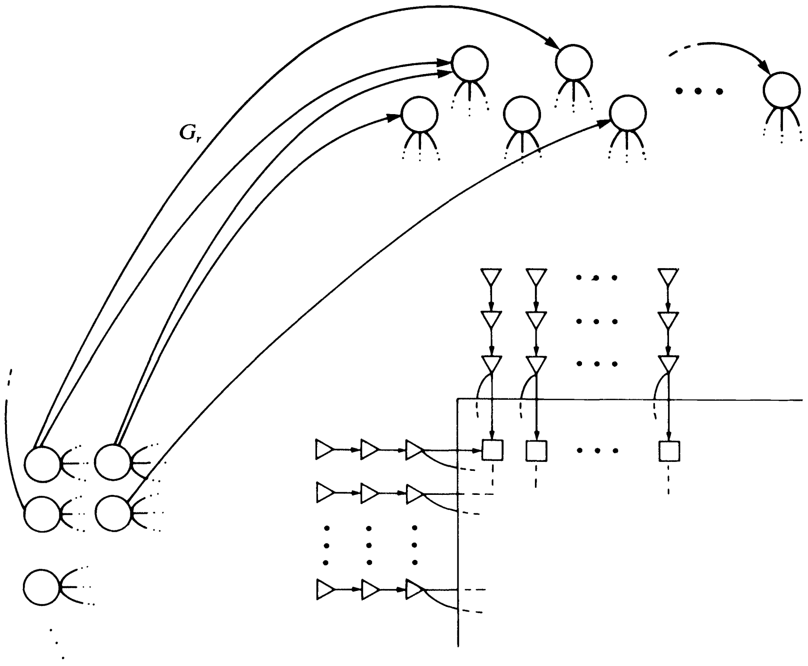


FIG. 2(b). *G_r—the "relational" subgraph.*

by:

$\Phi_2(x) = \Phi_1(x) \equiv \forall y. \neg E(x, y)$,

third $(x) \equiv \exists y. \Phi_2(y) \wedge E(x, y)$,

second $(x) \equiv \exists y.$ third $(y) \wedge E(x, y) \wedge$ outdegree $(x) = 1$,

first $(x) \equiv \exists y.$ second $(y) \wedge E(x, y) \wedge$ outdegree $(x) = 1$,

relational $(x) \equiv x$ points at a "third", a "second" and a "first",

row-relational $(x) \equiv x$ points at a relational,

$\Psi_2(x, y, z) \equiv \exists$ two relational elements, and 12 others, such that the construction described above holds.

Writing Ψ_2 explicitly is somewhat long, but straightforward from the above definitions.

The size of τ_2 is $2h^k + h^2 + 2kh$, or $2n^{k/2} + n + 2k\sqrt{n}$. That is because (again we need only consider the encoding of the full relation over A) we have h row coordinates, and by our original τ_1 there are $h^k + kh$ vertices representing the k-ary relations over the rows. The same holds for the columns, and we get the extra h^2 vertices from the elements of A that are in the matrix.

τ_3: *The totem matrix translation.* A slight modification of τ_2; let $R \in \mathcal{R}_k(n)$, $G = \tau_2(R)$. Observe the subgraph $G_r \subset G$, the subgraph of G generated by the relational nodes (see Fig. 2(b)). This is a bipartite graph, where every edge goes from a row-relational node to a column-relational one. The basic idea is that we can "fold" this bipartite graph into an unrestricted graph by interpreting each relational node as a row-relational node whenever it is the source of an edge, and as a column-relational node whenever it is a sink. We do this by having only one set of coordinates (see Fig. 3), having the "third" element in each coordinate point at a column in the matrix, and the "second" point at a row. Now an edge $E(r_{i_1 i_2 i_3}, r_{j_1 j_2 j_3})$ corresponds to the

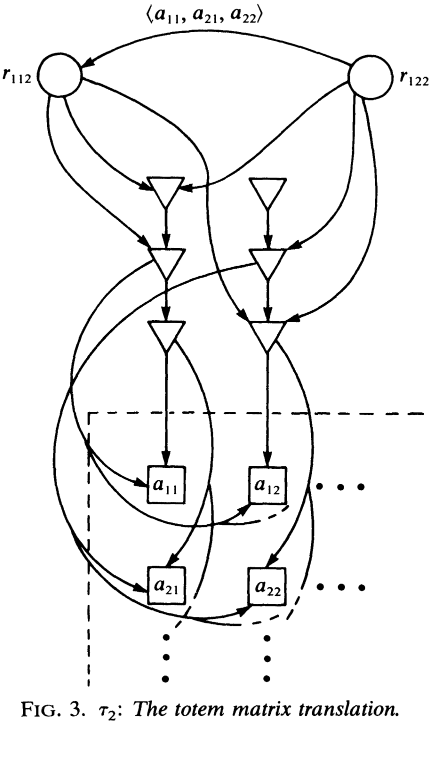


FIG. 3. τ_2: *The totem matrix translation.*

original triple $\langle a_{i_1j_1}, a_{i_2j_2}, a_{i_3j_3}\rangle$. Figure 3 shows how $\langle a_{11}, a_{21}, a_{22}\rangle$ is encoded. The reader can modify δ_2 to get δ_3. There are no new notions involved, and it is a straightforward modification.

Since we now have only one set of h coordinates and the relational elements over them, we arrive at $\text{size}_{\tau_3}(n) = n^{k/2} + n + kn^{1/2}$.

τ_4: *The diagonal translation.* In this last translation we try to reduce the size as much as we can. The reader that is not interested in this matter will lose nothing relevant to later parts by skipping it.

For our final translation, geared to reduce the number of nodes used for the coordinates, we use the following observations:

(a) If the coordinates are ordered then totem poles are not needed anymore. A relational element need only point at the coordinates it involves, and indicate in what order it treats them. This indication must make unambiguous the cases when the relational element points at less than k coordinates. This specification can be done by singling out $\log_2 k \times k! \approx (k+1) \log_2 k$ nodes, and having each relational element point at a subset of them. Note that this trick reduces the number of nodes per coordinate, at the cost of making the size of δ_4 very large (many disjuncts appear in it—doing things the straightforward way, $(k!)^2$). Only one coordinate is needed for every row and one for every column ($2\sqrt{n}$ vs. $k\sqrt{n}$). The coordinates can be ordered by having each one of them point at all the coordinates it precedes (this way, given two coordinates, $c_1 > c_2$ iff $E(c_1, c_2)$).

(b) After applying (a) to the totem matrix translation, we have $2\sqrt{n}$ nodes devoted to the coordinates. If we organize our original nodes in the upper triangle of a matrix then we can use the elements on the diagonal for coordinates. Specifying two coordinates will unambiguously determine a single element of our upper triangle. Specifying the same coordinate twice determines the diagonal element serving as the coordinate itself. This requires only $\sqrt{2n}$ coordinates, instead of $2\sqrt{n}$.

(c) We can use the matrix elements that are not on the diagonal as relational vertices. This reduces the size of our translations by $n - \sqrt{2n}$.

We do need to add a few nodes to the graph that will serve to distinguish between different kinds of nodes. We will need to distinguish diagonal nodes, nondiagonal matrix nodes, nonmatrix relational nodes. We can partition the nodes of a graph unambiguously in a first order definable manner by building what we call a D-tree in which each vertex is definable, and then having those vertices each stand for a part in the partition, and point at all the members of that part. The idea in constructing a D-tree is to begin with a root and have it point to a single son. This son now will have two sons, one that points back at him and one that does not. This can now be extended to a vertex definable binary tree, or the arity of the nodes can be increased, if we wish, by more careful schemes of backwards pointing. Every nontree node in the graph should be pointed at by some vertex of the D-tree, to avoid ambiguity. In our case, corresponding to part (a), one of the D-tree vertices will start a $(k+1) \log_2 k$ ring that will be used by the relational elements to indicate the tuple ordering among their coordinates.

Figure 4 is an example of such an encoding. We will again refrain from constructing δ_4 explicitly. The interested reader can follow our description and fill in the formal details.

We now have: $\text{size}_{\tau_4}(n) = n^{k/2} + \sqrt{2}n^{1/2} + (k+1) \log_2 k + o(n^{1/2} + k)$. We have thus arrived at an upper bound with the same asymptotic behavior as our lower bounds from Lemmas 1 and 2. By padding the matrix whenever n is not a square or a triangular number, this fact does not change, although the expression for the *size* then is not as small and as clean as the above. In fact, $\text{size}_{\tau_4}(n)$ above is quite close to the lower

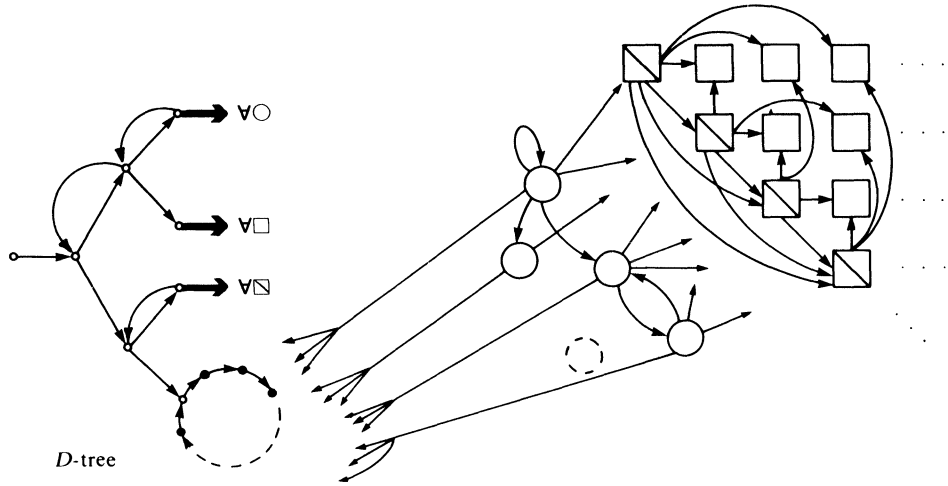


FIG. 4. *The diagonal translation.*

bound of Lemma 2. As we have mentioned earlier, we conjecture that the true lower bound is of the form $n^{k/2}+\Omega(n^{1/2})$, although this is still an open question.

4. Perfect encodings and applications. Lemma 1 states that for any encoding σ, $\text{size}_\sigma(n) \geqq n^{k/2}$. We call an encoding *perfect* if equality holds and furthermore every edge in the graph corresponds to a tuple in the relation. Do concise perfect encodings exist?

Lemma 2 implies that translations cannnot be perfect encodings. On the other hand, since an encoding is just a 1–1 mapping: $\mathcal{R}_k \rightarrow \mathcal{G}$, given effective enumerations of $\mathcal{R}_k$ and of $\mathcal{G}$ of the right type, one can construct a perfect encoding. These enumerations, as stated earlier, might be neither concise nor natural.

Our constructions in the bitotem and totem matrix translations give an interesting way of relating relations to graphs. Let us call the subgraph of an encoding graph generated by the relational nodes alone, its relational subgraph (denoted G_r). Notice that its edges are in 1–1 correspondence with the tuples in the original relation. An edge appears in G_r if and only if its corresponding tuple appears in the relation. In the bitotem case G_r is a bipartite graph with $\leqq n^{k/2}$ nodes in each part. In the totem matrix case this subgraph has $\leqq n^{k/2}$ nodes. In both cases, every node in G_r corresponds to a k-tuple of "half elements", and each edge makes a k-tuple out of two such creatures. The main difference is that in the bitotem case we have a bipartite graph, whereas in the totem matrix case, an unrestricted one.

The encoding mapping each relation to the G_r subgraph of its totem matrix graph is therefore a perfect encoding. Another perfect encoding that comes to mind is an encoding where every node in the graph would correspond to a $k/2$-tuple, and an edge would just specify how to glue them together. Of course, for an odd k a node would stand for a ($[k/2]$ and a half)-tuple where the half would be treated via a matrix. It is easy to see how such an encoding leads to translations parallel to our bitotem and totem matrix ones. In the bitotem parallel, we would need totem-poles of only half the height.

In fact, at this point it should be clear that there are many alternative ways to define encodings, translations and perfect encodings, and the machinery developed in § 3 can form the conceptual basis for many of them.

A very pleasing property of perfect encodings is that generating random graphs and generating random relations now become equivalent matters, the perfect encoding supplying the means to convert one into the other.

The G_rs in the bitotem case are the bipartite equivalent of a perfect encoding. The bitotem and totem matrix cases are so closely related that roughly anything you wish to do with one you can do with the other, and in some cases it will be easier to work with the bitotem encoding. In the bitotem case a random bipartite graph will correspond to a random relation.

E. Shamir suggested the following application:

In the bitotem setting, as we have mentioned, G_r is a bipartite graph. It is known [1] that if we choose at random $m \log m$ edges between two disjoint sets of vertices V_1 and V_2 of cardinality m, we will have a matching with probability $\rightarrow 1$ as $m \rightarrow \infty$. If we take all $(n^{k/2})$ row-relational nodes to be V_1, the column-relational ones to be V_2, we would have $m = n^{k/2}$, and with $n^{k/2} \log n^{k/2} (= (k/2) n^{k/2} \log n)$ random edges, there would be a matching. The property induced by a matching in the G_r on the encoded relation is that every row and every column would be represented at least $n^{(k-1)/2}$ times in every component in the relation (a "selection set" with threshold $n^{(k-1)/2}$). A somewhat more delicate analysis can be done to show that $(1/2k)\sqrt{n} \log n$ k-tuples chosen at random are enough to promise that at least one member of each row and column appear in the relation (again, with probability $\rightarrow 1$ as $n \rightarrow \infty$).

5. Generalizations. All the translations presented in § 3 generalize in a natural way from $k = 3$ to any $k \geqq 2$, and from encoding k-ary relations in graphs to the general case of encoding k-ary relations in l-ary relations, where $k \geqq l \geqq 2$. In the case of $l > 2$, an l-cube would replace our matrix, and the primary diagonal of the l-cube would replace the matrix's diagonal. The sizes of the totem l-cube and the l-cube diagonal translations would now be:

$$\text{size}_{\tau_3}(n)_{k,l} = n^{k/l} + n + kn^{1/l}, \qquad \text{size}_{\tau_4}(n)_{k,l} = n^{k/l} + (kn)^{1/l} + o(n^{1/l} + k).$$

Fagin [2] has independently shown equivalent ways of translating from k to $k-1$ in $O(n^{k/k-1})$.

6. Conclusions. We have defined the notion of an encoding of k-ary relations by graphs and introduced the notions of translations, perfect encodings and the size of an encoding. Lower bounds on the sizes of encodings and translations were given, and encodings that match the lower bound were presented along with translations that are larger than the lower bound only in lower order terms. A conjecture regarding the possible strengthening of the lower bound for translations was given, hinting at the connection this problem has to first order expressiveness. An example of how our constructions may be used to relate properties of graphs to those of relations was given. Perfect encodings provide a simple and straightforward way to convert a random graph generator into a random relation generator.

We believe that the notions introduced here may help convert between properties of graphs and those of relations. Relatively little has been done in pursuing this path. Our encodings and others give insight into the fundamental links and differences between graphs and k-ary relations.

Acknowledgments. Work on this paper was initiated as an attempt to answer a seminar question raised by Gaifman. We thank him also for observing Lemma 2. We would like to thank E. Shamir for his encouragement and his contribution to § 4. E. Belaga, D. Dolev, R. Fagin, E. Mayr and M. Vardi have contributed through fruitful discussions.

REFERENCES

[1] P. Erdös and A. Rényi, *On random matrices* I, Publ. Math. Inst. Hung. Acad. Sci., 8A (1963), pp. 455–461.

[2] R. Fagin, *A spectrum hierarchy*, Z. Math. Logik und Grundlagen Math., 21 (1975), pp. 123–134.

[3] H. Gaifman, *On local and non-local properties*, Proc. the Herbrand Logic Colloquium, Marseilles, 1981.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

012

THRESHOLD DIMENSION OF GRAPHS*

MARGARET B. COZZENS† AND ROCHELLE LEIBOWITZ‡

Abstract. This paper examines the problem of determining the threshold dimension of a graph. There exists numerous characterizations of threshold graphs, those graphs of threshold dimension one, as well as fast polynomial time algorithms to test if a graph is a threshold graph. Yannakakis [1982] proved that, in general, determining the threshold dimension is a hard problem by proving that for fixed $k \geqq 3$, determining if the threshold dimension of a graph is less than or equal to k is an NP-complete problem. In this paper we compute the threshold dimension of several classes of graphs and obtain upper and lower bounds on the threshold dimension of a graph. Counterexamples to a conjecture on threshold dimension are provided.

AMS subject classification. 05

1. Threshold graphs.

1.1. Introduction. Threshold graphs were introduced by Chvátal and Hammer in 1973 as a class of graphs for which there is a particularly simple method of distinguishing independent sets from nonindependent sets. Let G be a graph with vertex set: $\{v_1, v_2, v_3, \cdots, v_n\}$, and edge set E. For any subset S of V, define a characteristic vector $(x_1, x_2, \cdots, x_n)$ such that

$$x_i = \begin{cases} 1 & \text{if } v_i \in S, \\ 0 & \text{if } v_i \notin S, \end{cases} \qquad i = 1, 2, \cdots n.$$

Each subset S of V corresponds to a corner of the unit hypercube in $\mathbb{R}^n$. We want to know if a hyperplane exists that cuts n-space in half so that the corners of the hypercube corresponding to independent sets lie on one side of the hyperplane and the corners of the hypercube corresponding to nonindependent sets lie on the other side of the hyperplane. If such a hyperplane exists, the graph is said to be a *threshold graph*. Equivalently, $G = (V, E)$ is a threshold graph if there exists a threshold assignment (a, t) consisting of a labeling a of the vertices by nonnegative integers and an integer threshold t such that

$$(1) \qquad S \text{ is independent} \Leftrightarrow \sum_{v \in S} a(v) \leqq t \qquad (S \subseteq V).$$

Clearly, complete graphs are threshold graphs, by taking $a(v) = 1$ for each $v \in V$ and $t = 1$. The stars $K_{1,n}$ are also threshold graphs by taking $a(v) =$ degree of v, and $t = n$. We will see further examples of threshold graphs later in this paper.

There exist many characterizations of threshold graphs. We will present only a few of them here, notably the ones most frequently used. Threshold graphs have a nice forbidden subgraph characterization as seen in the first theorem.

THEOREM 1 (Chvátal and Hammer [1973]). *A graph is a threshold graph if and only if it has no generated subgraphs isomorphic to $2K_2$, P_3, or Z_4. These graphs are shown in Fig.* 1.

* Received by the editors July 6, 1983. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

‡ Department of Mathematics, Wheaton College, Norton, Massachusetts 02766.

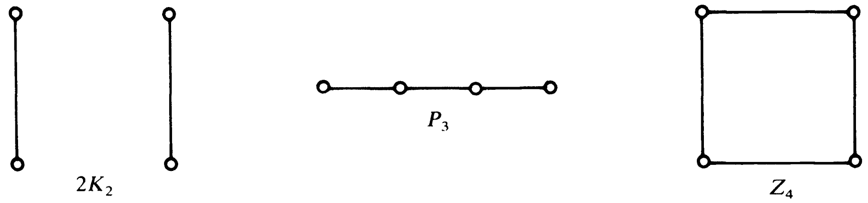


FIG. 1

Equivalently, Theorem 1 could be expressed as follows.

COROLLARY 1.1. *A graph is a threshold graph if and only if it does not contain the configuration shown in Fig. 2, where no line between vertices allows for the possibility of the edge existing or not.*

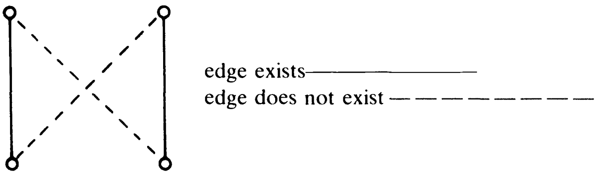


FIG. 2

THEOREM 2 (Chvátal and Hammer [1973]). *A graph is a threshold graph if and only if for each subset S of G there exists a vertex $u \in S$ such that u is adjacent to all vertices in $S-\{u\}$ or to none of them.*

Both Theorems 1 and 2 are testable conditions. It is easy to see from either of these theorems that the complement of a threshold graph is a threshold graph, and that any generated subgraph of a threshold graph is a threshold graph.

There are two types of orderings related to threshold graphs. For a graph $G=(V, E)$, define an equivalence relation R on the vertices of G by:

$$xRy \Leftrightarrow N(x)-\{y\}=N(y)-\{x\}$$

where $N(x)$ is the set of vertices adjacent to x. We can now define a partial order on the set of equivalence classes E_a of R, and call it the *vicinal preorder* associated with G:

$$E_a \rho E_b \Leftrightarrow E_a \neq E_b \text{ and } N(b) \subseteq N(a) \cup \{a\}.$$

For example, if G is the graph shown in Fig. 3, then the equivalence classes are: $E_a=\{a, c\}=E_c$, $E_b=\{b\}$, $E_d=\{d\}$, and $E_e=\{e\}$. Thus, we can take the set of equivalence classes to be $\{E_a, E_b, E_d, E_e\}$ and the partial order is $\{(E_a, E_e), (E_d, E_e)$ $(E_d, E_b)\}$.

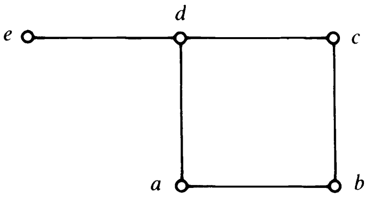
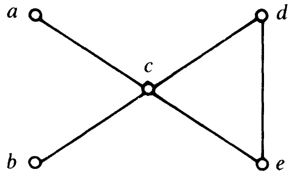


FIG. 3

THEOREM 3 (Peled and Simeone [1981]). *A graph G is a threshold graph if and only if the vicinal preorder associated with G is a linear order.*

It is possible to associate with each threshold graph G a second type of ordering, called an M-ordering and a corresponding 0–1 sequence. An ordering of the vertices $u_1, u_2, u_3, \cdots, u_n$ is an *M-ordering* of $V(G)$ with *M-sequence* $M_1, M_2, \cdots, M_{n-1}$ if for $i<j$, $\{u_i, u_j\} \in E(G) \Leftrightarrow M_i = 1$ and $\{u_i, u_j\} \notin E(G) \Leftrightarrow M_i = 0$. Figure 4 gives an example of a threshold graph with an M-ordering and associated M-sequence.



M-ordering: c, b, a, e, d
M-sequence: 1, 0, 0, 1

FIG. 4

If G is a threshold graph, Theorem 2 provides us with a way to get an M-ordering with corresponding M-sequence. By Theorem 2, there exists a vertex $u \in V(G)$ such that u is adjacent to every vertex of G, or none of the vertices of G. Choose this vertex and call it u_1. If u_1 is adjacent to all vertices of G then $M_1 = 1$, otherwise $M_1 = 0$. Choose a vertex in $V - \{u_1\}$ that is adjacent to all vertices in $V - \{u_1\}$ or to no vertex in $V - \{u_1\}$ and call it u_2 with $M_2 = 1$ in the former case, $M_2 = 0$ in the latter case. We continue in this manner to get an M-ordering of vertices, $u_1, u_2, \cdots, u_n$, with a corresponding M-sequence. In general, the M-orderings of the vertices are not unique. For example, the graph shown in Fig. 4 has three other M-orderings c, a, b, d, e; c, a, b, e, d; and c, b, a, d, e. But all four M-orderings have the same M-sequence $1, 0, 0, 1$. In fact the M-sequence is always unique. In the next section we give an application of threshold graphs and M-orderings to attitude measurement theory.

Threshold graphs have been studied by numerous authors and various applications exist in addition to the one described in the next section. Golumbic [1980] discusses an application to synchronized parallel processing due to Henderson and Zalcstein [1977]. For a more extensive summary of the properties and characterizations of threshold graphs, the reader should see Golumbic [1980].

1.2. Guttman scales. In this section we present an application of M-orderings of threshold graphs to finding Guttman scales. A Guttman scale is a linear ordering of subjects and items such that a subject agrees with all items following it and disagrees with all items preceding it. Guttman scales have been used widely in educational testing, opinion scaling, etc. We will define a graph G_p and show that a Guttman scale exists if and only if G_p is a threshold graph. Furthermore, if G_p is a threshold graph, a certain kind of M-ordering of G_p is a Guttman scale.

THEOREM 4. *A Guttman scale exists if and only if it is not the case that subject x agrees with item a but not item b while subject y agrees with b but not a.*

Form a bipartite graph P representing the situation as follows: $V(P) = T \cup I$ where T is the set of subjects and I is the set of items. There exists an edge between $x \in T$ and $a \in I$ if and only if x agrees with a. Rephrasing Theorem 4, we have:

COROLLARY 4.1. *A Guttman scale exists if and only if $2K_2$ is not a generated subgraph of P.*

The above theorem and corollary give tests for determining the existence of a Guttman scale but they do not give an ordering of the subjects and items. A certain kind of M-ordering of a particular threshold graph, called G_p, will be a Guttman scale. G_p is formed from P by connecting every pair of subjects. That is, T generates a clique of G_p.

LEMMA 1. *$2K_2$ and Z_4 are never generated subgraphs of G_p.*

THEOREM 5. *P has a Guttman scale if and only if G_p is a threshold graph.*

Proof. Suppose P does not have a Guttman scale. Then $2K_2$ is a generated subgraph of P. Now P_3 is a generated subgraph of G_p and G_p is not a threshold graph.

Suppose G_p is not a threshold graph. By Lemma 1, G_p contains P_3 as a generated subgraph. Keeping in mind that T generates a clique of G_p and I is a set of isolated vertices, we must have both u and v as subjects and both r and s as items. Hence $2K_2$ is a generated subgraph of P, implying that P does not have a Guttman scale by Corollary 4.1. Q.E.D.

We can assume that no two subjects agree with exactly the same items and no two items are agreed with by exactly the same subjects. For if a Guttman scale exists on the remaining vertices, all subjects agreeing with exactly the same items can be placed consecutively, and similarly for all items agreed with by exactly the same subjects. Thus, we have a Guttman scale on all vertices. And conversely, if a Guttman scale exists on all vertices, then a Guttman scale exists on a subset of vertices.

COROLLARY 5.1. *Suppose that no two subjects agree with exactly the same items and no two items are agreed with by exactly the same subjects. Suppose G_p is a threshold graph. Then the following two statements are true:*

a) *There exists an M-ordering of G_p, $u_1, u_2, \cdots, u_{n-1}, u_n$, satisfying the following condition:*

$$(*) \qquad \{u_{n-1}, u_n\} \in E(G_p) \text{ if and only if } u_{n-1} \text{ is a subject.}$$

b) *Suppose $M = u_1, u_2, \cdots, u_{n-1}, u_n$ is an M-ordering of G_p satisfying $(*)$. Then M defines a Guttman scale.*

Proof. (a) Suppose $M = u_1, u_2, \cdots, u_{n-1}, u_n$ is an M-ordering of G_p which does not satisfy $(*)$. Thus, either $\{u_{n-1}, u_n\} \in E(G_p)$ and u_{n-1} is an item or $\{u_{n-1}, u_n\} \notin E(G_p)$ and u_{n-1} is a subject. Suppose $\{u_{n-1}, u_n\} \in E(G_p)$ and u_{n-1} is an item. Since the set of items forms an independent set, u_n must be a subject. Note that for any M-ordering $v_1, v_2, \cdots, v_{n-1}, v_n$ of a threshold graph G, $v_1, v_2, \cdots, v_{n-2}, v_n, v_{n-1}$ is another M-ordering of G. Thus, $N = u_1, u_2, \cdots, u_{n-2}, u_n, u_{n-1}$ is an M-ordering of G_p. It is easy to see that N satisfies $(*)$. A similar argument follows for $\{u_{n-1}, u_n\} \notin E(G_p)$ and u_{n-1}, a subject.

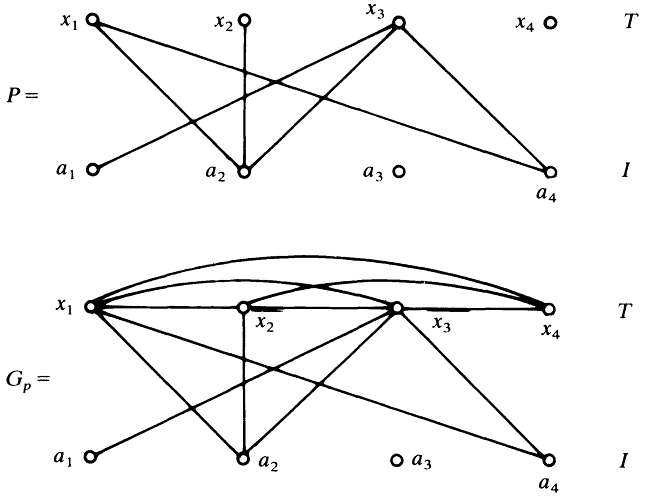
(b) We assume that $|T| \geqq 2$ and $|I| \geqq 2$, otherwise we must have a Guttman scale and it is trivial to find the ordering. It is sufficient to show that for any M-ordering $M = u_1, u_2, \cdots, u_n$ satisfying $(*)$, its associated M-sequence satisfies for $i \neq n$:

$$M_i = 0 \Leftrightarrow u_i \text{ is an item,}$$

$$M_i = 1 \Leftrightarrow u_i \text{ is a subject.}$$

The proof is by induction on i. The result is obvious for $i = 1$ since $|T| \geqq 2$ and $|I| \geqq 2$. Assume the result is true for all $j < i$. Show it for i. Let $i \leqq n-1$. Suppose u_i is a subject. If a subject u_k follows u_i in M, then $M_i = 1$ since $\{u_i, u_k\} \in E(G_p)$. If no subject follows u_i in M, then u_k is an item for all $k > i$. If $i < n-1$, then, in particular, u_{n-1} and u_n are items. Since $M_j = 1$ for all subjects j with $j < i$, then either $M_i = 1$ or $M_i = 0$ imply that u_{n-1} and u_n are agreed with by exactly the same subjects, which is a contradiction. Thus $i=n-1$ and so, by (*), $M_i=1$. The argument is similar if u_i is an item. Q.E.D.

We illustrate an example in Fig. 5. Note that no two subjects agree with exactly the same items and no two items are agreed with by exactly the same subjects. Also, $\{u_{n-1}, u_n\} \not\in E(G_p)$ and u_{n-1} is an item, where $u_{n-1}=a_2$ and $u_n=x_4$. G_p is a threshold graph and the given M-ordering is a Guttman scale.



$a_3, x_3, a_1, x_1, a_4, x_2, a_2, x_4$
0, 1, 0, 1, 0, 1, 0

FIG. 5

2. Threshold dimension.

2.1. General results. The original motivation for studying threshold graphs was determining if a single inequality would be satisfied by the characteristic vectors of independent subsets of the vertices of a graph and not satisfied by nonindependent subsets of the vertices of the graph. The obvious question to ask now is the following: what is the least number k of linear inequalities

$$\begin{aligned} &a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n \leqq t_1, \\ &a_{k1}x_1+a_{k2}x_2+\cdots+a_{kn}x_n \leqq t_n, \end{aligned} \tag{2}$$

such that S is independent if and only if its characteristic vector $x=(x_1, x_2, \cdots, x_n)$ satisfies (2). Analogous to the hyperplane dividing the hypercube into two halves, each inequality corresponds to one side of a hyperplane. Thus S is independent if and only if x lies on the "good" side of each of the k hyperplanes. The *threshold dimension*, $t(G)$, of a graph G is the least number k of linear inequalities (2) such that S is an independent subset of $V(G)$ if and only if its characteristic vector satisfies (2). Without loss of generality we can assume that all the numbers a_{ij} and t_i in (2) are nonnegative integers. If G is a threshold graph then $t(G) \leqq 1$. $t(G)=0$ if and only if $V(G)$ is an independent set (G consists of isolated vertices only). If H is a generated subgraph of G then any independent subset of H is an independent subset of G so $t(H) \leqq t(G)$.

Chvátal and Hammer [1973] show that the definition of threshold dimension can be stated in an equivalent way. For a graph $G=(V, E)$, a set of threshold graphs $G_i=(V, E_i)$ $i=1, 2, \cdots, t$, with $E=E_1 \cup E_2 \cup \cdots \cup E_t$ is called a *threshold cover* of G.

THEOREM 6. *Let G be a graph. $t(G)=k$ if and only if k is the least integer such that a threshold cover of size k exists.*

Other types of dimension of a graph are defined in terms of intersections of supergraphs. The following corollary relates the threshold dimension of a graph G to intersections of threshold supergraphs of the complement of G, denoted $\tilde{G}$.

COROLLARY 6.1. *Let G be a graph. $t(G)$ is the least integer k such that $\tilde{G}$ is the intersection of k threshold graphs.*

Since, if isolated vertices are added to a threshold graph, the graph remains a threshold graph, it suffices to look for graphs to form a threshold cover such that the union of the vertices of G_i equals V. Figure 6 shows two graphs and their corresponding threshold covers.

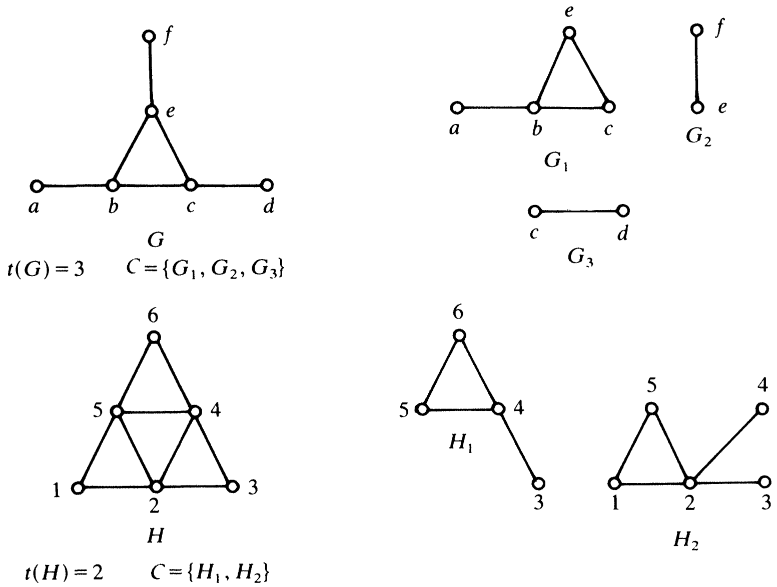


FIG. 6

Since $H=G$, Fig. 6 shows that in general the threshold dimension of a graph is not equal to the threshold dimension of its complement.

The next theorem about threshold dimension provides a method of computing the threshold dimension of some classes of graphs. Let $\alpha(G)$ denote the size of the largest independent set in G.

THEOREM 7 (Chvâtal and Hammer [1973]). *If G is a graph with n vertices, then $t(G) \leqq n-\alpha(G)$. Moreover, equality holds if G contains no triangle.*

COROLLARY 7.1. *For the following graphs we have*:

(i) $t(Z_n)=\lceil n/2 \rceil$ $(n>3)$[1],

(ii) $t(P_n)=\lceil n/2 \rceil$ *P_n is the path with n edges,*

(iii) $t(K_{m,n})=\min\{m, n\}$.

We also have the following for $K(m_1, m_2, \cdots, m_p)$ the complete p-partite graph with $m_1 \leqq m_2 \leqq \cdots \leqq m_p$.

THEOREM 8. $t(K(m_1, m_2, \cdots, m_{p-1}, m_p))=m_{p-1}$.

Proof. Since $K(m_{p-1}, m_p)$ is a generated subgraph of $K(m_1, m_2, \cdots, m_{p-1}, m_p)$ and $m_{p-1} \leqq m_p$, from Corollary 7.1 we have that $t(K(m_{p-1}, m_p))=m_{p-1} \leqq$

[1] $\lceil x \rceil$ = the least integer greater than or equal to x.

$t(K(m_1, m_2, \cdots, m_{p-1}, m_p))$. We will show the reverse inequality by covering $K(m_1, m_2, \cdots, m_{p-1}, m_p)$ with m_{p-1} threshold graphs and using Theorem 6.

Let M_i be the set of m_i independent vertices, and arbitrarily order these vertices as $a_{i1}, a_{i2}, \cdots, a_{im_{i-1}}, a_{im_i}$. Therefore $\{a_{ik}, a_{jl}\}$ is an edge if and only if $i \neq j$. Define subgraphs G_k of $K(m_1, m_2, \cdots, m_p)$ as follows: $V(G_k) = V(K(m_1, m_2, \cdots, m_p))$ and $E(G_k) = \{\{a_{ik}, a_{jl}\} | i < j\}$. Since $i < j \leqq m_p$ and $a_{ik} \in M_i$, $k \leqq m_{p-1}$. Thus we have defined m_{p-1} subgraphs. Clearly each edge is in one of these subgraphs, so $\{G_1, G_2, \cdots, G_{m_{p-1}}\}$ cover the edges of $K(m_1, m_2, \cdots, m_p)$. It remains to show that each G_k is a threshold graph.

Let e_1 and e_2 be edges of G_k. Then $e_1 = \{a_{ik}, a_{jl}\}$ and $e_2 = \{a_{i'k}, a_{j'l'}\}$ for $i < j$ and $i' < j'$ and some l and l'. We consider 2 cases.

Case 1. $i \neq i'$, say $i < i'$. Thus $i < i' < j'$. But now $\{a_{ik}, a_{i'k}\} \in E(G_k)$ and $\{a_{ik}, a_{j'l}\} \in E(G_k)$ and the subgraph generated, shown in Fig. 7, is not P_3, Z_4 or $2K_2$.

Case 2. $i = i'$. But now e_1 and e_2 share a vertex in common and cannot generate P_3, Z_4 or $2K_2$.

Therefore each G_k is a threshold graph and $C = \{G_1, G_2, \cdots, G_{m_{p-1}}\}$ is a threshold cover of G. Hence $t(G) \leqq m_{p-1}$. Therefore $t(G) = m_{p-1}$. Q.E.D.

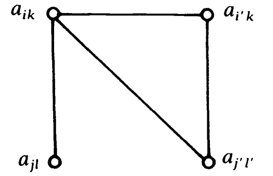


FIG. 7

Using the concept of threshold cover, it is possible to bound the threshold dimension of a graph from above. Let θ_e be the minimum number of cliques needed to cover the edges of a graph. Then $t(G) \leqq \theta_e(G)$, since every clique is a threshold graph.

2.2. Complexity and related results. As illustrated in the last proof, determining if a graph is a threshold graph or not reduces to determining if any 2 edges generate a P_3, Z_4 or $2K_2$. It is therefore reasonable to construct a new graph G^* from G as follows:

$$V(G^*) = \{\{x, y\} | \{x, y\} \in E(G)\} \quad \text{and}$$

$$\{\{x, y\}, \{u, v\}\} \in E(G^*) \Leftrightarrow \{x, y, u, v\} \text{ generates } P_3 \text{ or } Z_4 \text{ or } 2K_2.$$

If $\chi(G^*)$ denotes the chromatic number of G^*, then it is easy to see that $t(G) \geqq \chi(G^*)$. For a long time it was conjectured that indeed this might be an equality. If equality existed, even for $\chi(G^*) = 2$, then we would have a polynomial time algorithm for computing whether or not $t(G) \leqq 2$. Our first example shows that unfortunately $t(G) > \chi(G^*)$ for some G, and our second example shows that $t(G) - \chi(G^*)$ can be arbitrarily large.

Example 1 is shown in Fig. 8 and consists of 2 sets of vertices, $S_1 = \{a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3\}$ and $S_2 = \{A_1, A_2, A_3, B_1, B_2, B_3, C_1, C_2, C_3\}$. S_1 generates a complete subgraph K_9 and S_2 is an independent set. Call class A edges those edges connecting $\{A_1, A_2, A_3\}$ and $\{a_1, a_2, a_3\}$ and similarly define class B edges and class C edges. Ab edges are those edges connecting $\{A_1, A_2, A_3\}$ and $\{b_1, b_2, b_3\}$; Bc edges are those edges connecting $\{B_1, B_2, B_3\}$ and $\{c_1, c_2, c_3\}$ and Ca edges are those edges connecting $\{C_1, C_2, C_3\}$ and $\{a_1, a_2, a_3\}$.

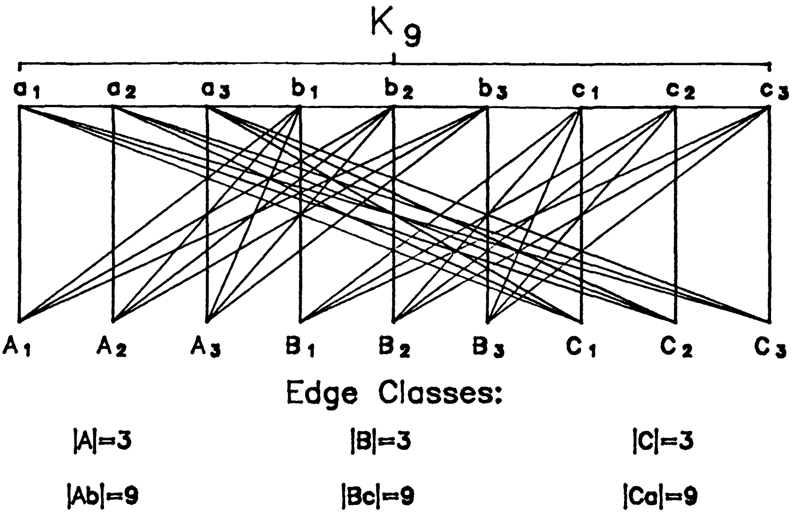

K_9
a_1 a_2 a_3 b_1 b_2 b_3 c_1 c_2 c_3
A_1 A_2 A_3 B_1 B_2 B_3 C_1 C_2 C_3
Edge Classes:
|A|=3
|B|=3
|C|=3
|Ab|=9
|Bc|=9
|Ca|=9

FIG. 8

LEMMA 2. *For the graph shown in Fig.* 8, $t(G) > \chi(G^*)$.

Proof. We will show that $t(G) \geqq 5$ and $\chi(G^*) = 4$, thus proving $t(G) > \chi(G^*)$. First, consider the edges of G. No two class A edges can be in the same threshold graph for $\{A_i, a_i, a_j, A_j\}$, $i \neq j$ generates a P_3 subgraph of G. Similarly, no two class B edges, and no two class C edges can be in the same threshold graph. A class A edge and a class Bc edge cannot be in the same threshold graph for they, too, would generate a P_3. Similarly, class B and Ca edges, and class C and Ab edges, may not be in the same threshold graph. No three vertical edges can be in the same threshold graph, for they would have to be in distinct classes, say edges $\{a_i, A_i\}$, $\{b_j, B_j\}$ and $\{c_k, C_k\}$. But now an Ab edge, a Bc edge, and a Ca edge must be present, otherwise P_3 is generated as a subgraph, contradicting the condition on pairs of edges. Therefore at least $\frac{9}{2}$ different threshold graphs are needed to cover the edges of G. Since $t(G)$ is integer valued, $t(G) \geqq 5$.

We now show that $\chi(G^*) = 4$. Figure 9 shows G^* with a 4-coloring. In the previous part we observed which pairs of edges could not be in the same threshold graph, corresponding to those edges which as vertices of G^* have an edge between them. Any two K_9 edges, any two Ab edges, any two Bc edges, and any two Ca edges can be in the same threshold graph, thereby corresponding to independent sets of vertices in G^*. A Bc edge and a Ca edge of G may be in the same threshold graph only when the C of the Ca edge has the same index as the c of the Bc edge. Similarly, the same is true for Ca edges and Ab edges, and for Ab edges and Bc edges. Thus the Bc, Ca, and Ab vertices generate a proper subgraph of $K(9, 9, 9)$ in G^*. As long as each set of Bc, Ca, and Ab vertices gets a different color, and the Bc color is different from the three A colors, and similarly for Ca and B and Ab and C, we have a proper 4-coloring of G^*. Therefore $\chi(G^*) \leqq 4$. Since K_4 is a generated subgraph of G^*, $\chi(G^*) = 4$. Therefore $t(G) > \chi(G^*)$. Q.E.D.

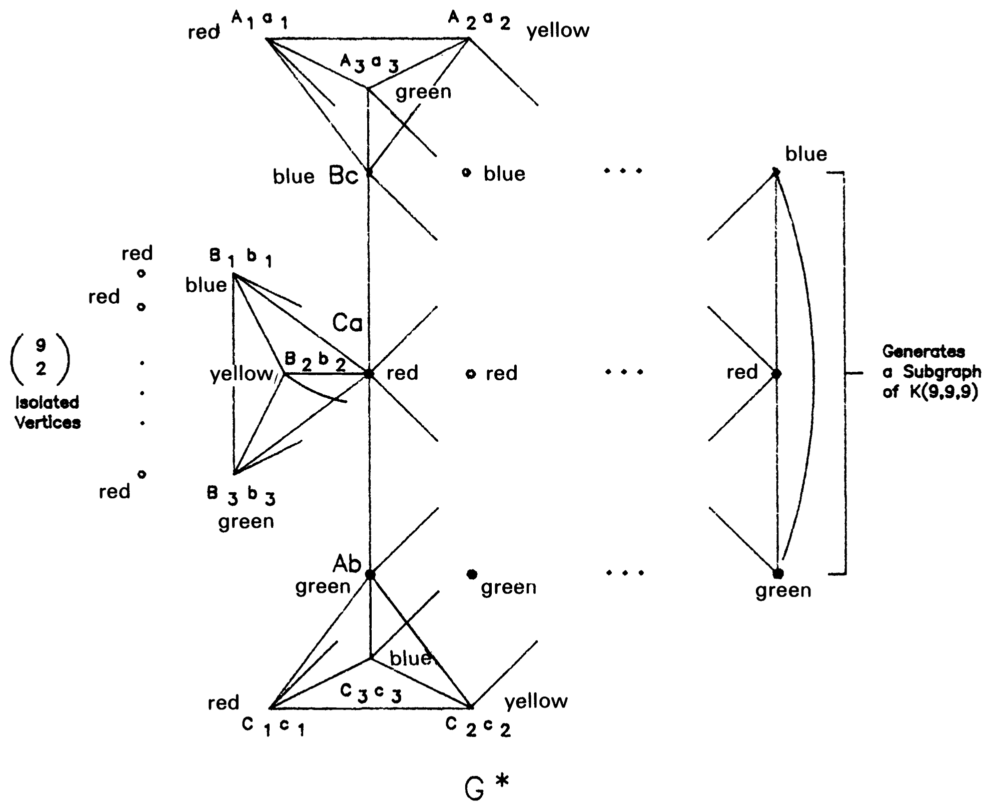


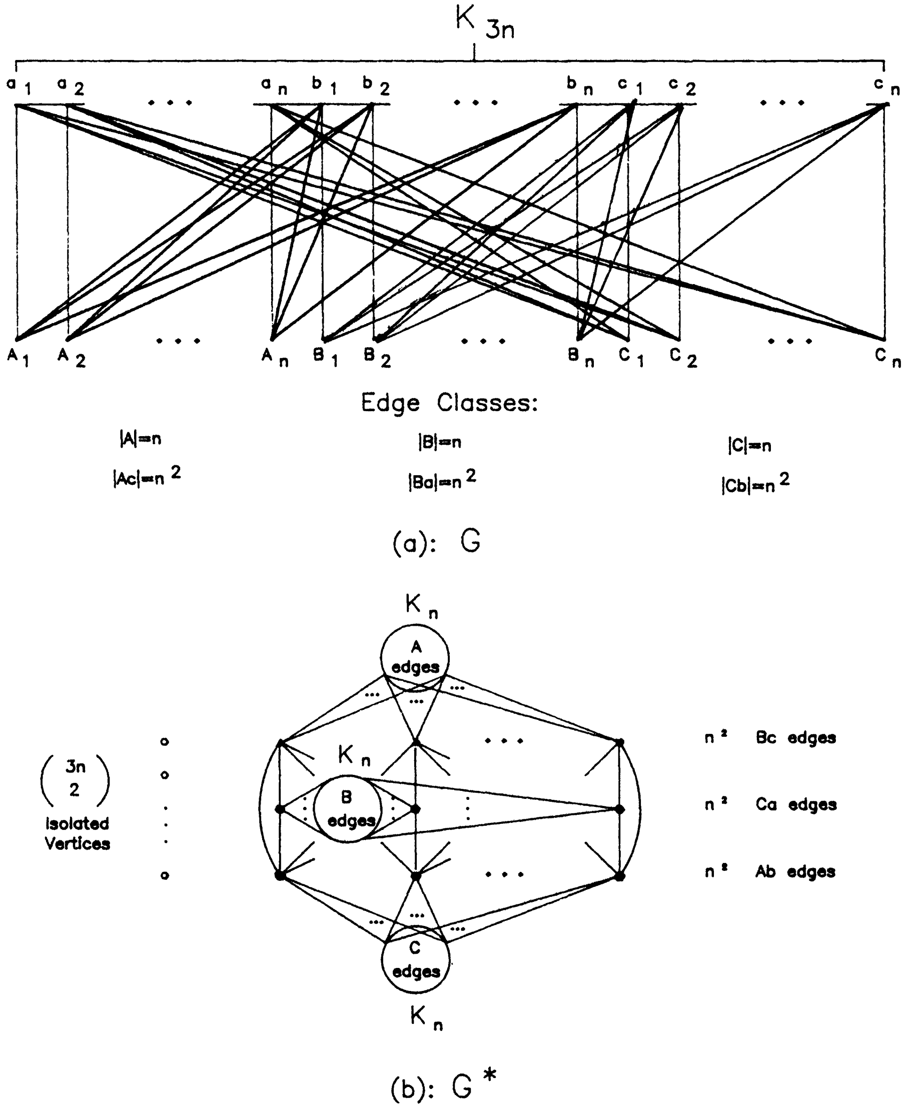
FIG. 9

We can now generalize the example shown in Fig. 8.

THEOREM 9. *There exist graphs G for which $t(G)-\chi(G^*)$ is arbitrarily large.*

Proof. The graph shown in Fig. 10a is a generalization of the one shown in Fig. 8, where instead of a_1, a_2, a_3 we have $a_1, a_2, \cdots, a_n$, instead of A_1, A_2, A_3 we have $A_1, A_2, \cdots, A_n$, and similarly for the b's, B's, c's, and C's. Hence we have the small lettered vertices generating K_{3n} and the large lettered vertices generating an independent set of $3n$ elements. There are now $3n$ vertical edges and $3n^2$ diagonal edges. By the same reasoning as in Lemma 2, no 3 vertical edges can be in the same threshold graph, so $t(G) \geqq 3n/2$. As shown in Fig. 10b, $\chi(G^*) = n+1$. Therefore $t(G)-\chi(G^*) \geqq n/2-1$, which can be arbitrarily large. Also, $t(G)-\chi(G^*)>0$, all $n \geqq 3$. Q.E.D.

The difficulties encountered in computing the threshold dimension of a graph are not surprising since Chvátal and Hammer [1977] pointed out that the problem of computing threshold dimension is NP-complete. Since computing $\alpha(G)$ is NP-complete for triangle-free graphs, Theorem 7 says computing $t(G)$ is NP-complete for triangle-free graphs. Yannakakis [1982] showed the stronger result that determining if $t(G) \leqq k$ is NP-complete for all fixed $k \geqq 3$. The case $k=2$ had remained open until now. Ibaraki and Peled [1982] gave sufficient conditions for the threshold dimension of a graph to be less than or equal to two, by showing that for split graphs G ($V(G)$ can be partitioned into a clique and an independent set) if $\chi(G^*)=2$ then $t(G)=2$; and for general G, if $\chi(G^*)=2$ and G^* has at most two nonsingleton components then $t(G)=2$. They conjectured that if $\chi(G^*)=2$, then $t(G)=2$. For general graphs G, this would provide


K_{3n}
a_1 a_2 . . . a_n b_1 b_2 . . . b_n c_1 c_2 . . . c_n
A_1 A_2 . . . A_n B_1 B_2 . . . B_n C_1 C_2 . . . C_n
Edge Classes:
$|A|=n$
$|Ac|=n^2$
$|B|=n$
$|Ba|=n^2$
$|C|=n$
$|Cb|=n^2$
K_n
A edges
K_n
B edges
C edges
K_n
$\binom{3n}{2}$ Isolated Vertices
n^2 Bc edges
n^2 Ca edges
n^2 Ab edges

(a): G

(b): G *

FIG. 10

a polynomial time algorithm for determining if $t(G)=2$. Cozzens and Leibowitz, in a forthcoming paper, show that this conjecture is false, by showing that determining if $t(G) \leqq 2$ is an NP-complete problem. The problem of determining if $t(G) \leqq 2$ is shown to be NP-complete by a transformation from the problem of partitioning the edge set of a graph into two triangle-free subgraphs.

In certain cases, the threshold dimension of a graph is related to other types of dimensions. A digraph $D=(\underline{X}, A)$ is called a *Ferrers digraph* when there exists a linear order $(\underline{X}, L)$ such that for every $x, y, z \in \underline{X}$, if $(x, y) \in L$ and $(y, z) \in A$ then $(x, z) \in A$. The *Ferrers dimension* of a digraph D is the smallest number of Ferrers digraphs whose intersection is G. Cogis [1982a] shows that the problem of determining the Ferrers dimension of a digraph is polynomially equivalent to finding the threshold dimension of a split graph. Cozzens and Leibowitz [to appear] discuss various relationships between the threshold dimension of particular classes of graphs and the dimension of certain corresponding partial orders. Since computing if either the Ferrers dimension or the partial order dimension is less than or equal to 2 is not NP-complete, yet determining if the threshold dimension is less than or equal to 2 is NP-complete, neither of these correspondences answer all of the questions regarding threshold dimension. Therefore it is necessary to develop more techniques to compute the threshold dimension for at least some classes of graphs, and to set bounds on the threshold dimension for all graphs.

2.3. Reduction of *G*. In this section we develop a test for threshold dimension 1 which will generalize to some results about threshold dimension k. In § 1.1 we defined a relation R on the vertices of a graph G as follows:

$$xRy \Leftrightarrow N(x)-\{y\}=N(y)-\{x\}.$$

Recall that R is an equivalence relation on $V(G)$. The reduction of G into its equivalence classes is denoted G^R; E_x is the equivalence class containing x. Formally define:

$$V(G^R)=\{E_x|x\in V(G)\},$$

$$E(G^R)=\{\{E_x, E_y\}|E_x\neq E_y, \{x, y\}\in E(G)\}.$$

Figure 11 shows a graph and its reduction.

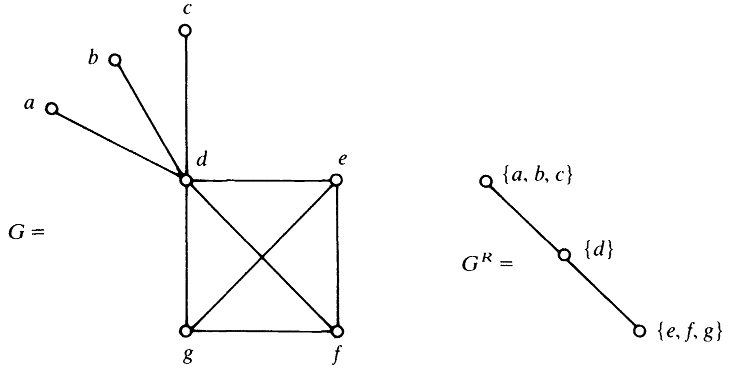


FIG. 11

LEMMA 3. *If G is a threshold graph, then G^R is a threshold graph.*

Proof. G^R is isomorphic to a generated subgraph of G. The result follows since a generated subgraph of a threshold graph is a threshold graph.

The converse of this lemma is not true. For example, Z_4^R is isomorphic to K_2 which is a threshold graph. Hence, some conditions must be satisfied by G^R in order to use thresholdness of G^R as a test of thresholdness of G. These conditions will be discussed later.

Let $u_1, u_2, \cdots, u_{n-1}, u_n$ be an M-ordering of $V(G)$. The associated M^**-sequence* is:

$$M_i^*=\begin{cases} M_i, & 1\leq i\leq n-1,\\ M_{n-1}, & i=n.\end{cases}$$

Hence we have defined a zero-one sequence of length n such that for $i<j$, $\{u_i, u_j\}\in E(G)$ if and only if $M_i^*=1$ and $\{u_i, u_j\}\notin E(G)$ if and only if $M_i^*=0$.

THEOREM 10. *Suppose G is a threshold graph. For all $x, y \in V(G)$, the following are equivalent*:

(i) $\deg(x) = \deg(y)$.

(ii) xRy.

(iii) *Given any M-ordering and the associated M^*-sequence, $M^*(x) = M^*(y) = M^*(z)$ for all z lying between x and y in the M-ordering.*

Proof. We show (i)$\Rightarrow$(ii)$\Rightarrow$(iii)$\Rightarrow$(i).

(i)$\Rightarrow$(ii) Suppose $\deg(x) = \deg(y)$ and suppose not xRy. Then there are a, b such that $a \in N(x) - N(y)$ and $b \in N(y) - N(x)$. Then the configuration shown in Fig. 2, is a subconfiguration of G.

(ii)$\Rightarrow$(iii) Suppose xRy and suppose there exists an M-ordering and associated M^*-sequence such that $M^*(x) \neq M^*(y)$, or there exists a $z(\neq x, y)$ between x and y such that $M^*(x) = M^*(y) \neq M^*(z)$. Without loss of generality, let x appear before y in the M-ordering. If $M^*(x) = 1$ and $M^*(y) = 0$, then $y = u_n$ and $\{x, y\} \in E(G)$. Now $x \neq u_{n-1}$ since $M^*_{n-1} = M^*_n = 0$, while $M^*(x) = 1$. But $u_{n-1} \in N(y) - N(x)$, a contradiction. If $1 = M^*(x) = M^*(y) \neq M^*(z) = 0$, then $z \in N(x) - N(y)$, a contradiction. Finally if $0 = M^*(x) = M^*(y) \neq M^*(z) = 1$, then $z \in N(y) - N(x)$, a contradiction.

(iii)$\Rightarrow$(i) We have two cases, $M^*(x) = M^*(y) = M^*(z) = 1$ for all z lying between x and y or $M^*(x) = M^*(y) = M^*(z) = 0$ for all z lying between x and y. Note that if $M^*(u) = 0$, $\deg(u)$ is the number of ones before u in the M^*-sequence; if $M^*(u) = 1$, $\deg(u)$ is the number of ones before u in the M^*-sequence plus the number of terms after u in the M^*-sequence. Using this observation, the result follows easily for both cases. Q.E.D.

COROLLARY 10.1. *If G is a threshold graph, then*:

(a) *Equivalence classes are either cliques or independent sets.*

(b) *Any M-ordering of G lists elements of equivalence classes consecutively.*

(c) *Any M-sequence of G^R is a sequence of alternating* 0's *and* 1's.

(d) *There is only one possible M-sequence of G^R, and that is either* $0, 1, 0, 1, \cdots$ or $1, 0, 1, 0, \cdots$.

THEOREM 11. *If G is a threshold graph, then G^R has exactly two M-orderings $E_1, E_2, \cdots, E_{n-1}, E_n$ and $E_1, E_2, \cdots, E_n, E_{n-1}$.*

Proof. Suppose there exist two M-orderings $E_1, E_2, \cdots, E_{n-1}, E_n$ and $F_1, F_2, \cdots, F_n$. By Corollary 10.1 (*d*), they have the same M-sequence, either $1, 0, 1, 0, \cdots$ or $0, 1, 0, 1, \cdots$. Therefore they have the same M^* sequence. Let i be the smallest integer such that $F_i \neq E_i$. Let $L_1, L_2, \cdots, L_n$ denote the M^*-sequence of the E's, while $M_1, M_2, \cdots, M_n$ denotes the M^*-sequence of the F's. Since $\{E_j: j = 1, 2, \cdots, i-1\} = \{F_j: j = 1, 2, \cdots, i-1\}$ and the sequence of M's is the same as the sequence of L's, E_i is equivalent to F_i in G^R. Suppose $F_i = E_{i+k}$, $k > 0$. Then $L^*_i = L^*_{i+1} = \cdots = L^*_{i+k}$ by Theorem 10. By Corollary 10.1, $i = n-1$ and $k = 1$, allowing only the two sequences stated in the theorem as possibilities. In fact, both are possibilities, for in any M-ordering of a threshold graph, the last two vertices are equivalent. Hence, G^R has exactly two M-orderings, those listed above. Q.E.D.

If G^R is a threshold graph, *unfolding* an M-ordering $E_1, E_2, \cdots, E_n$ of G^R gives a listing of the vertices of G, $x_{11}, x_{12}, \cdots, x_{1p_1}, x_{21}, \cdots, x_{2p_2}, \cdots, x_{n1}, \cdots, x_{np_n}$, such that $x_{ij} \in E_i$, $j = 1, 2, \cdots, p_i = |E_i|$. That is, we list the vertices by equivalence classes, with arbitrary ordering within each equivalence class. An M-ordering of G^R *admits* an M-ordering of G if by unfolding the M-ordering of G^R we (always)[2] get an

[2] If some unfolding of an M-ordering M of G^R is an M-ordering of G, then any unfolding of M is an M-ordering of G.

M-ordering of G. An M-ordering of G^R which admits an M-ordering of G is said to be *admissible.*

THEOREM 12. *If G is a threshold graph, every M-ordering of G is obtained from an M-ordering of G^R by unfolding.*

Proof. Let $M = v_1, v_2, \cdots$ be an M-ordering of G. By Corollary 10.1(b) equivalent vertices are listed consecutively. Then, we can order the equivalence classes as they appear in M. Call this order of equivalence classes L. It is easy to see that L is an M-ordering of G^R. We can unfold L to get back M. Q.E.D.

COROLLARY 12.1. *If G is a threshold graph, G^R has at least one admissible M-ordering.*

COROLLARY 12.2. *If G is a threshold graph, G^R has exactly one admissible M-ordering.*

Proof. Let $E_1, E_2, \cdots, E_{n-1}, E_n$ and $E_1, E_2, \cdots, E_n, E_{n-1}$ be the two M-orderings of G^R. By Corollary 10.1(a), E_{n-1} and E_n are both either a clique or an independent set. Since E_{n-1} and E_n are different, it follows that either E_{n-1} or E_n is a clique, while the other is an independent set, and $|E_{n-1}|+|E_n|>2$. Without loss of generality, say E_{n-1} is the clique. To show that exactly one of the M-orderings of G^R is admissible, we consider two cases.

Case 1. Suppose E_{n-1} and E_n are not adjacent in G^R.

(a) Suppose $|E_{n-1}|>1$. In the unfolding of $E_1, E_2, \cdots, E_{n-1}, E_n$, we get $x_{11}, x_{12}, \cdots, x_{1p_1}, \cdots, x_{n-1,1}, x_{n-1,2}, \cdots, x_{n,1}, \cdots$. In the M-ordering of G, $M^*(x_{n-1,1}) = 1$ since $\{x_{n-1,1}, x_{n-1,2}\} \in E(G)$. Hence, $\{x_{n-1,1}, x_{n,1}\} \in E(G)$, a contradiction. Hence, $E_1, E_2, \cdots, E_{n-1}, E_n$ is not admissible. However, by Corollary 12.1, $E_1, E_2, \cdots, E_n, E_{n-1}$ is admissible.

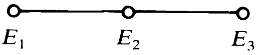
(b) Suppose $|E_n|>1$. Then $|E_{n-1}|>1$, otherwise $E_{n-1}UE_n$ would be an equivalence class, a contradiction. Thus, we are back in part (a).

Case 2. E_{n-1} and E_n are adjacent in G^R. The proof is similar to Case 1. Q.E.D.

COROLLARY 12.3. *If G is a threshold graph, it has a unique M-ordering of the vertices, up to equivalence, and hence a unique M^*-sequence (a unique M-sequence).*

Proof. This is obvious by Theorem 12 and Corollary 12.2. Q.E.D.

For an example to illustrate admissibility, see G in Fig. 4 and G^R in Fig. 12. The two M-orderings of G^R are E_2, E_1, E_3 and E_2, E_3, E_1 but by Case 1(a) of the proof of Corollary 12.2, only E_2, E_1, E_3 is admissible.



$E_1 = \{a, b\}$, $E_2 = \{c\}$, $E_3 = \{e, d\}$

FIG. 12

As we previously mentioned, it is possible for G^R to be a threshold graph, while G is not; for example $G = Z_4$. If we add a restriction on the M-sequence of G^R, the thresholdness of G^R will test the thresholdness of G.

THEOREM 13. *Suppose G^R is a threshold graph and $E_1, E_2, \cdots, E_n$ is an M-ordering of G^R. Suppose the associated M-sequence satisfies the following conditions for $i \leqq n-1$:*

(i) *E_i is not a clique implies $M_i = 0$;*

(ii) *E_i is not an independent set implies $M_i = 1$.*

Then G is a threshold graph.

Proof. We shall show that unfolding the M-ordering $M = E_1, E_2, \cdots, E_n$ gives an M-ordering N for G, and hence G is a threshold graph. Suppose x belongs to E_i.

If $|E_i|>1$, then there is $y\neq x$ in E_i. If E_i is a clique, then by (ii), $M_i=1$. Thus, for all $j>i$, E_i is adjacent to E_j. Hence, it is clear that x is adjacent to all vertices following it in N. Similarly, if E_i is an independent set, then by (i), $M_i=0$. Then it is clear that x is not adjacent to any vertex following it in N. Finally, suppose $E_i=\{x\}$. If $M_i=1$, then clearly x is adjacent to all vertices following it in N. If $M_i=0$, then clearly x is not adjacent to any vertex following it in N. Q.E.D.

Suppose each vertex of a graph G gets a label C, I, or B. Call this labelling L. G is *1^*-threshold via L* if there exists an M-ordering $v_1, v_2, \cdots, v_n$ of G and associated M-sequence satisfying the following conditions for $i\leqq n-1$:

(i) If v_i is labelled I, then $M_i=0$.

(ii) If v_i is labelled C, then $M_i=1$.

Note that G is a threshold graph if and only if G is 1^*-threshold via L, for some L. For if G is a threshold graph, simply define L from an M-ordering of G, satisfying conditions (i) and (ii) above.

By Corollary 10.1(a), equivalence classes under the relation R on $V(G)$ are either cliques or independent sets. Hence, we get a *canonical labelling Lc* of G^R:

if E_i is a singleton set, E_i gets label B,
if E_i is a clique with $|E_i|>1$, E_i gets label C, and
if E_i is an independent set with $|E_i|>1$, E_i gets label I.

COROLLARY 13.1. *G is a threshold graph if and only if G^R is 1^*-threshold via Lc.*

Proof. Sufficiency is shown by Theorem 13. Suppose G is a threshold graph. From an M-sequence in G, we can get an M-sequence in G^R by representing consecutive 1's and consecutive 0's by a single 1 and single 0, respectively. Vertices belonging to a clique E_i with $|E_i|>1$ must appear consecutively in an M-ordering of G and have 1's associated with them. That gives a 1 to those E_i in an M-ordering of G^R. Similarly, vertices belonging to an independent set E_i with $|E_i|>1$ have 0's associated with them, so the corresponding E_i gets 0 in an M-ordering of G^R. Q.E.D.

We now generalize threshold dimension 1 to threshold dimension k.

LEMMA 4. $t(G^R)\leqq t(G)$.

Proof. G^R is isomorphic to a generated subgraph of G. Q.E.D.

The inequality may be a strict inequality. $t(Z_4^R)=1$ while $t(Z_4)=2$. Analogous to the need for 1^*-thresholdness of G^R, we define k^*-thresholdness.

Suppose each vertex of a graph G gets a label C, I, or B. Call this labelling L. G is *k^*-threshold via L* if and only if there are subgraphs $G_1, G_2, \cdots, G_k$ of G which are 1^*-threshold via L and cover the edge set and vertex set of G.

Recall that Lc is the canonical labelling of a graph G^R. The smallest k such that G^R is k^*-threshold via Lc is called the *star-threshold dimension* of G^R, denoted $t^*(G^R)$. If it is impossible to cover the edge set and vertex set of G^R with subgraphs which are 1^*-threshold via Lc, we say $t^*(G^R)=\infty$. Figure 13 gives some graphs G with $t(G)$ and their reductions G^R with $t(G^R)$ and $t^*(G^R)$. The canonical labelling of each G^R is given. Note that if G is isomorphic to G^R, then each vertex of G^R is a singleton set and so gets the label B.

Even though $t(G^R)$ is not an upper bound of $t(G)$, $t^*(G^R)$ is an upper bound of $t(G)$, as shown in the next theorem.

THEOREM 14. $t(G)\leqq t^*(G^R)$.

Proof. If $t^*(G^R)=\infty$, the inequality is true since $t(G)\leqq|V(G)|$. Suppose $t^*(G^R)=k$. Therefore it is possible to cover the edge set of G^R by k subgraphs which are 1^*-threshold via Lc. Let $\bar{G}_1, \bar{G}_2, \cdots, \bar{G}_k$, all 1^*-threshold via Lc, cover the edge set and vertex set of G^R. Consider subgraph $\bar{G}_i$. It has an M-ordering of its vertices $E_{i1}, E_{i2}, \cdots, E_{il_i}$ and an associated M-sequence such that 1's in the M-sequence

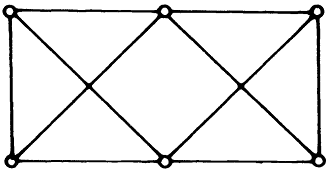
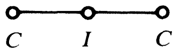
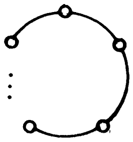
G	$t(G)$	G^R	$t(G^R)$	$t^*(G^R)$
$K(\underbrace{2,2,\ldots,2}_{n})$	2	K_n with each vertex labelled I	1	∞
	2		1	2
P_n	$\left\lceil \frac{n}{2} \right\rceil$	P_n	$\left\lceil \frac{n}{2} \right\rceil$	$\left\lceil \frac{n}{2} \right\rceil$
 $Z_n (n \geqq 4)$	$\left\lceil \frac{n}{2} \right\rceil$	Z_n	$\left\lceil \frac{n}{2} \right\rceil$	$\left\lceil \frac{n}{2} \right\rceil$

FIG. 13

correspond to cliques in the M-ordering, and 0's correspond to independent sets. Let $W(\bar{G}_i)$ be the subgraph of G formed from $\bar{G}_i$ as follows:

$$V(W(\bar{G}_i)) = \bigcup_{j=1}^{l_i} E_{ij}.$$

Suppose $x \in E_{ij}$, $y \in E_{ik}$. If $j \neq k$, then x is adjacent to y in $W(\bar{G}_i) \Leftrightarrow E_{ij}$ is adjacent to E_{ik} in $\bar{G}_i$. If $j = k$ and $x \neq y$, then: x is adjacent to y in $W(\bar{G}_i) \Leftrightarrow E_{ij}$ is a clique.

Now we observe that unfolding the M-ordering $M = E_{i1}, E_{i2}, \cdots, E_{il_i}$ of $\bar{G}_i$ gives an M-ordering N of $W(\bar{G}_i)$, which implies that $W(\bar{G}_i)$ is a threshold graph. The proof is the same as the proof of Theorem 13.

It remains to show that $\bigcup_{i=1}^{k} W(\bar{G}_i)$ covers the vertex set and edge set of G. Suppose v is a vertex of G which belongs to $E_l \in V(G^R)$. Since the $\bar{G}_i$'s cover the vertex set of G^R, $E_l \in V(\bar{G}_m)$ for some m. By definition of the vertex set of $W(\bar{G}_m)$, and since $v \in E_l$ and $E_l \in V(\bar{G}_m)$, we have that v is a vertex of $W(\bar{G}_m)$. Thus, the $W(\bar{G}_i)$'s cover the vertex set of G. Consider $\{v_i, v_j\} \in E(G)$. If both v_i and v_j belong to $E_l \in V(G^R)$, then $\{v_i, v_j\} \in E(G)$ implies E_l is a clique. Now $E_l \in V(\bar{G}_m)$ for some

m. By definition of $W(\bar{G}_m)$, $\{v_i, v_j\}$ is an edge of $W(\bar{G}_m)$. Next, suppose $v_i \in E_i$ and $v_j \in E_j$, $i \neq j$. Then $\{v_i, v_j\} \in E(G)$ implies $\{E_i, E_j\} \in E(G^R)$, which implies $\{E_i, E_j\} \in E(\bar{G}_m)$ for some m. By definition of $W(\bar{G}_m)$, $\{v_i, v_j\} \in E(W(\bar{G}_m))$. Thus, $\bigcup_{i=1}^{k} W(\bar{G}_i)$ covers the edge set of G, so $t(G) \leqq k = t^*(G^R)$. Q.E.D.

Lemma 4 and Theorem 14 give the important result that $t(G^R) \leqq t(G) \leqq t^*(G^R)$. If G is isomorphic to G^R, that is, if each E_i is a singleton set, then each threshold subgraph of G^R is 1*-threshold via Lc. Thus, the numbers $t(G^R)$ and $t^*(G^R)$ are equal, and $t(G)$ must also equal these numbers. We summarize these results in the following corollary.

COROLLARY 14.1.

(a) $t(G^R) \leqq t(G) \leqq t^*(G^R)$.

(b) *If G is isomorphic to G^R, then $t(G) = t(G^R) = t^*(G^R)$.*

(c) *If $t(G^R) = t^*(G^R)$, then $t(G) = t(G^R) = t^*(G^R)$.*

If $t^*(G^R) = \infty$, we do not get a good upper bound on $t(G)$ but we do get the following result.

THEOREM 15. *If $t^*(G^R) = \infty$, then Z_4 is a generated subgraph of G.*

Proof. An edge of G^R is 1*-threshold via Lc if and only if it has the property that at least one of the endpoints is a clique. If this property holds for all edges of G^R, then $t^*(G^R) \leqq |V(G^R)| + |E(G^R)| < \infty$. If $t^*(G^R) = \infty$, then G^R has an edge with both endpoints being independent sets of cardinality at least two. That is, Z_4 is a generated subgraph of G. Hence, $t(G) > 1$. Q.E.D.

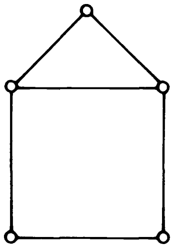
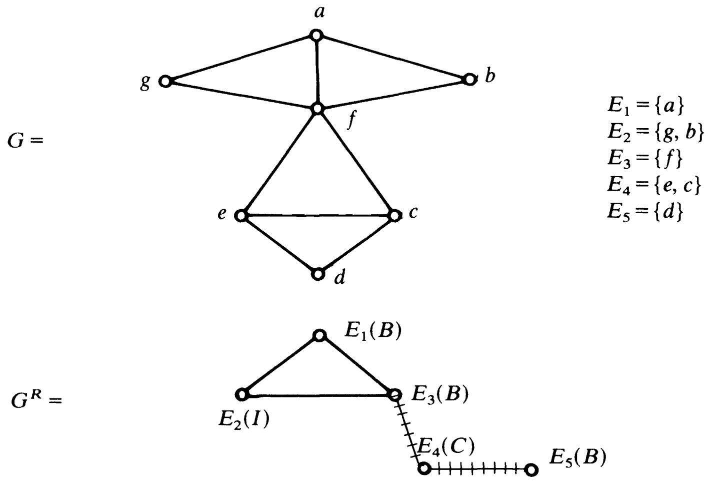


FIG. 14



$2 = t(G^R) \leqq t(G) \leqq t^*(G^R) = 2$

FIG. 15

The converse of Theorem 15 is not true. Z_4 is a generated subgraph of G shown in Fig. 14, while G^R is isomorphic to G. Hence $t(G)=t^*(G^R)=2$.

In Fig. 15, we have $t^*(G^R)=t(G^R)=2$ implying $t(G)=2$.

In this section we introduced new bounds on $t(G)$, the threshold dimension of a graph G, namely the lower bound $t(G^R)$ and the upper bound $t^*(G^R)$. In § 2.2, we discussed a lower bound $\chi(G^*)$ for $t(G)$. As shown by some examples, these bounds may not be good. This does not reduce their importance since determining if $t(G) \leqq k$ is NP-complete for all fixed $k \geqq 2$.

REFERENCES

C. Benzaken and P. L. Hammer [1978], *Linear separation of dominating sets in a graph*, Ann. Discrete Math, 3, pp. 1–10.

V. Chvátal and P. L. Hammer [1973], *Set packing and threshold graphs*, Univ. of Waterloo Res. Report, Corr 73-21.

——— [1977], *Aggregation of inequalities in integer programming*, Ann. Discr. Math., 1, pp. 145–162.

O. Cogis [1979], *Ferrers digraphs and threshold graphs*, Université Pierre et Marie Curie, Parix VI; Equipe graphes et optimisation combinatoire, Rapport de recherche No. 13, Fevrier, 1979.

——— [1982a], *Ferrers digraphs and threshold graphs*, Discrete Math., 38, pp. 33–46.

——— [1982b], *On the Ferrers dimension of a digraph*, Discrete Mathematics, 38, pp. 47–52.

C. H. Coombs [1964], *A Theory of Data*, John Wiley, New York.

M. Cozzens and R. Leibowitz, *Multidimensional scaling and threshold dimension*, to appear.

S. Földes and P. L. Hammer [1978], *On a class of matroid-producing graphs*, Colloq. Math. Soc. J. Bolyai (Combinatorics), 18, pp. 331–352.

M. Garey and D. Johnson [1979], *Computers and Intractability—A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco.

M. C. Golumbic [1978], *Threshold graphs and synchronizing parallel processes*, Colloq. Math. Soc. J. Bolyai (Combinatorics), 18, pp. 419–428.

——— [1980], *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York.

P. L. Hammer, T. Ibaraki and U. N. Peled, *Threshold numbers and threshold completions*, forthcoming in Ann. Discr. Math.

P. L. Hammer, T. Ibaraki and B. Simeone [1981], *Threshold sequences*, this Journal, 2, pp. 39–49.

P. B. Henderson and Y. Zalcstein [1977], *A graph-theoretic characterization of the PV chunk class of synchronizing primitives*, SIAM J. Comput., 6, pp. 88–108.

T. Ibaraki and U. N. Peled, *Sufficient conditions for graphs to have threshold number* 2, forthcoming in Ann. Discr. Math.

R. Leibowitz [1978], *Interval counts and threshold numbers of graphs*, PhD Thesis, Mathematics Dept., Rutgers Univ., New Brunswick, NJ.

J. Orlin [1977], *The minimal integral separator of a threshold graph*, Ann. Discr. Math., 1, pp. 415–419.

U. N. Peled [1977], *Matroidal graphs*, Discr. Math., 20, pp. 263–286.

——— [1980], *Threshold graph enumeration and series-product identities*, Proc. Eleventh Southeastern Conference on Combinatorics, Graph Theory, and Computing, pp. 735–738.

U. N. Peled and B. Simeone, *Box threshold graphs*, to appear.

S. Poljak [1974], *A note on stable sets of colorings of graphs*, Comm. Math. Univ. Carolinae, 15, pp. 307–309.

K. T. Rawlinson and R. C. Entringer [1979], *Class of graphs with restricted neighborhoods*, J. Graph Theory, 3, pp. 257–262.

H. Vantilborgh and A. Van Lamsweede [1972], *On an extension of Dijkstra's semaphore primitives*, Inform. Process. Lett., 1, pp. 181–186.

M. Yannakakis [1982], *The complexity of the partial order dimension problem*, this Journal, 3, pp. 351–358.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

013

INSTABILITY IN DISCRETE ALGORITHMS AND EXACT REVERSIBILITY*

GÉRARD Y. VICHNIAC†

Abstract. Some discrete algorithms (e.g., that derive from common discretization schemes for differential equations) are numerically exactly reversible: after any number of steps, the initial conditions can be recovered exactly, despite numerical roundoff. These algorithms constitute discrete dynamical systems on the numerical mesh that are deterministic in both directions of time. Because of the conservation of the information encoded in the initial conditions, and of the incompressibility of this information on the numerical mesh, these algorithms lead to instability in the presence of an attractor. This phenomenon is illustrated in examples involving several types of attractors and is compared with other mechanisms recently proposed for the explanation of instabilities in nonlinear finite-difference equations.

1. Introduction. Numerical errors resulting from roundoff in finite-precision arithmetic are often referred to as numerical "noise". Indeed, statistical analysis is a most natural tool for the study of the propagation of these errors, which are then treated as a stochastic process. Moreover, the apparently irreversible loss of the information contained in the digits discarded leads one to view roundoff as an irreversible (albeit often controllable) contamination of noise. But we know, of course, that there is no such thing as a genuine numerical "noise". In a digital computer, in particular, nothing happens at random. The roundoff process is a deterministic one, it obeys a specified rule. If genuinely random noise were present, repeated runs of the same program and data would yield different results, rendering any attempt to "debug" computer codes impossible.

It is then appropriate to view a computation on a digital computer as a dynamical system on the discrete set formed by the numerical mesh for the *dependent* variables. The spacing Δ for this mesh is usually much smaller than that for the independent variables, e.g., the time step h.

2. Reversible algorithms. It has been pointed out by Edward Fredkin [2] that roundoff in fixed point arithmetic does not necessarily entail a loss of information, i.e., some numerical schemes for initial-value problems are exactly reversible. These algorithms are then dynamical systems that are deterministic in both the forward and the backward directions of time. It turns out that some of these algorithms are in common use in the numerical treatment of initial-value ordinary differential equations.

Consider, for example, Newton's second law for a point mass m in a one-dimensional force field $F(x)$

$$m\frac{d^2x(t)}{dt^2}=F(x). \tag{2.1}$$

Replacing the seçond derivative by the simplest three-point formula, (2.1) is approximated by the finite-difference equation

$$m\frac{x_{n+1}-2x_n+x_{n-1}}{h^2}=F(x_n). \tag{2.2}$$

* Received by the editors August 10, 1983, and in final form March 5, 1984. This work was presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27-29, 1983. This work was supported in major part by an IBM postdoctoral fellowship, and also by grants from DARPA (nos. N00114-75-C-0661 and N0014-83-K-025), from the National Science Foundation (no. 8214312-IST), and from the Department of Energy (no. DE-AC02-83ER13082).

† Massachusetts Institute of Technology, Laboratory for Computer Science, Cambridge, Massachusetts 02139.

Solving for x_{n+1}, we have the algorithm

$$x_{n+1}=\left\{h^2\frac{F(x_n)}{m}+2x_n\right\}-x_{n-1}, \tag{2.3}$$

where the braces indicate that the operations are performed in finite-precision arithmetic. Notice, however, that the braces do not enclose the last term in the right-hand side of (2.3). This is because in fixed-point arithmetic, subtraction is exact (as in integer arithmetic), provided that no overflow occurs.

If we now specify the force field $F(x_n)$ and the initial conditions x_0 and x_1 on the numerical mesh, we can iterate (2.3) for a arbitrary number of steps and obtain a sequence of numbers $x_0, x_1, \cdots, x_{N-1}, x_N$. We might ask, given the obtained values x_{N-1} and x_N, can we recover the exact values of the initial conditions x_0 and x_1, for any $F(x_n)$, for an arbitrarily severe roundoff, and an arbitrarily large number of steps, despite the apparently irreversible loss of information that occurred at every step of the construction of the sequence? Surprisingly enough, the answer to this question is affirmative. The entire sequence can be reconstructed backward from x_N and x_{N-1}, simply by solving (2.3) for x_{n-1}. Indeed, since the roundoff is a deterministic process, the expression in the braces in (2.3) is evaluated in the same way going upward or backward in the sequence.

3. Instability in the presence of a fixed point. Fredkin [2] also noticed that numerical reversibility entails remarkable phenomena when a fixed point occurs in exact arithmetic. Consider for example

$$x_{n+1}=kx_n, \qquad 0<k<1. \tag{3.1}$$

Unlike (2.3), this scheme is numerically reversible only in the limit of infinite-precision arithmetic, where it has the obvious solution

$$x_n=k^nx_0 \tag{3.2}$$

and admits zero as an attractive fixed point.

Consider now the second-order finite-difference equation

$$x_{n+1}=\left(k+\frac{1}{k}\right)x_n-x_{n-1}. \tag{3.3}$$

With initial conditions obeying

$$x_1=kx_0, \tag{3.4}$$

(3.3) is equivalent to (3.1). This is readily derived by adding to (3.1) its own shifted equation

$$x_{n-1}=\frac{1}{k}x_n.$$

Though equivalent in exact arithmetic, equations (3.1) and (3.3) have dramatically different behaviors in finite-precision arithmetic. With a starting x_0 large compared to the mesh size Δ, the sequence obtained from (3.1) first decreases and eventually settles at a constant value (that can be finite or null, depending on whether or not $k>\frac{1}{2}$). In any case, the "attractive" nature of zero is always qualitively reproduced by (3.1), even when severe roundoffs occur. This is not the case for sequences obtained in fixed point arithmetic from (3.3). We shall write this equation in the form

$$x_{n+1}=\left\{\left(k+\frac{1}{k}\right)x_n\right\}-x_{n-1} \tag{3.5}$$

to emphasize that the right-hand side is evaluated with roundoff. Just like (2.3), the algorithm (3.5) is exactly reversible despite roundoff, again because it can be solved exactly for x_{n-1}.

We shall see how this reversibility, together with the discreteness of the numerical mesh, are responsible for the following interesting behavior of (3.5). Starting with initial conditions large compared with Δ and that satisfy (3.4), we first obtain values in good agreement with the exact solution (3.2). But when the sequence approaches zero it either curves back and increases to $+\infty$, or, depending on x_0, it plunges to $-\infty$. (For small values of x_0, the sequence may also oscillate around zero.) Now, the exact numerical reversibility of (3.5) means that trajectories in the (x, t) plane cannot merge, in fact any couple of points (x_n, x_{n+1}) completely defines a whole trajectory. Therefore, the sequence generated by (3.5) cannot settle at a constant value because it would then "forget" its initial conditions. Also, since a neighborhood of given width around zero contains a finite number of numerical mesh points, it cannot accommodate an arbitrarily large number of distinct trajectories in the (x, t) plane. This description [2] in terms of the conservation of the information encoded in the initial conditions and of the impressibility of this information on the numerical mesh characterizes simply the instabilities of the algorithm (3.5). It will be helpful for the following discussion to interpret this instability in more familiar terms of numerical analysis.

Let us notice first that the finite-difference equation (3.2), being linear, has a general solution in closed form (see, e.g., [3]):

$$x_n=C_-z_-^n+C_+z_+^n, \tag{3.6}$$

where C_- and C_+ are constants depending on the initial conditions. The numbers z_- and z_+ are the roots of the quadratic equation

$$z^2-2Az+1=0, \tag{3.7}$$

where $A=\frac{1}{2}(k+1/k)$, i.e.,

$$z_-=A-\sqrt{A^2-1}, \qquad z_+=A+\sqrt{A^2-1}. \tag{3.8}$$

Notice that for distinct roots, $A>1$, therefore $z_-<1$ and $z_+>1$, and hence (3.6) is the sum of a decaying sequence and a growing sequence. The second-order equation (3.3) has two linearly independent solutions, whereas our original first-order equation (3.1) has the purely decaying solution. The spurious growing solution introduced with (3.3) is suppressed when the proper initial conditions are used; specifically, relation (3.4) cancels the coefficient C_+. But if at some step i the relation

$$x_{i+1}=kx_i$$

ceases to be exactly verified because of roundoff, the coefficient C_+ is then "switched on" and the spurious growing solution eventually swamps the exact decaying solution.

Notice finally that since $z_-z_+=1$, (3.6) can be written in the form

$$x_n=C_-z_-^n+C_+z_-^{-n}. \tag{3.9}$$

This interpretation of the instabilities of (3.5) in terms of an admixture of the *time-reversed* solution will be useful in the following.

4. Instabilities around a limit cycle. We saw in the last section that numerically reversible algorithms are unstable if a fixed point occurs in exact arithmetic. Fredkin's argument holds in fact to any type of attractor, and we shall study in this section the effects of the conservation of information when a *limit cycle* is present. It turns out that the reasoning based on the admixture of the time-reversed solution keeps its predictive power in that case too.

Consider in polar coordinates the system

$$\frac{dr}{dt}=r(1-r^2), \tag{4.1a}$$

$$\frac{d\theta}{dt}=-1. \tag{4.1b}$$

These two equations are uncoupled. Equation (4.1a) has a stable fixed point at $r=1$, whereas $r=0$ and $r=\infty$ are unstable fixed points. Equation (4.1b), whose solution is $\theta(t)=-t+\theta(0)$ endows a geometrical meaning to time, and turns the stable fixed point of (4.1a) into a limit cycle given by the unit circle. Under (4.1), points inside this circle will "spiral out" toward it at a constant angular velocity in the *clockwise* directions. Similarly, points outside the limit cycle will spiral in at the same clockwise angular velocity. Since the equations (4.1) carry points out of the origin and out of infinity, these loci are said to constitute the *outset* whereas the limit cycle forms the *inset.*

Using Cartesian coordinates ($x=r\cos x$, $y=r\sin x$), equations (4.1) can be written in the form of two coupled equations

$$\frac{dx}{dt}=y-x(1-x^2-y^2), \tag{4.2a}$$

$$\frac{dy}{dt}=-x+y(1-x^2-y^2). \tag{4.2b}$$

We shall now compare the behaviors of irreversible and reversible numerical schemes corresponding to these equations. First, we replace the derivatives by the simple two-point forward difference, i.e., for (4.2a)

$$\frac{dx}{dt}\leftarrow\frac{x_{n+1}-x_n}{h}. \tag{4.3}$$

By confining the dependent variables x and y to the nodes of a mesh of spacing size Δ, (taken, say, to be the inverse of some integer N), we obtain the numerically irreversible algorithm

$$x_{n+1}=\lfloor hN(y_n+x_n(1-x_n^2-y_n^2))+\tfrac{1}{2}\rfloor/N+x_n, \tag{4.4a}$$

$$y_{n+1}=\lfloor hN(-x_n+y_n(1-x_n^2-y_n^2))+\tfrac{1}{2}\rfloor/N+y_n, \tag{4.4b}$$

where $\lfloor a\rfloor$ stands for the largest integer that is less than or equal to a.

If instead of using (4.3) we now approximate the derivatives by central differences, e.g.,

$$\frac{dx}{dt}\leftarrow\frac{x_{n+1}-x_{n-1}}{2h}, \tag{4.5}$$

we obtain a conspicuously reversible scheme

$$x_{n+1}=\lfloor 2hN(y_n+x_n(1-x_n^2-y_n^2))+\tfrac{1}{2}\rfloor/N+x_{n-1}, \tag{4.6a}$$

$$y_{n+1}=\lfloor 2hN(-x_n+y_n(1-x_n^2-y_n^2))+\tfrac{1}{2}\rfloor/N+y_{n-1}. \tag{4.6b}$$

By keeping an explicit control of the roundoff in (4.4) and (4.6) we stress again that roundoffs are not an expression of the free will of the computer. Actually, since the finite precision is already included in the equations, actual simulations of these algorithms will, for Δ not too small, yield identical results when performed in double precision on a mainframe computer, or on a pocket calculator, or even on a cash register if $\Delta = 10^{-2}$ (and if (4.4) and (4.6) are modified in order to include explicit roundoff of each intermediate operation). In other words, simulations of (4.4) and (4.6) will be exact implementations of the mathematics contained in these equations.

Figures 1 and 2 show the result of the first 400 iterations of (4.4) and 200 iterations of (4.6), respectively, with $\Delta = 10^{-3}$ and $h = 0.05$ starting with $x_0 = 0$, $y_0 = 0.1$. We used (4.4) to get values for the x_1 and y_1, required to initialize (4.6). During the very first steps both curves behave as expected from (4.1): the variable point moves smoothly toward the unit circle in the clockwise direction. In the neighborhood of the limit cycle, however, their behavior differ sharply. While (4.4) gently settles on the limit cycle, and thus "forgets" the initial conditions in a irreversible way, the numerically reversible scheme (4.6) shows instabilities. Here again, conservation and incompressibility of the information forbid the system evolving by (4.6) to settle around the limit cycle. Different initial points lead to different forms of instability, hence the oscillations can be seen as the signature of the initial conditions. Also, since a second order finite-difference equation simulate a first order differential equation, the instability can be accounted for by an admixture of a spurious solution. Figure 2 suggests that the spurious solution is related to *time-reversed* solution of (4.1), just like in the linear case (cf. (3.9)). Indeed, the system displays a *counterclockwise* oscillatory rotation

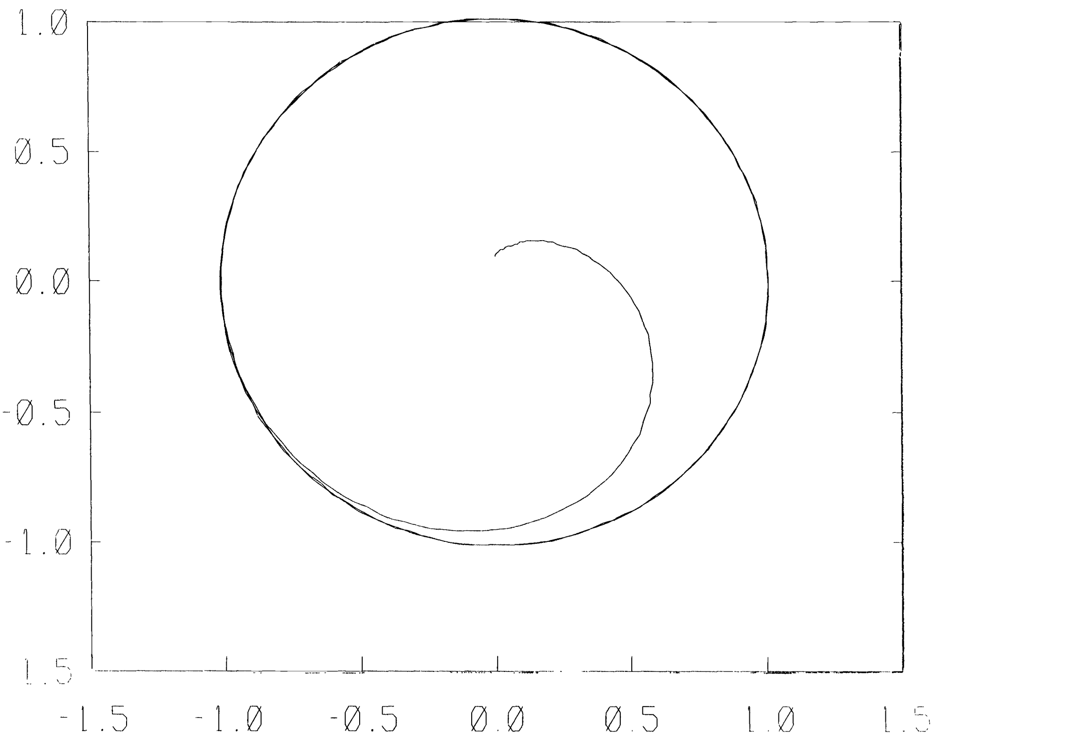


FIG. 1. *The first* 200 *iterations of the irreversible scheme* (4.4), *with parameters defined in the text.*

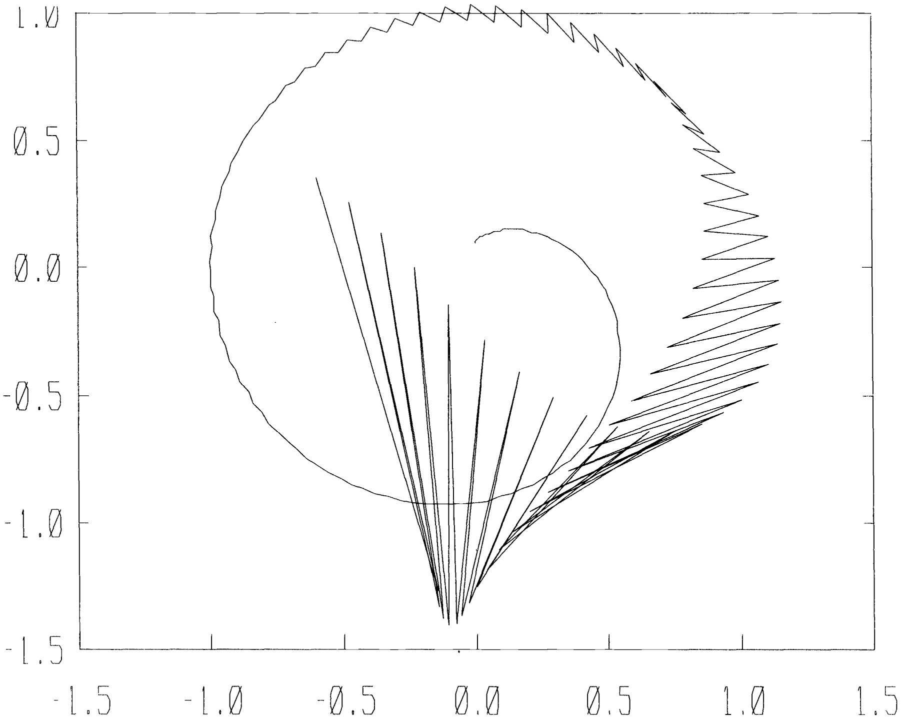


FIG. 2. *The first* 400 *iterations of the reversible scheme* (4.6).

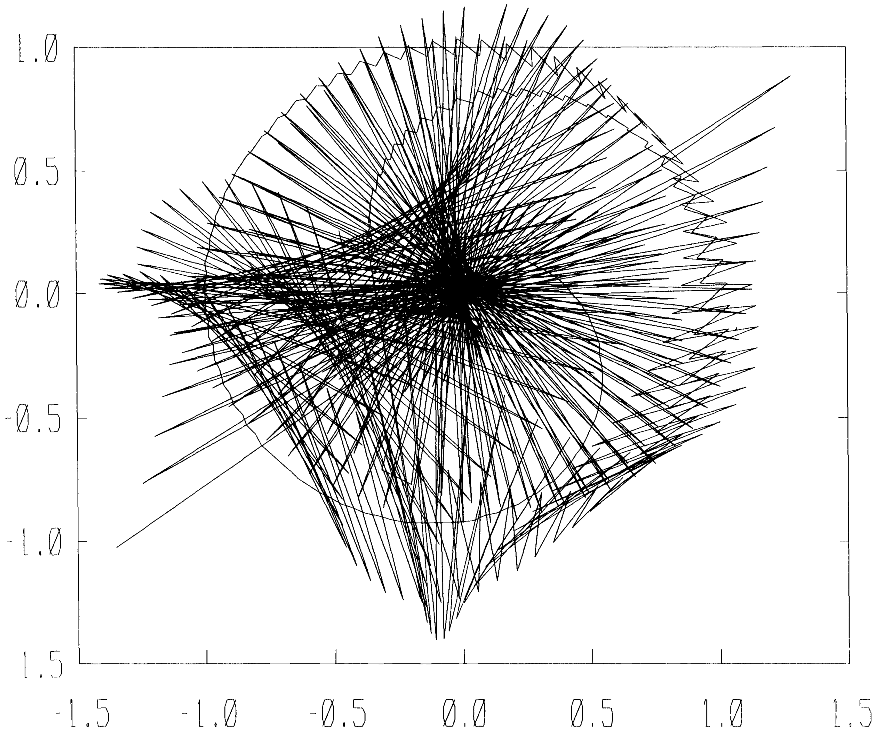


FIG. 3. *The first* 566 *iterations of the reversible scheme* (4.6).

around the normal *clockwise* revolution. Moreover, for the time-reversed solution, the nature of the inset and the outset is exactly reversed, and this has conspicuous consequences when this solution dominates. This happens twice during the first 566 iterations of (4.6), shown in Fig. 3. The system penetrates the circle of radius 0.1 at time steps 344, 346, and 348, displaying the attractive nature of the origin. The time-reversed component also clearly dominates beyond iterations 560, as it carries the whole system this time to the point at ∞. (The distance of the variable point from the origin, while still less than 1.5 at iteration 562, exceeds 14000 at time 571.) The time-forward solution dominates during the early stages of the evolution (see Fig. 2), bringing the system to the unit circle and keeping it there (within a distance of 0.1) between steps 61 and 153.

5. Discussion. Exact numerical reversibility yields instability in the presence of an attractor. This instability can also be accounted for by an admixture of the time-reversed solution. Other models for the onset of instability have recently been proposed [1], [4]. However, these models are based on the nonlinearity of the equations, whereas our description applies to the linear case as well.

Acknowledgments. This work stems from an idea of Edward Fredkin. I am thankful to him and also to Tommaso Toffoli and Norman Margolus for enlightening discussions, to Stephen Altschul for assistance in the study of the limit cycle, and to George Carrette for help in the preparation of the figures.

REFERENCES

[1] W. L. Briggs, A. C. Newell and T. Sarie, *Focusing: a mechanism for instability on nonlinear finite difference equations*, J. Comp. Phys., 51 (1983), pp. 83–106.

[2] E. Fredkin, private communication.

[3] P. Henrici, *Elements of Numerical Analysis*, John Wiley, New York, 1964.

[4] M. Yamaguti and S. Ushiki, *Chaos in numerical analysis of differential equations*, Physica, 3D (1981), pp. 618–626.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

014

HOW LARGE ARE TRANSITIVE SIMPLE MAJORITY DOMAINS?*

JAMES M. ABELLO† AND CHARLES R. JOHNSON‡

Abstract. It is well known that pairwise simple majority voting (smv) over n alternatives can lead to a nontransitive social outcome if the domain of possible (transitive) individual·preference orderings is unrestricted. A variety of domain restrictions which ensure transitivity have been offered (the best known being single-peaked preferences) but all of course, amount to some limitation of individual preferences. Call such restricted domains transitive smv domains (tsmv) and consider the following question: how many preference orderings may transitive smv domains contain and how much range of choice is consistent with ensured transitivity? This presents a combinatorial problem of an unusual sort; a lower bound of 2^{n-1} is easily proven but good upper bounds are generally harder to obtain. By using a graph representation of S_n (the symmetric group on n symbols) we have been able to construct transitive smv domains of cardinality $3\times 2^{n-2}-4$ for $n>4$ which to our knowledge is the best known general lower bound. The constructed sets have a relatively simple structure and the methods used here may be a valuable tool in the search for a maximum tsmv domain and in the general study of domains upon which simple majority vote is transitive.

Key words. Arrow's theorem, consistent sets, maximal chain, poset, simple majority rule, symmetric group, transitivity

Introduction. Transitive simple majority domains. Because of the classical anomaly which shows that simple majority voting produces a social relation which is not necessarily transitive and because of Arrow's theorem, there has been a natural thrust of research in the direction of determining domain restrictions under which majority vote avoids the intransitivity flaw. Two particular modes of addressing this issue are (1) sufficient conditions for domains over which unrestricted choice necessarily leads to transitivity (e.g. "single peaked" preferences [2]), and more generally, (2) characterization of those distributions of voters' profiles which happen to produce a transitive result (e.g. Inada [5]). Our goal is to determine how "big" transitive simple majority voting domains may be.

We restrict our attention to strict preference orderings (consistent sets), but it is clear that many of the ideas are extendable to nonstrict preferences either as they stand or with modification.

1. Problem formulation. Let $\langle \Sigma, \leqq \rangle$ be a totally ordered set of symbols of cardinality $|\Sigma| = n \in Z^+$, and S_Σ the set of permutations on Σ.

DEFINITION 1.1. A set $B \subset S_\Sigma$ is called *cyclic* if there are three symbols x_{i_1}, x_{i_2}, $x_{i_3} \in \Sigma$ and three permutations in B which, when restricted to the three symbols, are

$$\left\{\begin{matrix} x_{i_1} & x_{i_2} & x_{i_3} \\ x_{i_3} & x_{i_1} & x_{i_2} \\ x_{i_2} & x_{i_3} & x_{i_1} \end{matrix}\right\}.$$

DEFINITION 1.2.

i. Let $u, v, w \in S_\Sigma$; if the set $\{u, v, w\}$ is not cyclic it is called a *consistent* three-set.

ii. A subset C of S_Σ for which every three-subset is consistent is called a *consistent subset* of S_Σ.

[*] Received by the editors June 30, 1983. Portions of this work were presented at the SIAM Second Conference on the Applications of Discrete Mathematics, held at Massachusetts Institute of Technology, Cambridge, Massachusetts, June 27–29, 1983.

[†] Mathematics Department, University of California, San Diego, La Jolla, California 92093.

[‡] Institute for Physical Sciences and Technology, University of Maryland, College Park, Maryland 20742.

Example 1.1. Let $\Sigma=\{1,2,3,4\}$.

- The set $B=\{1234, 3124, 4231, 4321\}$ is cyclic because the permutations 1234, 3124 and 4231 when restricted to the symbols $\{1,2,3\}$ are

$$\left\{\begin{matrix}123\\312\\231\end{matrix}\right\}.$$

- The set $C=\{1234, 4123, 4321, 4312\}$ is consistent because every three-subset is consistent. Notice that this requires checking the consistency of the $\binom{|C|}{3}$[1] three-subsets of C. Moreover for each three-subset it is necessary to check each of the $\binom{|\Sigma|}{3}$ triples of symbols of Σ for the noncyclicity condition. It should be clear at this point that this task is computationally expensive even for moderately large values of $n=|\Sigma|$.

Comments. It is important to note that it is very easy for a set to be cyclic because cyclicity is a concept which depends only on a subset. On the other hand, it is a very well-known fact that the consistent sets are those over which unrestricted choice necessarily produces transitive relations under simple majority vote [4].

Problem formulation. The problem that we are interested in is how large a set may be and still necessarily lead to transitive simple majority vote results. More precisely the problem we propose to study here is:

Find the cardinality and study the structure of a *maximum consistent subset of* S_Σ.

Since the unfortunate aspect of domain restrictions is that a sufficiently rich realm of choice may not remain, it is natural to ask how much freedom of choice is consistent with transitivity, and we take the size (in terms of number of orderings) of the domain as a rough measure of the degree of choice.

Observation (Johnson [6]). When the number of alternatives is n, consistent sets containing 2^{n-1} distinct preference orderings can be constructed by the following inductive technique. To the set constructed for $(n-1)$ add alternative n at the left and right of each preference ordering. The process may be begun for $n=2$ with the singleton 1, and then it is clear that 2^{n-1} orderings result for general n. For $n=2$, 3, 4 these sets are shown below to see how they evolve:

$$n=2 \qquad \{12, 21\}$$

$$n=3 \qquad \left\{\begin{matrix}123, 213\\312, 321\end{matrix}\right\}$$

$$n=4 \qquad \left\{\begin{matrix}1234, 2134\\4123, 4213\\3124, 3214\\4312, 4321\end{matrix}\right\}.$$

To see that these sets are consistent, we note that for the triple $i<j<k$, neither of the subpermutations (j, k, i) or (i, k, j) can occur in any ordering because of the construction. Thus the definitional requirement for cyclicity cannot be satisfied by either

$$x_{i_1}=i,\ x_{i_2}=j,\ x_{i_3}=k \quad \text{or} \quad x_{i_1}=k,\ x_{i_2}=j,\ x_{i_3}=i$$

for any triple (i, j, k), so that the set is consistent.

[1] $\binom{n}{k}$ denotes the binomial coefficient.

A very different general class of consistent sets (EXP (P)) of cardinality 2^{n-1} have been constructed by Abello [1]. These sets are not only consistent, they are also maximal with respect to the noncyclicity property and they play a central role in this study.

The remainder of this paper is organized as follows. The next section presents basic general properties of consistent sets. In § 3 we present mainly without proofs the main properties of EXP (P) which are fully discussed in [1]. Section 4 contains a general theorem about maximal consistent sets which is the key to construct consistent sets of cardinality greater than 2^{n-1} whose existence was questioned since 1978 [6].

2. General properties of consistent sets.

DEFINITION 2.1. Let $A \subseteq S_\Sigma$ and let Σ' be any set of symbols such that $\Sigma' \cap \Sigma = \emptyset$. For any $w' \in S_{\Sigma'}$, $w'A \equiv \{w \in S_{\Sigma \cup \Sigma'}\colon w = w'v$ for some $v \in A\}$. Here, $w'v$ denotes the concatenation of w' and v. Aw' can be defined in a similar way to $w'A$.

The following are some elementary properties of consistent sets:

FACT 2.1.

i) *Any subset of a consistent set is consistent.*

ii) *Any superset of a cyclic set is cyclic.*

iii) *The intersection of consistent sets is a consistent set but their union is not always consistent.*

iv) *If* $C \subset S_\Sigma$, $\Sigma' \cap \Sigma = \emptyset$ *and* $w \in S_{\Sigma'}$ *we have:* wC *and* Cw *are consistent subsets of* $S_{\Sigma \cup \Sigma'}$ *iff* C *is a consistent subset of* S_Σ.

DEFINITION 2.2. i) If $p \in S_\Sigma$, by $T(p)$ we will denote the set of ordered triples of symbols of Σ determined by p and by $\mathcal{P}(p)$ we will denote the set of ordered pairs determined by p. (Notice that $\mathcal{P}(p) \neq \varepsilon(p)$. See Definition 3.1).

Example 2.1. For $p = 2314$

$$T(p) = \{(2,3,1), (2,3,4), (2,1,4), (3,1,4)\}$$

and

$$\mathcal{P}(p) = \{(2,3), (2,1), (2,4), (3,1), (3,4), (1,4)\}.$$

It is not difficult to see that $|T(p)| = \binom{|\Sigma|}{3}$ for $|\Sigma| \geqq 3$.

ii) If $C \subseteq S_\Sigma$, $T(C) \equiv \bigcup_{p \in C} T(p)$ and $\mathcal{P}(C) \equiv \bigcup_{p \in C} \mathcal{P}(p)$.

We will state without proof the following simple but useful result.

LEMMA 2.1.

i) $|T(S_\Sigma)| = P(|\Sigma|), 3)$ (*the number of different* 3-*permutations out of a set of* $|\Sigma|$-*elements*).

ii) *If* C *is a consistent subset of* S_Σ *then* $|T(C)| \leqq 4\binom{|\Sigma|}{3}$.

3. S_Σ, LC(p), EXP (p).

3.1. S_Σ as a poset vs. consistent sets.

Graph representation. Consider a graph $G_n = (V, E)$ where $V = S_\Sigma$, $n = |\Sigma|$ and two vertices u, v are joined by an edge iff there exists an adjacent transposition l such that $u = l(v)$. This graph can be represented by a convex polyhedron sometimes called a "permutohedron" (Guilbaud and Rosentiehl [3]). When two vertices u, v are adjacent the arc is directed from u to v if $u = u_1 \cdots u_i u_{i+1} \cdots u_n$ and $v = u_1 \cdots u_{i+1} u_i \cdots u_n$ with $u_i < u_{i+1}$. It is clear that the degree of $u = n - 1$, $\forall u \in G_n$ (see Fig. 1).

DEFINITION 3.1. If $u = u_1 \cdots u_n$ let $\varepsilon(u)$ be the set of pairs (u_i, u_j) which do not introduce an inversion, i.e.,

$$\varepsilon(u) = \{(u_i, u_j) \qquad i < j,\ u_i < u_j\}.$$

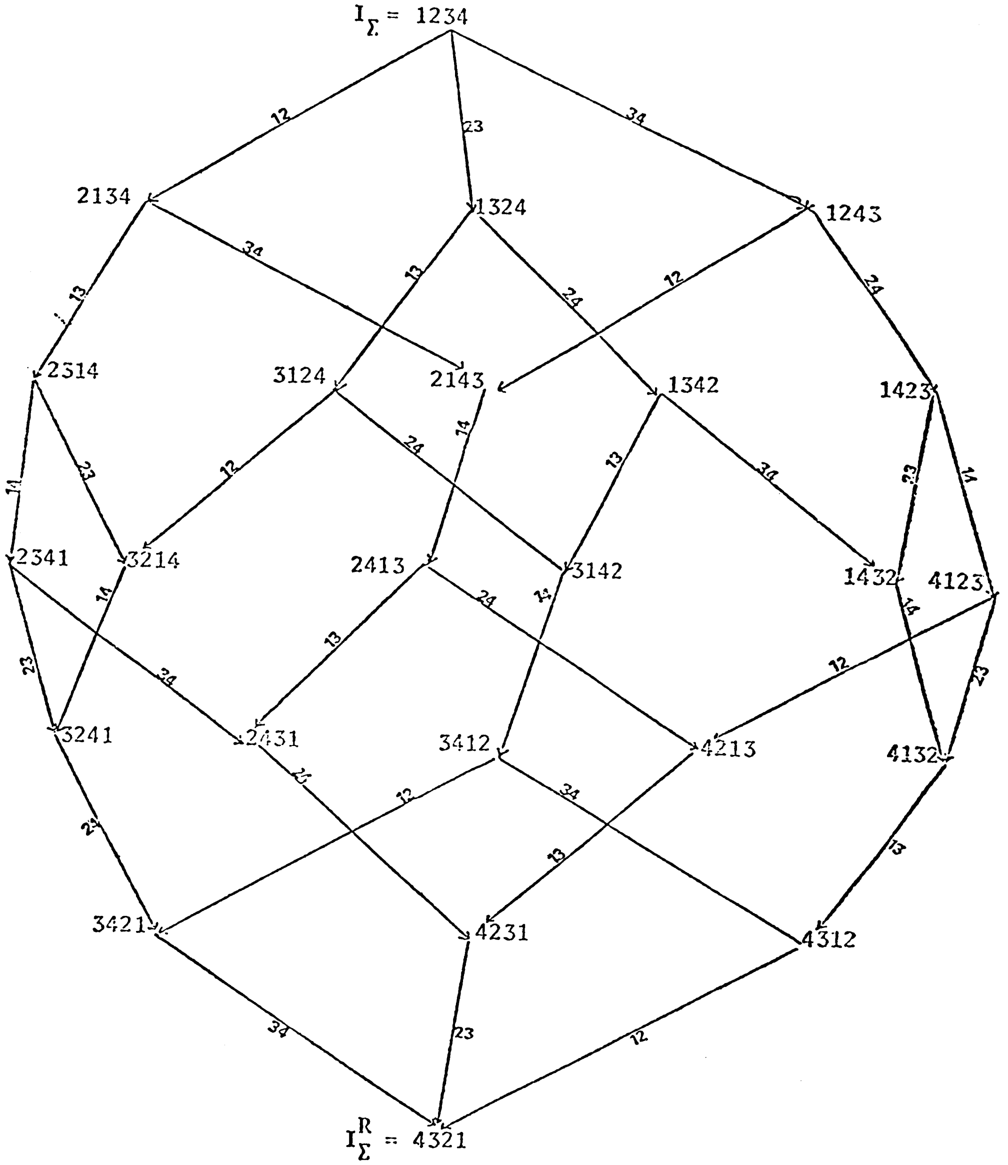


FIG. 1. *Representation of the graph of the permutations of* $\Sigma = \{1, 2, 3, 4\}$. *The relevant transpositions are indicated on each edge.*

Notice that $|\varepsilon(u)| = n(n-1)/2 - \text{INV}(u)$ where INV (u) denotes the number of inversions of u.

DEFINITION 3.2. For $u, v \in S_\Sigma$, $u \geqq v$ iff $\varepsilon(u) \supseteq \varepsilon(v)$.

FACT 3.1. $\geqq$ *is an order relation on* S_Σ *and* $\langle S_\Sigma, \geqq \rangle$ *is a poset with maximum element* I_Σ (*the identity in* S_Σ) *and minimum element* I_Σ^R (*the reverse of* I_Σ).

The following lemma gives the first relation between the poset $\langle S_\Sigma, \geqq \rangle$ and the class of consistent subsets of S_Σ.

LEMMA 3.1. *If L is a chain in* $\langle S_\Sigma, \geqq \rangle$ *then L is a consistent subset of* S_Σ.

Proof (*by contradiction*). Assume that L is cyclic. Then there are three permutations u, v, w in L and three symbols x_i, x_j, x_k in Σ such that

$$u=\cdots x_i\cdots x_j\cdots x_k\cdots,$$
$$v=\cdots x_j\cdots x_k\cdots x_i\cdots,$$
$$w=\cdots x_k\cdots x_i\cdots x_j\cdots.$$

We can assume, w.l.o.g., that $x_i<x_j<x_k$ (the only other essentially different case is $x_i>x_j>x_k$ which can be treated similarly).

(i) $\varepsilon(u)$ contains the ordered pairs $(x_i, x_j)(x_i, x_k)(x_j, x_k)$ and at least two of these pairs do not belong to $\varepsilon(v)$, thus $\varepsilon(v)\not\supseteq\varepsilon(u)$ which means that $v\not\geqq u$. Similarly $\varepsilon(w)\not\supseteq\varepsilon(u)$; then $w\not\geqq u$. On the other hand $\varepsilon(v)$ contains (x_j, x_k) which does not belong to $\varepsilon(w)$; then $\varepsilon(w)\not\supseteq\varepsilon(v)$, thus $w\not\geqq v$.

(ii) Also, $\varepsilon(w)$ contains (x_i, x_j) which does not belong to $\varepsilon(v)$; then $\varepsilon(v)\not\supseteq\varepsilon(w)$ and $v\not\geqq w$.

(i) and (ii) together give us that v and w are not comparable. Thus they cannot be in the same chain and therefore u, v, w are not in the same chain (a contradiction). Q.E.D.

Example 3.1. The set $\{1234, 1243, 1423, 4123, 4132, 4312, 4321\}\subset S_{\{1,2,3,4\}}$ is consistent because it is a chain in $\langle S_{\{1,2,3,4\}}, \geqq\rangle$ (see Fig. 1).

LEMMA 3.2. *If L is a maximal chain in $\langle S_\Sigma, \geqq\rangle$ then $|L|=n(n-1)/2+1$, where $n=|\Sigma|$ and $|T(L)|=4\binom{n}{3}$.*

3.2. LC(p)—The canonical set generated by a permutation p.

DEFINITION 3.3. i) Let $p=p_1p_2\cdots p_n\in S_\Sigma$. We will denote by l_k the adjacent transposition such that

$$l_k(p_1p_2\cdots p_n)=p_1p_2\cdots p_{k-1}\underbrace{p_{k+1}p_k}_{\text{adjacent transposition}}p_{k+2}\cdots p_n$$

(of course l_k is defined for $1\leqq k<n$).

ii) With $p\in S_\Sigma$ we will associate the n permutations $b^j(p)$ defined as follows:

$$b^0(p)=p,$$
$$b^j(p)=l_{n-j}(b^{j-1}(p))\quad \text{for } j=1,\cdots,n-1\quad (\text{note that } n=|\Sigma|).$$

$b^j(p)$ is a permutation obtained from p by the successive application of j adjacent transpositions. The set $B^0(p)=\{b^j(p), j=0, 1,\cdots, n-1\}$ will be called the base set generated by p.

LEMMA 3.3. *If I_Σ denotes the identity in S_Σ then $B^0(I_\Sigma)$ is a chain in $\langle S_\Sigma, \geqq\rangle$.*

Proof. It follows from the fact that $b^j(I_\Sigma)$ contains one inversion more than $b^{j-1}(I_\Sigma)$ for $j=1, 2,\cdots, n-1$. Q.E.D.

Note. For simplicity in notation we will refer to $B^0(I_\Sigma)$ as B^0 and to $b^j(I_\Sigma)$ as b^j. With this convention if $b^j(i)$ denotes the ith symbol in b^j and if $I_\Sigma=x_1x_2\cdots x_n$ then $b^j(n-j)=x_n$ for $j=0, 1,\cdots, n-1$. In other words x_n is at the $(n-j)$th position in the permutation b^j.

DEFINITION 3.4. Again let $p=p_1p_2\cdots p_n\in S_\Sigma$. The *canonical set* generated by p is the set LC $(p_1p_2\cdots p_n)$ defined by

$$\mathrm{LC}\,(p_1p_2\cdots p_n)\equiv\left[\bigcup_{j=0}^{n-3} p_n\cdots p_{n-j}B^0(p_1p_2\cdots p_{n-j-1})\right]\cup B^0(p_1p_2\cdots p_n).$$

Let us illustrate with one example the preceding definitions.

Example 3.2. Let $\Sigma=\{1,2,3,4\}$ and $I_\Sigma=1234$.

$$\begin{aligned} B^0 \equiv B^0(I_\Sigma) &= \{b^0(1234), b^1(1234), b^2(1234), b^3(1234)\} \\ &= \{123\underline{4}, 12\underline{4}3, 1\underline{4}23, \underline{4}123\}. \end{aligned}$$

Now by Lemma 3.3 B^0 is a chain in $\langle S_\Sigma, \geqq\rangle$ so we can represent B^0 as in Fig. 2. Notice that the symbol 4 is moved toward the left one position at a time until it reaches the first place.

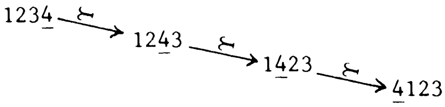


FIG. 2.

Finally let us compute LC (I_Σ) = LC (1234). By definition we have

$$\begin{aligned} \text{LC}\,(1234) &= [4B^0(123) \cup 43B^0(12)] \cup B^0(1234) \\ &= [4\{12\underline{3}, 1\underline{3}2, \underline{3}12\} \cup 43\{1\underline{2}, \underline{2}1\}] \\ &\quad \cup \{123\underline{4}, 12\underline{4}3, 1\underline{4}23, \underline{4}123\} \\ &= \{4123, 4132, 4312, 4321, 1234, 1243, 1423\}. \end{aligned}$$

Notice that LC $(1234) = 4\text{LC}\,(123) \cup 1B^0(234)$. In fact this is always the case as is stated in the following lemma whose proof follows readily from the definitions.

LEMMA 3.4.

1. $B^0(p_1 \cdots p_n) = p_1 B^0(p_2 \cdots p_n) \cup p_n(p_1 p_2 \cdots p_{n-1})$.
2. LC $(p_1 \cdots p_n) = p_n$ LC $(p_1 p_2 \cdots p_{n-1}) \cup p_1 B^0(p_2 \cdots p_n)$.
3. LC (I_Σ) *is a maximal chain in* $\langle S_\Sigma, \geqq\rangle$.

The contents of the preceding lemma are expressed graphically in Fig. 3.

COROLLARY 3.1. *If* $p \in S_\Sigma$ *then* LC (p) *is a consistent subset of* S_Σ.

Proof. Follows from the consistency of LC (I_Σ). Q.E.D.

Remark. Notice that LC (p) is not always a chain in $\langle S_\Sigma, \geqq\rangle$; however the consistency of LC (I_Σ) implies the consistency of LC $(p) \subset S_\Sigma$.

Notation. For $x, y \in \Sigma$ and $A \subset \Sigma - \{x, y\}$ we will say that an ordered triple (x, y, z) is of the form (x, y, A) if $z \in A$. Similarly with (x, A, y) and (A, x, y).

The importance of all the preceding machinery will become transparent in the next result which gives vital information about those consistent sets containing LC (p).

THEOREM 3.1. *Let C be a consistent set such that* LC $(p) \subset C$ *for some* $p \in S_\Sigma$, *and let* $\rho_1 \in \{2, \cdots, n-1\}$.

If $w \in S_\Sigma$ *is such that* $w_1 = p_{\rho_1}$ *then* $w \notin C$.

Proof. Let $w \in S_\Sigma$: $w_1 = p_{\rho_1}$ and $w_n = p_j$ for some $j \in \{1, 2, \cdots, n\}$.

Case 1. $\rho_1 < j$. In this case $2 \leqq \rho_1 < j \leqq n$ and w is bounded to contain triples of the form

(i) $(p_{\rho_1}, \{p_1, \cdots, p_{\rho_1 - 1}\}, p_j)$ for some j and ρ_1: $2 \leqq \rho_1 < j \leqq n$.

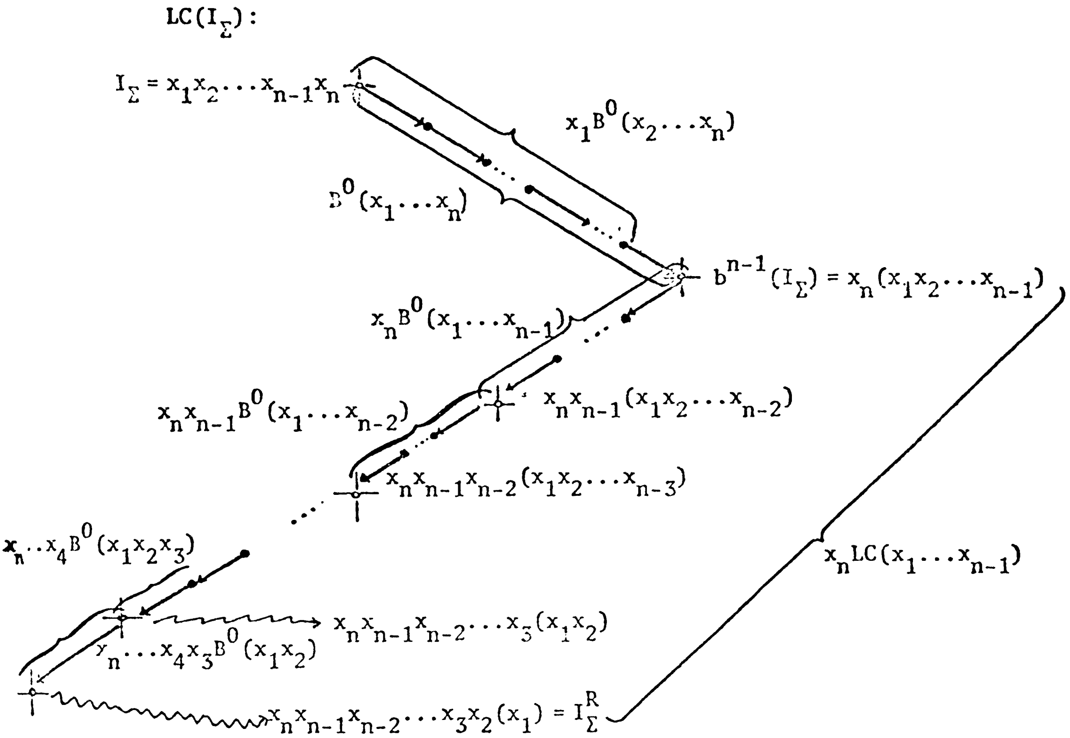

$LC(I_\Sigma)$:
$I_\Sigma = x_1x_2\cdots x_{n-1}x_n$
$x_1B^0(x_2\cdots x_n)$
$B^0(x_1\cdots x_n)$
$b^{n-1}(I_\Sigma) = x_n(x_1x_2\cdots x_{n-1})$
$x_nB^0(x_1\cdots x_{n-1})$
$x_nx_{n-1}B^0(x_1\cdots x_{n-2})$
$x_nx_{n-1}(x_1x_2\cdots x_{n-2})$
$x_nx_{n-1}x_{n-2}(x_1x_2\cdots x_{n-3})$
$x_n\cdots x_4B^0(x_1x_2x_3)$
$x_nLC(x_1\cdots x_{n-1})$
$x_nx_{n-1}x_{n-2}\cdots x_3(x_1x_2)$
$x_n\cdots x_4x_3B^0(x_1x_2)$
$x_nx_{n-1}x_{n-2}\cdots x_3x_2(x_1) = I^R_\Sigma$

FIG. 3. *Structure of* LC (I_Σ). *The big black dots represent permutations*; *every arrow represents an adjacent transposition*; *those permutations which are the beginning and/or the end of a subchain are denoted by*$-\!\circ\!-$; *the maximum element of* LC (I_Σ) *is* I_Σ *and the minimum is* I^R_Σ; *every left bracket { groups a subchain and it is worth noticing that they are interlaced. Notice that the cardinality of the subchains decreases from n down to* 2. *The right brackets } express the content of Lemma* 3.4, *namely that* LC $(x_1\cdots x_n)$ *is partitioned in the two sets* $x_1B^0(x_2\cdots x_n)$ *and* x_n LC $(x_1\cdots x_{n-1})$.

On the other hand $T(\mathrm{LC}(p))$ contains triples of the form

(ii) $(p_j, p_{\rho_1}, \{p_1, \cdots, p_{\rho_1-1}\})$

and

(iii) $(\{p_1, \cdots, p_{\rho_1-1}\}, p_j, p_{\rho_1})$.

Therefore $w \notin C$ because C is consistent and (i), (ii) and (iii) are not consistent triples. Now, for a fixed value of ρ_1 there are $(n-\rho_1) * (n-2)!$ of such w permutations.

Case 2. $\rho_1 > j$. $p = p_1p_2 \cdots p_n$ and $q = p_n(p_1 \cdots p_{n-1})$ are elements of LC (p) and the set $\{p, q, w\}$ is cyclic; therefore $w \notin C$.

The two cases considered above give us that for a fixed value of $\rho_1 \in \{2, \cdots, n-1\}$, if $w \in S_\Sigma$ is such that $w_1 = p_{\rho_1}$ then $w \notin C$ because all the $(n-1)!$ such permutations are not consistent with LC (p). Q.E.D.

COROLLARY 3.2. *If C is a consistent set,* LC $(p) \subset C$ *for some* $p \in S_\Sigma$ *and if* $v \in C$, *then* $v_1 = p_1$ *or* $v_1 = p_n$.

Proof. Immediate from the theorem. Q.E.D.

3.3. EXP (p)—The expansion of a permutation p. In this section we will define and study the properties of a special set denoted by EXP (p) which plays a vital role in this paper.

DEFINITION 3.5. *The expansion of a permutation*—EXP (p).

If $x_i \in \Sigma$ then EXP $(x_i) \equiv x_i$ (the expansion of a symbol is itself), and if $p = p_1 \cdots p_n \in S_\Sigma$ then

$$\text{EXP}(p_1 \cdots p_n) \equiv p_1 \,\text{EXP}(p_2 \cdots p_n) \cup p_n \,\text{EXP}(p_1 \cdots p_{n-1}).$$

Note. It is straightforward to prove by induction on n, that $|\text{EXP}(p_1 \cdots p_n)| = 2^{n-1}$.

Example 3.3. Let $\Sigma = \{1, 2, 3, 4\}$ and $p = 1234$.

$$\begin{aligned}\text{EXP}(1234) &= 1\,\text{EXP}(234) \cup 4\,\text{EXP}(123)\\ &= 1\{2\,\text{EXP}(34) \cup 4\,\text{EXP}(23)\} \cup 4\{1\,\text{EXP}(23) \cup 3\,\text{EXP}(12)\}\\ &= 1\{2(34), 2(43), 4(23), 4(32)\} \cup 4\{1(23), 1(32), 3(12), 3(21)\}\\ &= \{1234, 1243, 1423, 1432\} \cup \{4123, 4132, 4312, 4321\}.\end{aligned}$$

Later on we will see that EXP (p) has a lot of structure built in and that LC (p) is a subset of it.

The following technical facts about EXP (p) will be used later on. (The proofs follow readily from the definitions of EXP (p) and LC (p).)

LEMMA 3.5.

i) $$\text{EXP}(p_1 \cdots p_n) = \left[\bigcup_{j=1}^{n-2} p_1 \cdots p_j \,\text{EXP}(p_{j+1} \cdots p_n)\right] \cup p_n \,\text{EXP}(p_1 \cdots p_{n-1}),$$

ii) $$p_1 \,\text{EXP}(p_2 \cdots p_n) = \left[\bigcup_{j=1}^{n-2} p_1 \cdots p_j p_n \,\text{EXP}(p_{j+1} \cdots p_{n-1})\right] \cup \{p_1 p_2 \cdots p_n\},$$

iii) LC $(p_j \cdots p_n) \subset$ EXP $(p_j \cdots p_n)$ *for* $j \in \{1, 2, \cdots, n\}$,

iv) $p_1 \cdots p_j$ LC $(p_{j+1} \cdots p_n) \subset$ EXP $(p_1 \cdots p_n)$ *for* $j \in \{1, 2, \cdots, n-1\}$.

At this point it is convenient to illustrate what we know about EXP (p) by drawing its graphical representation. Let us take now $\Sigma = \{1, 2, 3, 4, 5\}$ and $p = 12345$; then EXP (12345) may be represented as in Fig. 4.

Maximality and consistency of EXP (p). Now, we will see that EXP (p) is more than just a nice set, it is a *maximal consistent set.*

THEOREM 3.2. EXP (p) *is a consistent subset of* S_Σ, *for any* $p \in S_\Sigma$.

Proof sketch. (i) LC $(p) \subset$ EXP (p) and LC (p) is a consistent subset of S_Σ by Lemma 3.5 and Corollary 3.1; thus T(LC (p)) is a consistent set of triples such that T(LC $(p)) \subset T$(EXP (p)).

(ii) One can prove that in fact T(LC $(p)) = T$(EXP (p)) which implies that EXP (p) is a consistent subset of S_Σ. The idea is the following: EXP $(p_1 \cdots p_k p_{k+1}) = p_{k+1}$ EXP $(p_1 \cdots p_k) \cup p_1$ EXP $(p_2 \cdots p_{k+1})$ by the definition of EXP.

Thus if one proves that

(I) $$T(p_{k+1} \,\text{EXP}(p_1 \cdots p_k)) \subset T(\text{LC}(p_1 \cdots p_{k+1}))$$

and

(II) $$T(p_1 \,\text{EXP}(p_2 \cdots p_{k+1})) \subset T(\text{LC}(p_1 \cdots p_{k+1})),$$

then $T(\text{EXP}(p_1 \cdots p_k p_{k+1})) \subset T(\text{LC}(p_1 \cdots p_{k+1}))$ which implies that $T(\text{EXP}(p_1 \cdots p_k p_{k+1})) = T(\text{LC}(p_1 \cdots p_{k+1}))$; so the main job is in the proof of (I) and (II). Q.E.D.

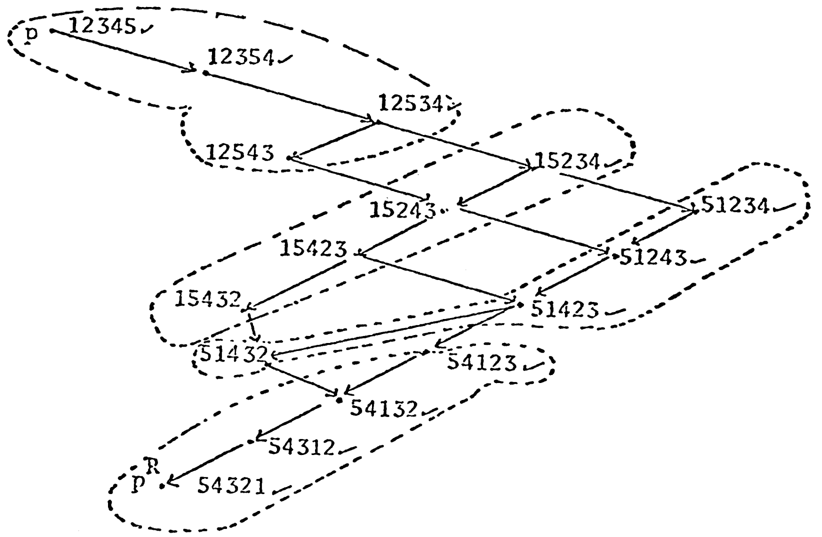


FIG. 4. *Graphical representation of* EXP (12345). *Each arrow represents an adjacent transposition.* • |EXP (12345)| = 16 *and* LC (12345) ⊂ EXP (12345). • *The elements of* LC (12345) *have a check mark at their right.* • *We have encircled those elements which have a common prefix of length two, namely* 12 EXP (345), 15 EXP (234), 51 EXP (234) *and* 54 EXP (123).

DEFINITION 3.6. If $\Sigma = \{x_1, \cdots, x_n\}$ and $Q \subset S_\Sigma$, let

$$Q_{x_i} \equiv \{v \in S_{\Sigma-\{x_i\}}: x_i v \in Q\}.$$

It is clear that if $T \subset Q \subset S_\Sigma$ then $T_{x_i} \subset Q_{x_i} \subset S_{\Sigma-\{x_i\}}$ and that T is consistent iff T_{x_i} is consistent.

THEOREM 3.3. EXP (p) *is a maximal consistent subset of* S_Σ.

Proof. We know by the preceding theorem that EXP (p) is a consistent subset of S_Σ. To prove that it is a maximal consistent subset of S_Σ we will prove the following equivalent statement:

If C is a consistent subset of S_Σ: EXP $(p) \subset C$ *for some* $p \in S_\Sigma$ *then* $C =$ EXP (p).

(i) First, notice that for any n, if $|\Sigma| = n$ and EXP $(p) \subset C$ then LC $(p) \subset C$, (LC $(p) \subset$ EXP (p) by Lemma 3.5) thus $w \in C \Rightarrow w_1 = p_1$ or $w_1 = p_n$ by Corollary 3.2.

Now the proof is by induction on $n = |\Sigma|$ (the cardinality of Σ).

Basis. If $n = 2$, $w = p_1 p_2$ or $w = p_2 p_1$ by (i) above and in both cases $w \in$ EXP $(p_1 p_2)$, so $C \subset$ EXP $(p_1 p_2) \Rightarrow C =$ EXP (p).

Induction hypothesis. Assume that for $n = k$, EXP $(p_1 \cdots p_k) \subset C \Rightarrow C \subset$ EXP $(p_1 \cdots p_k)$. Let us prove it for $n = k+1$.

Let $w \in C$; then $w_1 = p_1$ or $w_1 = p_{k+1}$ by (i). Thus

(ii) $$w = p_1 v \quad \text{with } v \in C_{p_1} \subset S_{\Sigma-\{p_1\}}$$

or

$$w = p_{k+1} v \quad \text{with } v \in C_{p_{k+1}} \subset S_{\Sigma-\{p_{k+1}\}}$$

(see the definition of C_{p_i} in Definition 3.6).

Now, EXP $(p_1 \cdots p_{k+1}) \subset C \Rightarrow$ EXP $(p_2 \cdots p_{k+1}) \subset C_{p_1}$ and EXP $(p_1 \cdots p_k) \subset C_{p_{k+1}}$ by the definition of EXP. But C_{p_1} and $C_{p_{k+1}}$ are consistent subsets of $S_{\Sigma-\{p_1\}}$ and $S_{\Sigma-\{_{k+1}\}}$ respectively because C is a consistent subset of S_Σ, so $C_{p_1} \subset$ EXP $(p_2 \cdots p_{k+1})$

and $C_{p_{k+1}} \subset \text{EXP}\,(p_1 \cdots p_k)$ by the induction hypothesis. Therefore $v \in \text{EXP}\,(p_2 \cdots p_{k+1})$ or $v \in \text{EXP}\,(p_1 \cdots p_k)$ by (ii) above, which implies that $p_1 v \in p_1 \text{EXP}\,(p_2 \cdots p_{k+1}) \subset \text{EXP}\,(p_1 \cdots p_{k+1})$ or $p_{k+1} v \in p_{k+1} \text{EXP}\,(p_1 \cdots p_k) \subset \text{EXP}\,(p_1 \cdots p_{k+1})$, so in either case $w \in \text{EXP}\,(p_1 \cdots p_{k+1})$. Q.E.D.

The following corollary assures us that no matter how hard we try to find consistent sets containing LC (p) the most we can get is EXP (p).

COROLLARY 3.3. *Let C be a consistent subset of S_Σ such that* LC $(p) \subset C$ *for some* $p \in S_\Sigma$. *Then* $C \subseteq \text{EXP}\,(p)$. *In other words* $|\text{EXP}\,(p)| = \text{maximum}_{C \supset \text{LC}\,(p)} \{|C|$: *$C$ is consistent*$\}$.

Proof (*by contradiction*). Let C a consistent subset of S_Σ such that $\text{LC}\,(p) \subset C$ for some $p \in S_\Sigma$ and assume that $C \not\subseteq \text{EXP}\,(p)$. First, $C \not\supset \text{EXP}\,(p)$ because EXP (p) is a maximal consistent subset of S_Σ by the preceding theorem; therefore $C - \text{EXP}\,(p) \neq \emptyset$ and $\text{EXP}\,(p) \not\subset C$. If $v \in C - \text{EXP}\,(p)$ then v is consistent with LC (p) because $\text{LC}\,(p) \subset C$; however $v \notin \text{EXP}\,(p)$.

Now, if $T(v) \subset T(\text{LC}\,(p))$ then $v \cup \text{EXP}\,(p)$ is a consistent set because $T(\text{LC}\,(p)) \subset T(\text{EXP}\,(p))$, but this contradicts the maximality of EXP (p). Therefore $T(v) \not\subset T(\text{LC}\,(p))$, so there exists a triple in $T(v)$ which is not in $T(\text{LC}\,(p))$ and still does not introduce inconsistencies in $T(C) \supsetneq T(\text{LC}\,(p))$ because by assumption C is consistent. Then $|T(C)| > |T(\text{LC}\,(p))| = 4\binom{n}{3}$, which contradicts the fact that for every consistent set C, $|T(C)| \leqq 4\binom{n}{3}$ by Lemma 2.1.

Conclusion. There is no $v \in C - \text{EXP}\,(p)$, thus $C \subseteq \text{EXP}\,(p)$. Q.E.D.

At this point we know that in order to find consistent sets of cardinality bigger than $|\text{EXP}\,(p)| = 2^{n-1}$ (if any) we must avoid any set containing LC (p). This is precisely the purpose of the next section.

4. Searching for consistent sets of cardinality bigger than 2^{n-1} (if any). The next theorem opens the door to new consistent sets. It is the first general theorem about maximal consistent sets.

THEOREM 4.1. *Let* $|\Sigma| = n$ *and let* $A \subset S_\Sigma$ *be a maximal consistent set such that* $|T(A)| = 4\binom{n}{3}$; *assume there is a* $w \in A$ *and a fixed ordered pair of symbols* (p_i, p_j), $p_i, p_j \in \Sigma$ *such that* $T(A - \{w\})$ *does not contain any triple of the form* (p_i, p_k, p_j), (p_k, p_i, p_j) *or* (p_i, p_j, p_k).

If the set $T(w) - T(A - \{w\}) = \{(p_k, p_i, p_j)\}$ *has cardinality* $(n-2)$ *and if there is a maximal consistent set* A^0: $A^0 \subset S_{\Sigma - \{p_i, p_j\}}$ *and* $p_j p_i A^0 \subset A$ *then the set* $A' = (A - \{w\}) \cup p_i p_j A^0$ *is a maximal consistent subset of* S_Σ.

Comment. This result allows us to obtain consistent sets of bigger cardinality than a given consistent set A, if A satisfies the hypothesis of the theorem, so it may be a helpful tool in the search for a *maximum* consistent set.

Proof of Theorem 4.1.

Part I. *A' is consistent.* The proof is by contradiction.

Assume that A' is not consistent. Then there are in $T(A')$ triples of the form

1) (p_r, p_s, p_t), (p_s, p_t, p_r) and (p_t, p_r, p_s),

or

2) (p_r, p_t, p_s), (p_s, p_r, p_t) and (p_t, p_s, p_r).

Without loss of generality we can assume that $T(A')$ contains triples of the form 1), and since $T(A') = T(A - \{w\}) \cup T(p_i p_j) A^0$ at most two out of the three triples can be in either $T(A - \{w\})$ or $T(p_i p_j A^0)$; $(p_i p_j A^0)$ is consistent because A^0 is consistent.

(i) Now, if any two of the three triples of the form 1) are in $T(p_i p_j A^0)$ they must be in $T(A^0)$ but $T(A^0) \subset T(A)$ so they belong to $T(A)$.

(ii) On the other hand, if the remaining triple is in $T(A - \{w\})$ it is also in $T(A)$ because $T(A - \{w\}) \subset T(A)$.

Therefore by (i) and (ii) A is not consistent, a contradiction.

(iii) So at this point we know that two of the three triples must be in $T(A-\{w\})$ and the third one must be in $T(p_ip_jA^0)$. There is no loss of generality by taking (p_r, p_s, p_t) and $(p_s, p_t, p_r) \in T(A-\{w\})$ and $(p_t, p_r, p_s) \in T(p_ip_jA^0)$.

It is clear that $(p_t, p_r, p_s) \notin T(A^0)$ because otherwise it will contradict the consistency of A (remember that $T(A^0) \subset T(A)$ and $T(A-\{w\}) \subset T(A)$). Therefore we can assume that $(p_t, p_r, p_s) \in T(p_ip_jA^0) - T(A^0)$. Here there are two cases to be considered.

Case 1. $p_t = p_i$ or $p_t = p_j$ *and* $(p_r, p_s) \in \mathcal{P}(A^0)$. In this case $(p_t, p_r, p_s) \in T(p_jp_iA^0)$ because we are just choosing p_t to be p_i or p_j, so from the triples (p_t, p_r, p_s) point of view the relative order of p_i and p_j is irrelevant; but this being the case we have that $(p_t, p_r, p_s) \in T(A)$ because $T(p_jp_iA^0) \subset T(A)$ by hypothesis. Again this together with (ii) contradicts the consistency of A.

Case 2. $p_t = p_i$, $p_r = p_j$ *and* $p_s \in \Sigma - \{p_i, p_j\}$. Here we have by (ii) that $(p_s, p_i, p_j) \in T(A-\{w\})$, contradicting the hypothesis that $T(A-\{w\})$ does not contain any triple of the form $(-, p_i, p_j)$. Therefore A' must be consistent. Q.E.D.

Part II. *A' is maximal.* The proof is by contradiction.

Assume there exists a set C' which is consistent and such that $C' \supsetneq A'$.

Let $v \in C'$, $v \notin A'$. $T(w)$ contains all the triples of the form (p_k, p_i, p_j); thus $w \notin p_jp_iA^0 \subset A$, which implies that $p_jp_iA^0 \subset A-\{w\}$. This means that the only triples which are in $T(p_ip_jA^0) - T(A-\{w\})$ are those of the form $(p_i, p_j, -)$; there are $(n-2)$ of them, so

$$\text{(o)} \qquad |T(p_ip_jA^0) - T(A-\{w\})| = n-2.$$

On the other hand

$$\begin{aligned} |T(A)| &= |T(A-\{w\})| + |T(w)| - |T(A-\{w\}) \cap T(w)| \\ &= |T(A-\{w\})| + |T(w) - T(A-\{w\})| \\ &= |T(A-\{w\})| + (n-2) \end{aligned}$$

by hypothesis, which implies that

$$\text{(i)} \qquad |T(A-\{w\})| = |T(A)| - (n-2).$$

Putting (o) and (i) together we have that

$$\text{(ii)} \qquad |T(A')| = |T(A)| = 4\tbinom{n}{3}.$$

Now, $C' \supsetneq A'$ and $|T(A')| = 4\binom{n}{3} \Rightarrow T(C') = T(A')$ because C' is consistent, (see Lemma 2.1, part ii). Thus $T(v) \subset T(A') = T((A-\{w\}) \cup p_ip_jA^0)$. Let us consider possibilities for $T(v)$.

Case 1. $T(v) \subset T(A-\{w\})$. This contradicts the maximality of A; therefore

$$\text{(iii)} \qquad T(v) \not\subset T(A-\{w\}).$$

Case 2. If $T(v) \subset T(p_ip_jA^0)$ then $p_ip_jA^0 \cup \{v\}$ is a consistent set ($p_ip_jA^0$ is consistent because A^0 is consistent and $v \notin p_ip_jA^0$ because $v \notin A'$); thus v must be such that $v_1 = p_i$, $v_2 = p_j$ because if not $T(v)$ will contain a triple of the form

$$\text{(iv)} \qquad (p_t, p_j, p_i) \notin T(p_ip_jA^0).$$

Now let v' be the $(n-2)$ permutation obtained from v by deleting from it the symbols p_i and p_j. Then $A^0 \cup \{v'\}$ is a consistent subset of $S_{\Sigma-\{p_i,p_j\}}$ by (iv). On the other hand $v' \notin A^0$ because by assumption $v \notin A'$ so $A^0 \cup \{v'\}$ is a consistent subset which contains A^0, contradicting the maximality of A^0; therefore Case 2 is not possible, namely

$$\text{(v)} \qquad T(v) \not\subset T(p_ip_jA^0).$$

At this point we have by (iii) and (v) that $T(v) \not\subset T(A-\{w\})$ and $T(v) \not\subset T(p_ip_jA^0)$; however

$$\text{(vi)} \qquad T(v) \subset T(A-\{w\}) \cup T(p_ip_jA^0) = T(A').$$

Thus

$$\text{(vii)} \qquad T(v)-T(A-\{w\}) \neq \varnothing \quad \text{and} \quad T(v)-T(p_ip_jA^0) \neq \varnothing.$$

Now those triples in $T(v)-T(p_ip_jA^0)$ cannot be of the form:

(p_t, p_i, p_j) because $(p_t, p_i, p_j) \notin T(A-\{w\})$ (there are at most $(n-2)$ triples of this form and w got all of them; remember that (p_i, p_j) is fixed);

or

(p_i, p_t, p_j) because $(p_i, p_t, p_j) \notin T(A')$ by hypothesis.

By noticing that $T(w)$ contains triples of the form (p_k, p_i, p_j) (by hypothesis) we conclude that $w \notin p_jp_iA^0 \subset A$. Then $p_jp_iA^0 \subset A-\{w\}$; thus $T(p_jp_iA^0) \subset T(A-\{w\})$, which implies that

$$\text{(viii)} \quad T(p_ip_jA^0)-T(A-\{w\}) = \{(p_i, p_j, p_t) \text{ for some } p_t \in \Sigma-\{p_i, p_j\}\}.$$

Therefore those triples in $T(v)-T(A-\{w\})$ must be of the form (p_i, p_j, p_t) for some $p_t \in \Sigma-\{p_i, p_j\}$ by (vi) and (viii) above. So v must be such that $v_1 = p_i$, $v_2 = p_j$. However $v \notin p_ip_jA^0$; thus

$$\text{(ix)} \qquad v = p_ip_jv' \quad \text{with } v' \notin A^0$$

(v' is an $(n-2)$ permutation in $S_{\Sigma-\{p_i,p_j\}}$).

Now, we have that $T(v) \subset T(A')$ by (vi), and this implies that $T(v') \subset T(A')$ because $v = p_ip_jv'$. On the other hand $T(A^0) \subset T(A')$; thus $T(A^0) \cup T(v') \subset T(A')$. But A' is consistent, so $A^0 \cup v'$ is a consistent subset of $S_{\Sigma-\{p_i,p_j\}}$ such that $A^0 \cup v' \supsetneq A^0$ ($v' \notin A^0$ by (ix) which contradicts the maximality of A^0). Therefore (vi) is false, namely $T(v) \not\subset T(A')$; this means that C' contains a triple which is not in $T(A')$, so $|T(C')| > |T(A')| = 4\binom{n}{3}$ implies C' is not consistent, a contradiction.

Conclusion. There is no consistent set C': $C' \supsetneq A'$, so A' is a maximal consistent set. Q.E.D.

The preceding theorem gives us a way to identify some maximal consistent sets which are not maximum and it can be applied to EXP (p). Let us see how this can be done.

Remarks.

i. EXP (p) is a maximal consistent set:

$$|T(\text{EXP}\,(p)| = |T(\text{LC}\,(p))| = 4\binom{n}{3} \text{ where } n = |\Sigma|.$$

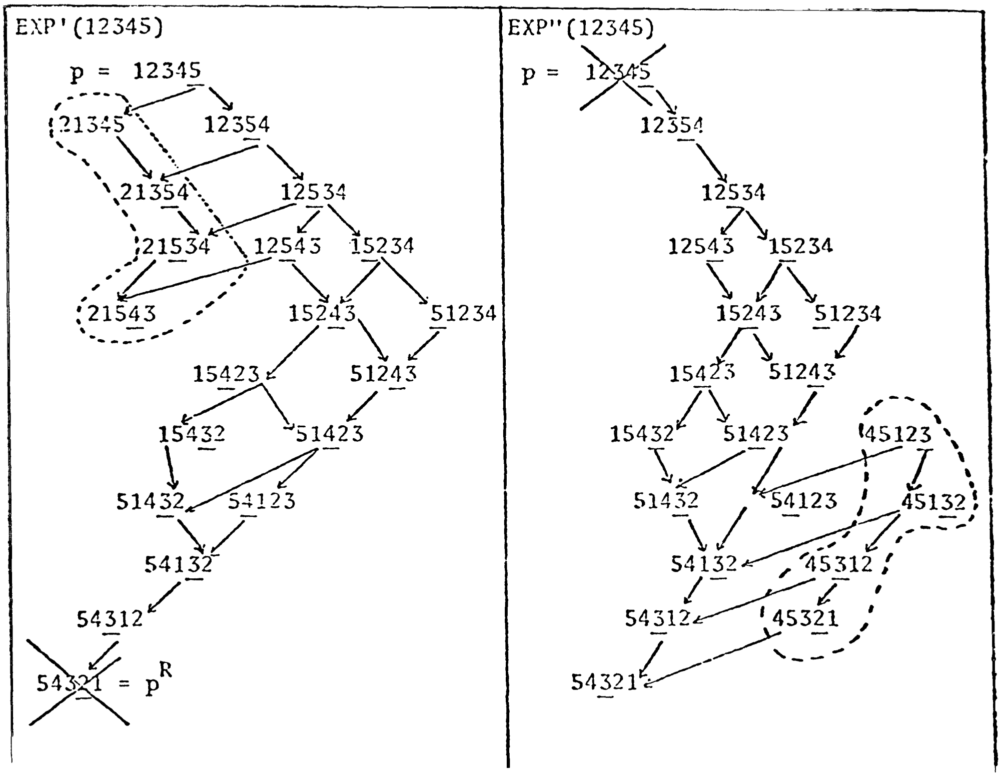
ii. p^R (the reverse of p) is the only permutation in EXP (p) which contains triples of the form (p_t, p_2, p_1), and $T(\text{EXP}\,(p)-p^R)$ does not contain any triple involving the ordered pair (p_2, p_1). (This can be seen by a straightforward inductive argument using the definition of EXP (p).) Moreover p^R contains all the $(n-2)$ possible triples of the form (p_t, p_2, p_1); thus $T(p^R)-T(\text{EXP}\,(p)-p^R)$ has cardinality $(n-2)$.

iii. EXP $(p_3 \cdots p_n)$ is a maximal consistent subset of $S_{\Sigma-\{p_1,p_2\}}$ such that p_1p_2 EXP $(p_3 \cdots p_n) \subset$ EXP (p) (see the definition of EXP (p)).

iv. Now, by taking $A = \text{EXP}\,(p)$ and $A^0 = \text{EXP}\,(p_3 \cdots p_n)$ we have that $\text{EXP}\,(p)$ and $\text{EXP}\,(p_3 \cdots p_n)$ satisfy the hypothesis of the preceding theorem by remarks i, ii and iii above. Therefore the set $\text{EXP}'\,(p) = (\text{EXP}\,(p) - p^R) \cup p_2p_1\,\text{EXP}\,(p_3 \cdots p_n)$ is a *maximal consistent subset of* S_Σ (see Fig. 5a below). It is worth noticing that $|\text{EXP}'\,(p)| \geqq |\text{EXP}\,(p)|$ with equality iff $|\Sigma| = 3$; specifically

$$\begin{aligned}|\text{EXP}'\,(p)| &= |\text{EXP}\,(p) - p^R| + |p_2p_1\,\text{EXP}\,(p_3 \cdots p_n)|\\ &= 2^{n-1} - 1 + |\text{EXP}\,(p_3 \cdots p_n)|\\ &= 2^{n-1} - 1 + 2^{n-3} = \tfrac{5}{4}2^{n-1} - 1.\end{aligned}$$

v. Similarly by noticing that the only permutation in $\text{EXP}\,(p)$ which contains the ordered pair (p_{n-1}, p_n) is p and that $T(p) - T(\text{EXP}\,(p) - p)$ has cardinality $(n-2)$, we can try to find another maximal consistent set different to $\text{EXP}'\,(p)$, and indeed that is the case because $\text{EXP}\,(p_1 \cdots p_{n-2})$ is a maximal consistent subset of $S_{\Sigma - \{p_{n-1}, p_n\}}$ such that $p_np_{n-1}\,\text{EXP}\,(p_1 \cdots p_{n-2}) \subset \text{EXP}\,(p)$. Therefore the set $\text{EXP}''\,(p) = (\text{EXP}\,(p) - p) \cup p_{n-1}p_n\,\text{EXP}\,(p_1 \cdots p_{n-2})$ (see Fig. 5b) is a *maximal consistent subset of* S_Σ, whose cardinality is again $\tfrac{5}{4}2^{n-1} - 1$.



(a) $\text{EXP}'\,(p) = (\text{EXP}\,(p) - p^R)$
$\cup\, p_2p_1\,\text{EXP}\,(p_3p_4p_5)$.

- *The encircled elements are those which have been added to* $\text{EXP}\,(p)$, *namely* $p_2p_1\,\text{EXP}\,(p_3p_4p_5)$.
- *Notice that* p^R *has been crossed out.*
- $|\text{EXP}'\,(12345)| = \tfrac{5}{4}2^4 - 1 = 19$.

(b) $\text{EXP}''\,(p) = (\text{EXP}\,(p) - p)$
$\cup\, p_4p_5\,\text{EXP}\,(p_1p_2p_3)$.

- *The encircled elements are those which have been added to* $\text{EXP}\,(p)$, *namely* $p_4p_5\,\text{EXP}\,(p_1p_2p_3)$.
- *Notice that p has been crossed out.*
- $|\text{EXP}''\,(12345)| = \tfrac{5}{4}2^4 - 1 = 19$.

FIG. 5

Let us put the preceding remarks together in the following corollary:

COROLLARY 4.1. $\mathrm{EXP}'(p) \equiv (\mathrm{EXP}(p) - p^R) \cup p_2p_1 \,\mathrm{EXP}(p_3 \cdots p_n)$ *and* $\mathrm{EXP}''(p) \equiv (\mathrm{EXP}(p) - p) \cup p_{n-1}p_n \,\mathrm{EXP}(p_1 \cdots p_{n-2})$ *are maximal consistent subsets of* S_Σ *of cardinality* $\frac{5}{4}2^{n-1} - 1$, *which is bigger than* $|\mathrm{EXP}(p)|$ *for* $n > 3$.

Until now our efforts to obtain consistent sets with more than 2^{n-1} elements have been successful. By studying carefully the structure of $\mathrm{EXP}'(p)$ and $\mathrm{EXP}''(p)$ we can obtain an even bigger consistent set. Let us see how this can be done.

Remarks.

1. $\mathrm{LC}(p) \supset p_n \cdots p_{n-j}B^0(p_1 \cdots p_{n-j-1})$ for $j = 0, 1, 2, \cdots, (n-2)$ and if $j = n-3$, $p_n \cdots p_{n-j}B^0(p_1 \cdots p_{n-j-1}) = p^R$ (by the definition of $\mathrm{LC}(p)$). Thus

$$\begin{aligned}\mathrm{LC}(p) - p^R &\supset p_n \cdots p_{n-j}B^0(p_1 \cdots p_{n-j-1}) \quad \text{for } j = 0, 1, 2, \cdots, (n-4)\\ &\supset p_n \cdots p_{n-j}\underline{p_1}B^0(p_2 \cdots p_{n-j-1}),\end{aligned}$$

and by the definition of $p_n \cdots p_{n-j}\underline{p_1}B^0(p_2 \cdots p_{n-j-1})$ the permutation $p_n \cdots p_{n-j}\underline{p_1}\, \underline{p_{n-j-1}}(p_2 \cdots p_{n-j-2}) \in p_n \cdots p_{n-j}p_1\overline{B^0}(p_2 \cdots p_{n-j-1})$ which contains the triple $(\overline{p_1}, \overline{p_{n-j-1}}, p_2)$. Thus

$$\text{(i)} \qquad T(\mathrm{LC}(p) - p^R) \supset \{(p_1, p_{n-j-1}, p_2) \text{ for } j = 0, 1, 2, \cdots, n-4\}.$$

On the other hand, $p_1p_n(p_2 \cdots p_{n-1})$ is in $\mathrm{LC}(p)$ and

(ii) this permutation contains the triple (p_1, p_n, p_2).

Thus

$$\text{(iii)} \qquad T(\mathrm{LC}(p) - p^R) \supset \{(p_1, p_t, p_2) \text{ for } t = n, n-1, \cdots, 3\}$$

by (i) and (ii).

Now, it is clear that p does not contain any triple of the form (p_1, p_t, p_2). So we have that $T(\mathrm{LC}(p) - \{p, p^R\}) \supset \{(p_1, p_t, p_2)$ for $t = n, n-1, \cdots, 3\}$ from (iii), therefore

$$\text{(iv)} \qquad T(\mathrm{EXP}(p) - \{p, p^R\}) \supset \{(p_1, p_t, p_2), t = n, n-1, \cdots, 3\}$$

because $\mathrm{LC}(p) \subset \mathrm{EXP}(p)$.

2. The only permutations in $\mathrm{EXP}''(p)$ which contain triples of the form (p_t, p_2, p_1) are

$$\text{(v)} \qquad p_{n-1}p_n(p_{n-2} \cdots p_2p_1) \quad \text{and} \quad p^R = p_{n-1}p_n(p_1 \cdots p_{n-2})^R.$$

(see Remark ii preceding Corollary 4.1).

On the other hand $p_2p_1 \,\mathrm{EXP}(p_3 \cdots p_n)$ contains all the triples of the form (p_2, p_1, p_t), $t \in \{3, \cdots, n\}$.

3. The only permutations in $\mathrm{EXP}'(p)$ which contain triples of the form (p_t, p_{n-1}, p_n) are

$$\text{(vi)} \qquad p \quad \text{and} \quad p_2p_1(p_3 \cdots p_{n-1}p_n) = l_1(p).$$

(See Remark (v) preceding Corollary 4.1 and recall the definition of $l_1(p)$ in § 3, Definition 3.3.)

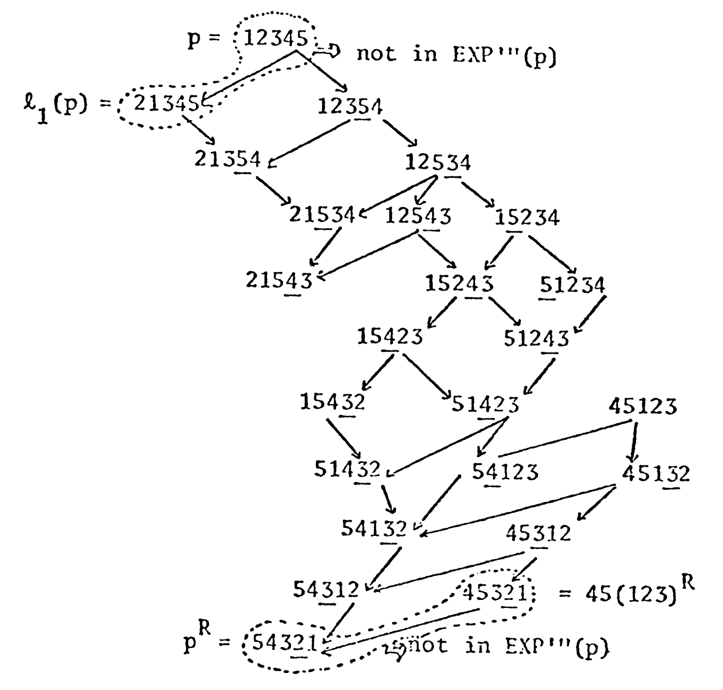
Also we have that $p_{n-1}p_n \,\mathrm{EXP}(p_1 \cdots p_{n-2})$ contains all the triples of the form (p_{n-1}, p_n, p_t) for $t \in \{1, 2, \cdots, n-2\}$, and it is clear that

$$\text{(vii)} \qquad T(\mathrm{EXP}(p) - \{p, p^R\}) \supset \{(p_n, p_t, p_{n-1}) \text{ for } t \in \{1, \cdots, n-2\}\}$$

because $p_n(p_1 \cdots p_{n-2})p_{n-1}$ is an element of $\mathrm{LC}(p) - \{p, p^R\} \subset \mathrm{EXP}(p) - \{p, p^R\}$.

4. (iv), (v), (vi) and (vii) together give us that

$$\mathrm{EXP}'(p) \cup \mathrm{EXP}''(p) - \{p_{n-1}p_n(p_1 \cdots p_{n-2})^R, l_1(p)\}$$



EXP''' (p) *with* $p=12345$

FIG. 6. *A maximal consistent set of cardinality* $\frac{3}{2}2^{n-1}-4$. • *Here* $n=5$; *then* $|\text{EXP}'''(12345)|=20$. • *The elements enclosed by dotted lines do not belong to* EXP''' (p).

(see Fig. 6) is a consistent subset of S_Σ; in fact we can prove that it is maximal if $|\Sigma|=n>4$ by following the same line of argument as the one given by Abello [1, Thm. 5.1].

THEOREM 4.2. *If* $p\in S_\Sigma$, *the set*

$$\text{EXP}'''(p)=\begin{cases}\text{EXP}'(p) & \textit{for } |\Sigma|=3,4,\\ (\text{EXP}'(p)-\{p, l_1(p)\})\cup p_{n-1}p_n\,\text{EXP}(p_1\cdots p_{n-2}) & \\ \qquad -p_{n-1}p_n(p_1\cdots p_{n-2})^R & \textit{for } |\Sigma|>4\end{cases}$$

is a maximal consistent subset of S_Σ *of cardinality*

$$|\text{EXP}'''(p)|=\begin{cases}\frac{5}{4}2^{|\Sigma|-1}-1 & \text{for } |\Sigma|=3,4,\\ \frac{3}{2}2^{|\Sigma|-1}-4 & \text{for } |\Sigma|>4.\end{cases}$$

TABLE 1

n	Biggest known consistent set cardinality	set
3	4 (this is the maximum)	EXP (p)
4	9 (this is the maximum)	EXP' (p)
5	20	EXP''' (p)
6	44	EXP''' (p)
7	92	EXP''' (p)
8	188	EXP''' (p)
9	380	EXP''' (p)

Conclusion. Any *maximum* consistent set $M \subset S_\Sigma$ must have

$$\text{Cardinality}\begin{cases} =\frac{5}{4}2^{|\Sigma|-1}-1 & \text{for } |\Sigma|=3,4, \\ \geqq \frac{3}{2}2^{|\Sigma|-1}-4 & \text{for } |\Sigma|>4. \end{cases} \qquad \text{Q.E.D.}$$

Table 1 gives some of the cardinalities of the consistent sets we have found which we believe are the best known values at present.

For $|\Sigma|>5$ an easy recurrence relation is: $|M_{n+1}|=2|M_n|+4$, where $|M_j|$ denote the cardinality of the biggest known consistent set for $|\Sigma|=j$.

Final notes.

1. Future work must be concentrated in finding upper bounds. At this point we are trying to prove that the cardinality of a consistent set must be less than or equal to 2^n. (This will have important implications for a complete characterization of the consistent sets from the point of view of Ward's and Inada's conditions.) In a following paper [9] some work in this direction will be presented.

2. A related question of interest is whether a consistent set of cardinality less than 2^n can be completed to a consistent set of cardinality 2^n.

3. Our results answer in general a question posed by Charles R. Johnson in 1978 about the cardinality of a maximum consistent set. He conjectured this to be 2^{n-1} [6].

4. Raynaud claims [8] that he has proved that the cardinality of a maximum consistent set which satisfies the "non- aumilieu" condition is 2^{n-1}. Our characterization of EXP (p) does not satisfy this condition even for a single triple of symbols, and yet EXP (p) is a maximal consistent set of cardinality 2^{n-1}. This suggests a path for a classification of the class of consistent sets.

Our results give an immediate construction of a subclass of consistent sets of bigger cardinality than those sets satisfying Inada's and Raynaud's conditions [4], [7]. Moreover, the constructed sets have a very nice structure which to our knowledge has not been found before.

5. The natural decision problem associated with maximum consistent sets has been untouched at present, and we suspect it does not belong to the class of NP-complete problems.

Acknowledgments. Thanks to Dr. Eugene Johnsen for his helpful comments and suggestions, to Ms. Ruth Cintron by her continuous encouragement and to Ms. Teddi Potter for the wonderful job typing the manuscript.

REFERENCES

[1] J. M. ABELLO, *Toward a maximum consistent set*, Technical report TRCS 11-81, Computer Science Dept., Univ. California, Santa Barbara, 1981.

[2] D. BLACK, *The Theory of Committees and Elections*, Cambridge Univ. Press, London, 1958.

[3] G. TH. GUILBAUD AND P. ROSENSTIEHL, *Analyse algebrique d'un scrutin*, Mathématiques et Sciences Humaines, numéro 4-1960.

[4] K. I. INADA, *A note on the simple majority decision rule*, Econometrica, 32 (1964), pp. 525–531.

[5] ———, *The simple majority decision rule*, Econometrica, 37 (1969), pp. 302–313.

[6] C. R. JOHNSON, *Remarks on mathematical social choice*, working paper 78-25, Dept. Economics and Institute for Physical Science and Technology, Univ. Maryland, College Park, 1978.

[7] H. RAYNAUD, *A proposition on some transitive conditions for the well-known majority rule*, Cahiers du Centre d'Etudes de Recherche Opérationnelle (1), 22 (1980), pp. 17–21. (In French.)

[8] ———, *Erratum*, Mathématiques et Sciences Humaines, numéro 73-1981, p. 122.

[9] J. M. ABELLO, *Intrinsic limitations of the majority rule, an algorithmic approach*, this Journal, 6 (1985), to appear.

SIAM J. ALG. DISC. METH.
Vol. 5, No. 4, December 1984

015

EXPLICIT SOLUTIONS TO OPTIMIZATION PROBLEMS ON THE INTERSECTIONS OF THE UNIT BALL OF THE l_1 AND l_∞ NORMS WITH A HYPERPLANE*

URIEL G. ROTHBLUM†

Abstract. In this paper we characterize the extreme points of intersections of the l_1 and l_∞ unit balls in R^n with a hyperplane. Such a characterization enables one to solve maximization problems of convex functions over these sets by enumeration. It turns out that the number of extreme points is $O(n^2)$ under the l_1 norm, and $O(2^N)$ under the l_∞ norm. In particular, solving corresponding optimization problems by enumeration is efficient only under the l_1 norm but not under the l_∞ norm. Still, we get explicit solutions for maximization problems of linear objectives on the intersection of the l_∞ unit ball with a hyperplane with a computational effort of $O(n)$. Applications of the results for computing bounds on the coefficients of ergodicity of square, nonnegative, irreducible matrices are discussed.

1. Introduction. In this paper we characterize the (finite set of) extreme points of the intersection of the l_1 and l_∞ unit balls in R^n with a hyperplane. It is well-known that every convex function over a compact, convex set attains a maximum at an extreme of that set. Hence, one can use a characterization of the extreme points of a compact, convex set having finitely many extreme points to solve, by enumeration, optimization problems where a convex function is maximized over the set. In particular, such an enumeration can be used for corresponding optimization problems with linear objectives.

The enumeration of the extreme points need not yield an efficient solution method when the number of extreme points is very large. Our characterization of the intersection of the l_1 unit ball in R^n with a hyperplane demonstrates that there are $n(n-1)$ extreme points; assuring that the solution of the corresponding optimization problems by enumeration can be done with reasonable computational effort. On the other hand, our characterization of the intersection of the l_∞ unit ball in R^n with a hyperplane demonstrates the number of extreme points is of the order of 2^n. (When the hyperplane consists of all points in R^n whose coordinate sum is zero, the number of the extreme points is $\binom{n}{k}$ where k is the largest integer smaller or equal $n/2$. In general, we have no closed expression for the number of corresponding extreme points; but, our characterization indicates that it is of the same order as the number of subsets of $\{1, \cdots, n\}$, a number which equals 2^n.) It follows that the solution of the corresponding optimization problems by enumeration of the extreme points is not a reasonable computational method. Still, we were able to obtain explicit optimal solutions to maximization problems of a linear objective over the intersection of the l_∞ unit ball in R^n with a hyperplane by a method whose computational effort is $O(n)$.

Our interest in optimization problems where a convex function is maximized over the intersection of unit balls under various norms with a hyperplane was startled by the study of bounds on coefficients of ergodicity of square, nonnegative, irreducible matrices. Specifically, let P be such a (square, nonnegative, irreducible) matrix and

* Received by the editors July 6, 1983, and in revised form December 31, 1983. This research was supported by the National Science Foundation under grant ENG-78-25182 and the Office of Naval Research under contract N00014-77-C-0518.

† Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa 32000, Israel.

let ρ be its *spectral radius*, i.e., ρ is the largest modulus of the eigenvalues of P. The Perron–Frobenius theorem (e.g., Gantmacher (1957)) assures that ρ is a simple positive eigenvalue of P, having a positive right eigenvector, which we denote w. The second largest modulus of the eigenvalues of P is called the *coefficient of ergodicity* of P and will be denoted $\xi(P)$. It is known to determine convergence properties of the sequence $\rho^{-N}P^N$ as $N\rightarrow\infty$ (e.g., Rothblum and Tan (1983, § 2)). Thus, upper bounds on the coefficient of ergodicity yield information about the convergence of the corresponding sequence of powers. Such bounds have been the focus of many studies (e.g., Rothblum and Tan (1983) and the list of references of that paper). Following the approach of Seneta (1979), (1983a), (1983b), Tan (1982), (1983), (1984) and Seneta and Tan (1983), it was shown in Rothblum and Tan (1983, Thm. 3.1) that for every norm $\| \ \|$ on R^n,

$$\xi(P)\leqq \max_{\substack{\|x\|\leqq 1\\ x^Tw=0\\ x\in R^n}} \|x^TP\|. \tag{1.1}$$

We remark that if $\|\cdot\|$ is a norm on C^n and the condition $x\in R^n$ in (1.1) is replaced by $x\in C^n$, then the corresponding (weaker) version of (1.1) is straightforward (e.g., Seneta (1979, p. 579) or Rothblum and Tan (1983, § 1)). The fact that the right-hand side of (1.1) is the maximum of a *real* optimization problem allows one to use mathematcial programming to compute these bounds. In particular, we observe that the function defined by $x\rightarrow\|x^TP\|$ is convex on the set $\mathscr{C}\equiv\{x\in R^n\colon \|x\|\leqq 1,\ x^Tw=0\}$, which is the intersection of the unit ball $\{x\in R^n\colon\|x\|\leqq 1\}$ with the hyperplane $\{x\in R^n\colon x^Tw=0\}$. We conclude that

$$\max_{x\in\mathscr{C}}\|x^TP\|=\max_{x\in\mathscr{F}}\|x^TP\|, \tag{1.2}$$

where $\mathscr{F}$ is the set of extreme points of $\mathscr{C}$. In particular, our characterization of the extreme points of the intersection of the l_1 unit ball with a hyperplane enable us to compute the right-hand side of (1.1) efficiently when the norm $\|\cdot\|$ is the l_1 norm. When the norm $\|\cdot\|$ is the l_∞ norm we have that, with P_i denoting the ith column of P,

$$\max_{x\in\mathscr{F}}\|x^TP\|_\infty=\max_{x\in\mathscr{F}}\max_{i=1,\cdots,n}|x^TP_i|=\max_{i=1,\cdots,n}\max_{x\in\mathscr{F}}x^TP_i, \tag{1.3}$$

where we used the symmetry of the set $\mathscr{F}$. Thus, the right-hand side of (1.1) can be computed by solving n optimization problems where a linear objective is maximized over the intersection of the l_∞ unit ball with a hyperplane. Our results show how to obtain these solutions explicitly. We remark that the explicit representation of the solutions to the right-hand side of (1.1) under the l_1 and l_∞ norms appear in Rothblum and Tan (1983, § 6).

The organization of this paper is as follows. We summarize some notational conventions in § 2. The characterization of the extreme points of the intersection of the l_1 unit ball with a hyperplane is given in § 3, whereas, the characterization of the extreme points of the intersections of the l_∞ unit ball with a hyperplane and the explicit solution of corresponding optimization problems with linear objectives are given in § 4. Finally, the Appendix contains explicit methods for solving the latter problems where the corresponding computational effort is $O(n\lg n)$ and $O(n)$, respectively.

2. Notational conventions. Throughout this paper we consider elements in, and subsets of, R^n for some fixed positive integer n. As usual, and as done in the introduction, the l_1 *norm* and l_∞ *norm* of a vector $x \in R^n$ will be denoted $\|x\|_1$ and $\|x\|_\infty$, respectively, i.e., $\|x\|_1 = \sum_{i=1}^n |x_i|$ and $\|x\|_\infty = \max_{1 \leqq i \leqq n} |x_i|$. The *convex hull* of a subset $\mathscr{B} \subseteq R^n$, denoted conv $\mathscr{B}$, is the set of all convex combinations of elements of $\mathscr{B}$, i.e., $\{\sum_{i=1}^n \alpha_i x^i \colon \alpha_i \geqq 0,\ x^i \in \mathscr{B},\ \sum_{i=1}^n \alpha_i = 1\}$. A set $\mathscr{B} \subseteq R^n$ is called *convex* if $\mathscr{B} = \operatorname{conv} \mathscr{B}$. An *extreme point* of a convex set $\mathscr{B} \subseteq R^n$ is a vector $x \in \mathscr{B}$ such that if $x = \alpha y + (1-\alpha) z$ for $y, z \in \mathscr{B}$ and $0 < \alpha < 1$, then necessarily $x = y = z$. Evidently, if x is an extreme point of a convex set $\mathscr{B}$ and $x \in \operatorname{conv} \mathscr{A}$ for some $\mathscr{A} \subseteq \mathscr{B}$, then $x \in \mathscr{A}$.

A vector $x \in R^n$ is called *nonnegative*, denoted $x \geqq 0$, if all the coordinates of x are nonnegative. A vector $x \in R^n$ is called *positive*, denoted $x \gg 0$, if all the coordinates of x are positive. Finally, a vector $x \in R^n$ is called *semipositive*, denoted $x > 0$, if $x \geqq 0$ and $x \neq 0$.

For a set $S \subseteq \{1, \cdots, n\}$, let e^S be the vector in R^n defined by $(e^S)_i = 1$ for $i \in S$ and $(eS)_i = 0$ for $i \in \{1, \cdots, n\} \backslash S$. In particular, $e^{\varnothing} = 0$. Of course if S consists of a single element, say j, e^S is the jth *unit vector* in R^n which we denote e^j, i.e., $e^{\{j\}} = e^j$.

3. The l_1 norm. The purpose of this section is to characterize the extreme points of the intersection of the l_1 unit ball in R^n with a hyperplane. We then use the results to give an explicit expression for the maximum of a convex function over such sets. The characterization of the extreme points of the corresponding sets when the hyperplane is the set of vectors in R^n whose coordinate-sum is zero, is implicit in Rothblum and Schneider (1980, Thm. 1).

THEOREM 3.1. *Let $0 \neq u \in R^n$ and let*

$$\mathscr{C} = \{x \in R^n \colon \|x\|_1 \leqq 1,\ x^T u = 0\}. \tag{3.1}$$

Then $\mathscr{C}$ is a compact convex set. Also, if $n = 1$, then $\mathscr{C} = \{0\}$ and if $n > 1$ then the set of extreme points of $\mathscr{C}$ is the set

$$\mathscr{F} = \{(|u_i| + |u_j|)^{-1}(u_j e^i - u_i e^j) \colon i, j = 1, \cdots, n, |u_i| + |u_j| \neq 0 \textit{ and } i \neq j\}. \tag{3.2}$$

Proof. The compactness and convexity of $\mathscr{C}$, as well as the fact that $\mathscr{C} = \{0\}$ when $n = 1$, are straightforward. Henceforth we will assume that $n > 1$. In this case $\mathscr{C} \neq \phi$ and $\mathscr{F} \neq \phi$. For notational convenience, for $i, j = 1, \cdots, n$ where $|u_i| + |u_j| \neq 0$ and $i \neq j$, we denote the vector $(|u_i| + |u_j|)^{-1}(u_j e^i - u_i e^j)$ by f^{ij}. Evidently, $\mathscr{F} = \mathscr{F}_2 \cup \mathscr{F}_1$ where

$$\mathscr{F}_2 \equiv \{f^{ij} \colon i, j = 1, \cdots, n,\ u_i u_j \neq 0 \text{ and } i \neq j\},$$

and

$$\mathscr{F}_1 \equiv \bigcup \{\{e^i, -e^i\} \colon i = 1, \cdots, n,\ u_i \neq 0\}.$$

We note that $\mathscr{F} \subseteq \mathscr{C}$ and for every $f \in \mathscr{F}$, $-f \in \mathscr{F}$ and $\|f\|_1 = 1$. Let $\mathscr{E}$ be the set of extreme points of $\mathscr{C}$. We will establish that $\mathscr{E} = \mathscr{F}$ by showing that $\mathscr{E} \subseteq \mathscr{F}$ and that $\mathscr{F} \subseteq \mathscr{E}$.

We will next establish that $\mathscr{E} \subseteq \mathscr{F}$. As $\mathscr{F} \subseteq \mathscr{C}$, the definition of extreme points of convex sets assures that it suffices to show that $\mathscr{E} \subseteq \operatorname{conv} \mathscr{F}$. We will prove, more generally, that $\mathscr{C} \subseteq \operatorname{conv} \mathscr{F}$. Specifically, we will show that if $x \in \mathscr{C}$, then $x \in \operatorname{conv} \mathscr{F}$. Our proof follows by induction on the number of nonzero coordinates of x, which we denote $\upsilon(x)$. First observe that if $\upsilon(x) = 0$, then $x = 0$ and for every element f in the nonempty set $\mathscr{F}$, $x = 2^{-1} f + 2^{-1}(-f) \in \operatorname{conv} \mathscr{F}$ (recall that if $f \in \mathscr{F}$, then $-f \in \mathscr{F}$). Next assume that $\upsilon(x) = 1$. Then x has the form αe^i for some $\alpha \neq 0$ and $i = 1, \cdots, n$. As $\alpha u_i = x^T u = 0$, we conclude that $u_i = 0$ and therefore both e^i and $-e^i$ are in $\mathscr{F}_1 \subseteq \mathscr{F}$. Hence, as $|\alpha| = \|\alpha e^i\|_1 = \|x\|_1 \leqq 1$, we have that

$$x = \alpha e^i = [2^{-1}(\alpha + 1)] e^i + [2^{-1}(1 - \alpha)](-e^i) \in \operatorname{conv} \mathscr{F}.$$

Next, assume that $\upsilon(x)=2$. Then x has the form $\beta e^i+\gamma e^j$ where $i, j=1,\cdots,n$, $i\neq j$ and $\beta\gamma\neq 0$. Evidently $|\beta|+|\gamma|=\|x\|_1\leqq 1$ and $\beta u_i+\gamma u_j=x^Tu=0$. We conclude that either $u_i=u_j=0$ or $u_iu_j\neq 0$. In the case where $u_i=u_j=0$, let $y\equiv[(|\beta|+|\gamma|)\beta/|\beta|]e^i$ and $z\equiv[(|\beta|+|\gamma|)\gamma/|\gamma|]e^j$. Evidently, $\|y\|_1=\|z\|_1=|\beta|+|\gamma|\leqq 1$. As $y^Tu=z^Tu=0$, we have that both y and z are in $\mathscr{C}$. As $\upsilon(y)=\upsilon(z)=1$ we conclude from the established induction assertion for the case where $\upsilon(\cdot)$ equals one, that both y and z are in conv $\mathscr{F}$. Hence, by the convexity of conv $\mathscr{F}$,

$$x=\beta e^i+\gamma e^j=[|\beta|/(|\beta|+|\gamma|)]y+[|\gamma|/(|\beta|+|\gamma|)]z\in\text{conv }\mathscr{F}.$$

In the remaining case, where $u_iu_j\neq 0$, both f^{ij} and $-f^{ij}=f^{ji}$ are in $\mathscr{F}_2\subseteq\mathscr{F}$. Let $\alpha\equiv\beta u_j^{-1}(|u_i|+|u_j|)$. As $\beta u_i+\gamma u_j=0$, we have that

$$\alpha f^{ij}=\beta u_j^{-1}(u_je^i-u_ie^j)=\beta e^i-\beta u_iu_j^{-1}e^j=\beta e^i+\gamma e^j=x.$$

As $0\leqq|\alpha|=|\alpha|\|f^{ij}\|_1=\|\alpha f^{ij}\|_1=\|x\|\leqq 1$, we conclude that

$$x=\alpha f^{ij}=[2^{-1}(\alpha+1)]f^{ij}+[2^{-1}(1-\alpha)](-f^{ij})\in\text{conv }\mathscr{F}.$$

Next assume that for some integer $t\geqq 3$, every $x\in\mathscr{C}$ with $\upsilon(x)<t$ is contained in conv $\mathscr{F}$, and consider $x\in\mathscr{C}$ with $\upsilon(x)=t$. If there exists an integer $q=1,\cdots,n$ with $x_q\neq 0$ and $u_q=0$, let

$$y\equiv x-x_qe^q.$$

Evidently, $\upsilon(y)=\upsilon(x)-1\geqq 2$ and $y^Tu=0$. Hence $y\neq 0$ and $y/\|y\|_1\in\mathscr{C}$. Next observe that $\upsilon(y/\|y\|_1)=\upsilon(y)<\upsilon(x)=t$, and therefore the induction hypothesis assures that $y/\|y\|_1\in\text{conv }\mathscr{F}$. Also observe that $0\leqq\|y\|_1=\|x\|_1-|x_q|\leqq 1-|x_q|<1$. Hence, $z\equiv[x_q/(1-\|y\|_1)]e^q$ satisfies $\|z\|_1=|x_q|/(1-\|y\|_1)\leqq 1$. As $z^Tu=0$, we have that $z\in\mathscr{C}$. Hence, as $\upsilon(z)=1$, the induction hypothesis implies that $z\in\text{conv }\mathscr{F}$. Now, the convexity of conv $\mathscr{F}$ and the fact that $0\leqq\|y\|_1\leqq 1$ assure that

$$x=y+x_qe^q=\|y\|_1(y/\|y\|_1)+(1-\|y\|_1)z\in\text{conv }\mathscr{F}.$$

Next assume that $u_i\neq 0$ for every $i=1,\cdots,n$ with $x_i\neq 0$. Let $\min\{|x_iu_i|\colon i=1,\cdots,n, x\neq 0\}=|x_qu_q|(>0)$. Since $x^Tu=0$, there exists some $r\in\{1,\cdots,n\}\backslash\{q\}$ with $(x_ru_r)(x_qu_q)<0$. As $u_qu_r\neq 0$ we have that $f^{qr}\in\mathscr{F}_2\subseteq\mathscr{F}$. Let

$$\beta=x_q/(f^{qr})_q\neq 0$$

and

$$y\equiv x-\beta f^{qr}.$$

Evidently, $\upsilon(y)\geqq\upsilon(x)-2\geqq 1$ and $y^Tu=0$. Hence, $y\neq 0$ and $y/\|y\|_1\in\mathscr{C}$. Next observe that $\upsilon(y/\|y\|_1)=\upsilon(y)<\upsilon(x)=t$, and therefore the induction hypothesis assures that $y/\|y\|_1\in\text{conv }\mathscr{F}$. Also observe that the selection of q and r implies that

$$\beta(f^{qr})_r/x_r=x_q(f^{qr})_r/x_r(f^{qr})_q=-(x_qu_q)/(x_ru_r)>0$$

and that

$$|\beta(f^{qr})_r/x_r|=|x_qu_q/x_ru_r|=|x_qu_q|/|x_ru_r|\leqq 1.$$

It follows that

$$|y_r|=|x_r-\beta(f^{qr})_r|=|x_r|-|\beta|\,|(f^{qr})_r|.$$

In particular,

$$0 \leqq \|y\|_1 = \|x\|_1 - (|x_q| + |x_r - y_r|) = \|x\|_1 - |\beta|[|(f^{qr})_q| + |(f^{qr})_r|]$$
$$= \|x\|_1 - |\beta|\|f^{qr}\|_1 = \|x\|_1 - |\beta| \leqq 1 - |\beta| < 1.$$

Hence, $z \equiv [\beta/(1-\|y\|_1)]f^{qr}$ satisfies $\|z\|_1 = |\beta|/(1-\|y\|_1) \leqq 1$. As $f^{qr} \in \mathscr{F} \subseteq \mathscr{C}$ we also have that $z^T u = 0$, assuring that $z \in \mathscr{C}$. As $v(z) = v(f^{qr}) = 2$, we conclude from the induction hypothesis that $z \in \operatorname{conv} \mathscr{F}$. Hence, the convexity of conv $\mathscr{F}$ and the fact that $0 \leqq \|y\|_1 \leqq 1$ assure that

$$x = y + \beta f^{qr} = \|y\|_1(y/\|y\|_1) + (1-\|y\|_1)z \in \operatorname{conv} \mathscr{F},$$

completing our proof that $\mathscr{C} \subseteq \operatorname{conv} \mathscr{F}$.

We next show that $\mathscr{F} \subseteq \mathscr{E}$. Let $f \in \mathscr{F}$ have a representation $f = \alpha x + (1-\alpha)y$, where x, $y \in \mathscr{C}$ and $0 < \alpha < 1$. We will show that $x = y = f$. As

$$1 = \|f\|_1 = \sum_{i=1}^n |f_i| = \sum_{i=1}^n |\alpha x_i + (1-\alpha)y_i| \leqq \sum_{i=1}^n (\alpha|x_i| + (1-\alpha)|y_i|) = \alpha\|x\|_1 + (1-\alpha)\|y\|_1$$
$$\leqq \alpha + 1 - \alpha = 1,$$

we conclude that for $i = 1, \cdots, n$, $|\alpha x_i + (1-\alpha)y_i| = \alpha|x_i| + (1-\alpha)|y_i|$, or equivalently, $x_i y_i \geqq 0$. First assume that $f \in \mathscr{F}_1$, i.e., $f = e^q$ or $f = -e^q$ for some $q = 1, \cdots, n$ where $u_q = 0$. Then, for $i \in \{1, \cdots, n\}\backslash\{q\}$, $0 = (e^q)_i = \alpha x_i + (1-\alpha)y_i$, and therefore, as $x_i y_i \geqq 0$, we conclude that $x_i = y_i = 0$. It follows that the only nonzero coordinate of both x and y is the q coordinate. Thus, $x = \gamma f$ and $y = \delta f$ for some scalars γ and δ. Evidently, $|\gamma| = \|x\|_1 \leqq 1$ and $|\delta| = \|y\|_1 \leqq 1$. As $f = \alpha x + (1-\alpha)y = [\alpha\delta + (1-\alpha)\delta]f$, we conclude that $\alpha\gamma + (1-\alpha)\delta = 1$, implying that $\gamma = \delta = 1$, i.e., $x = y = f$. It remains to consider the case where $f \in \mathscr{F}_2$, i.e., $f = f^{qr}$ for some q, $r \in \{1, \cdots, n\}$ where $u_q u_r \neq 0$ and $q \neq r$. For $i \in \{1, \cdots, n\}\backslash\{q, r\}$, $0 = (f^{qr})_i = \alpha x_i + (1-\alpha)y_i$, and therefore, as $x_i y_i \geqq 0$, we conclude that $x_i = y_i = 0$. It follows that the only possible nonzero coordinates of x and of y are the q and r coordinates. As x, $y \in \mathscr{C}$ and $u_q u_r \neq 0$ it follows that both x and y are proportional to $f^{qr} = f$, i.e., $x = \gamma f$ and $y = \delta f$ for some scalars γ and δ. The remainder of the proof follows the argument used in the case where $f \in \mathscr{F}_1$. □

We observe that if $u = 0$, then the set $\mathscr{C}$ defined by (3.1) is the unit ball under the l_1 norm, whose extreme points are known to be the unit vectors.

We next specialize Theorem 3.1 to the case where u is positive. It turns out that the corresponding expressions are simpler for this special case.

COROLLARY 3.2. *Let $n > 1$ and let $\mathscr{C}$ be given by* (3.1) *where $u \in R^n$ is a positive vector. Then the set of extreme points of $\mathscr{C}$ is the set*

$$\mathscr{F} \equiv \{(u_i + u_j)^{-1}(u_j e^i - u_i e^j)\colon i, j = 1, \cdots, n, i \neq j\}. \tag{3.3}$$

Moreover, if $u = e \equiv (1, \cdots, 1)^T \in R^n$, then the set of extreme points of $\mathscr{C}$ is the set

$$\mathscr{F} \equiv \{2^{-1}(e^i - e^j)\colon i, j = 1, \cdots, n, i \neq j\}. \tag{3.4}$$

The characterization of the extreme points of the set $\mathscr{C}$ (defined by (3.1) with respect to a vector $u \in R^n$) enables us to solve optimization problems where a convex function is maximized over the corresponding set.

COROLLARY 3.3. *Let $0 \neq u \in R^n$ and let $\mathscr{C}$ and $\mathscr{F}$ be given as in Theorem* 3.1, *where $n > 1$. Let h be a real valued convex function defined on $\mathscr{C}$. Then*

$$\max_{x \in \mathscr{C}} h(x) = \max_{x \in \mathscr{F}} h(x). \tag{3.5}$$

Proof. The conclusion of the corollary follows directly from Theorem 3.1 and the (well-known) fact that a convex function on a compact convex set attains a maximum at one of the extreme points of the set. □

4. The l_∞ norm. The purpose of this section is to characterize the extreme points of the intersection of the l_∞ unit ball in R^n with a hyperplane, and to give an explicit expression for the maximum of convex and linear functions over these sets. The characterization of the extreme points of the corresponding sets when the hyperplane is the set of vectors in R^n whose coordinate-sum is zero is implicit in Haviv and van der Heyden (1983).

THEOREM 4.1. *Let $u \in R^n$ and let*

$$\mathscr{C} = \{x \in R^n \colon \|x\|_\infty \leqq 1, x^T u = 0\}. \tag{4.1}$$

Then $\mathscr{C}$ is a compact convex set. Moreover, the set of extreme points of $\mathscr{C}$ is the set $\mathscr{F} \equiv \mathscr{F}_1 \cup \mathscr{F}_0$, where

$$\mathscr{F}_1 \equiv \left\{ e^S + \beta e^k - e^{S^c} \colon \phi \neq S \subseteq \{1, \cdots, n\}, k \in S, -2 < \beta \leqq 0, \right. \\ \left. \sum_{i \in S} u_i + \beta u_k - \sum_{i \in S^c} u_i = 0 \ \textit{where}\ \beta = 0 \ \textit{if}\ u_k = 0 \right\}, \tag{4.2}$$

and

$$\mathscr{F}_0 = \begin{cases} \phi & \textit{if}\ \sum_{i=1}^n u_i \neq 0, \\ \{-e^{\{1,\cdots,n\}}\} & \textit{if}\ \sum_{i=1}^n u_i = 0. \end{cases} \tag{4.3}$$

Proof. The compactness and convexity of $\mathscr{C}$ are straightforward. Let $\mathscr{E}$ be the set of extreme points of $\mathscr{C}$. We will show that $\mathscr{E} = \mathscr{F}$ by showing that $\mathscr{E} \subseteq \mathscr{F}$ and $\mathscr{F} \subseteq \mathscr{E}$.

We will next establish that $\mathscr{E} \subseteq \mathscr{F}$. As $\mathscr{F} \subseteq \mathscr{C}$, the definition of extreme points of convex sets implies that it suffices to show that $\mathscr{E} \subseteq \operatorname{conv} \mathscr{F}$. We will next prove that $\mathscr{C} \subseteq \operatorname{conv} \mathscr{F}$. Specifically, we will show that every $x \in \mathscr{C}$ is necessarily in $\operatorname{conv} \mathscr{F}$. Our proof follows by induction on the number of coordinates of x whose absolute value is less than one, a number which we denote $\mu(x)$. Now, if $\mu(x) = 0$, then $x = e^S - e^{S^c}$ for $S = \{i = 1, \cdots, n \colon x_i = 1\}$. It follows that if $S = \phi$, then $x = -e^{\{1,\cdots,n\}}$, and if $S \neq \phi$, then $x = e^S + \beta e^k - e^{S^c}$ for $\beta = 0$ and any $k \in S$. In the former case, $\sum_{i=1}^n u_i = -x^T u = 0$, assuring that $x \in \mathscr{F}_0 \subseteq \mathscr{F}$; and in the latter case, $\sum_{i \in S} u_i + \beta u_k - \sum_{i \in S^c} u_i \equiv x^T u = 0$, assuring that $x \in \mathscr{F}_1 \subseteq \mathscr{F}$. In either case, we conclude that $x \in \mathscr{F} \subseteq \operatorname{conv} \mathscr{F}$. Next assume that $\mu(x) = 1$. In this case $x = e^S + \beta e^k - e^{S^c}$ for $S = \{i = 1, \cdots, n \colon x_i > -1\}$, k being the (unique) index with $|x_k| < 1$ and $\beta = x_k - 1$. Evidently, $-2 < \beta < 0$ and, as $x \in \mathscr{C}$, $\sum_{i \in S} u_i + \beta u_k - \sum_{i \in S^c} u_i = x^T u = 0$. Hence, if $u_k \neq 0$, then $x \in \mathscr{F} \subseteq \operatorname{conv} \mathscr{F}$. In the remaining case $u_k = 0$, in which case let $y \equiv x - \beta e^k = e^S - e^{S^c} \in \mathscr{F}$ and $z \equiv x + (-2 - \beta) e^k = e^{S \setminus \{k\}} - e^{(S \setminus \{k\})^c} \in \mathscr{F}$ (the case where $S = \{k\}$, requiring a special argument). Hence, as $0 < 1 + 2^{-1}\beta < 1$, we have that

$$\begin{aligned} x &= (1 + 2^{-1}\beta)[x - \beta e^k] + (-2^{-1}\beta)[x + (-2 - \beta) e^k] \\ &= (1 + 2^{-1}\beta) y + (-2^{-1}\beta) z \in \operatorname{conv} \mathscr{F}. \end{aligned}$$

Next assume that for some integer $t \geqq 2$, every $x \in \mathscr{C}$ with $\mu(x) < t$ is contained in $\operatorname{conv} \mathscr{F}$, and consider $x \in \mathscr{C}$ with $\mu(x) = t$. As $\mu(x) = t \geqq 2$, there exist $q, r \in \{1, \cdots, n\}$ with $|x_q| < 1$, $|x_r| < 1$ and $q \neq r$. We first consider the case where $u_q = 0$. In this case,

$y \equiv x+(1-x_q)e^q$ and $z \equiv x+(-1-x_q)e^q$ are in $\mathscr{C}$ and $\mu(y)=\mu(z)=\mu(x)-1<t$. Hence, by the induction hypothesis, $y \in \operatorname{conv} \mathscr{F}$ and $z \in \operatorname{conv} \mathscr{F}$. Thus, the convexity of $\operatorname{conv} \mathscr{F}$ and the fact that $-1<x_q<1$, assure that

$$\begin{aligned} x &= [2^{-1}(1+x_q)][x+(1-x_q)e^q]+[2^{-1}(1-x_q)][x+(-1-x_q)e^q] \\ &= [2^{-1}(1+x_q)]y+[2^{-1}(1-x_q)]z \in \operatorname{conv} \mathscr{F}. \end{aligned}$$

A similar argument shows that $x \in \operatorname{conv} \mathscr{F}$ in the case where $u_r=0$. It remains to consider the case where $u_q \neq 0$ and $u_r \neq 0$. In this case, let

$$\alpha_q = \begin{cases} u_q(1-x_q) & \text{if } u_q>0, \\ u_q(-1-x_q) & \text{if } u_q<0. \end{cases}$$

Also, let

$$\alpha_r = \begin{cases} u_r(1+x_r) & \text{if } u_r>0, \\ u_r(-1+x_r) & \text{if } u_r<0. \end{cases}$$

For $p=q$ or $p=r$, $\alpha_p>0$ and for $0 \leqq \alpha < \alpha_p |x_p+\alpha u_p^{-1}| \leqq 1$ with equality holding if $\alpha=\alpha_p$. Let $\alpha_{qr}=\min\{\alpha_q, \alpha_r\}$ and let $y \equiv x+\alpha_{qr}(u_q^{-1}e^q-u_r^{-1}e^r)$. Then $\alpha_{qr}>0$, $|(y^{qr})_q| = |x_q+\alpha_{qr}u_q^{-1}| \leqq d$ and $|(y^{qr})_r|=|x_r+\alpha_{qr}u_r^{-1}| \leqq 1$. Moreover, either $|(y^{qr})_q|=1$ or $|(y^{qr})_r|= 1$, depending whether $\alpha_{qr}=\alpha_q$ or $\alpha_{qr}=\alpha_r$. As $(y^{qr})_i=x_i$ for $i \in \{1, \cdots, n\} \backslash \{q, r\}$, we conclude that $\|y^{qr}\|_\infty \leqq 1$ and $\mu(y^{qr})<\mu(x)=t$. Also observe that $(y^{qr})^T u = x^T u = 0$. Thus, $y^{qr} \in \mathscr{C}$. As $\mu(y^{qr})<t$, the induction hypothesis assures that $y^{qr} \in \operatorname{conv} \mathscr{F}$. Next, let α_{rq} and y^{rq} be defined correspondingly by reversing the roles of q and r. The above arguments show that $y^{rq} \in \operatorname{conv} \mathscr{F}$. We conclude from the convexity of $\operatorname{conv} \mathscr{F}$

$$\begin{aligned} x &= (\alpha_{rq}+\alpha_{qr})^{-1}(\alpha_{rq}x+\alpha_{qr}x) \\ &= (\alpha_{rq}+\alpha_{qr})^{-1}\{\alpha_{rq}[x+\alpha_{qr}(u_q^{-1}e^q-u_r^{-1}e^r)]+\alpha_{qr}[x+\alpha_{rq}(u_r^{-1}e^r-u_q^{-1}e^q)]\} \\ &= (\alpha_{rq}+\alpha_{qr})^{-1}\{\alpha_{rq}y^{qr}+\alpha_{qr}y^{rq}\} \in \operatorname{conv} \mathscr{F}, \end{aligned}$$

completing our proof that $\mathscr{C} \subseteq \operatorname{conv} \mathscr{F}$.

It remains to show that $\mathscr{C} \subseteq \mathscr{F}$. Let $f \in \mathscr{F}$ have a representation $f=\alpha x+(1-\alpha)y$, where x, $y \in \mathscr{C}$ and $0<\alpha<1$. We will show that $x=y=f$. As $\|x\|_\infty \leqq 1$, $\|y\|_\infty \leqq 1$ and

$$|f_i|=|\alpha x_i+(1-\alpha)y_i| \leqq \alpha|x_i|+(1-\alpha)|y_i| \leqq \alpha+1-\alpha=1,$$

we conclude that whenever $|f_i|=1$ we have that $|x_i|=|y_i|=1$ and $|\alpha x_i+(1-\alpha)y_i|= \alpha|x_i|+(1-\alpha)|y_i|$. Since the latter condition is equivalent to $x_iy_i \geqq 0$, we conclude that if $|f_i|=1$ then $x_i=y_i=f_i$. In the first of two cases assume that $f \in \mathscr{F}_0$. Then for $i \in \{1, \cdots, n\}$, $|f_i|=1$, assuring that $x_i=y_i=f_i$. Thus, $f=x=y$. In the remaining case we have that $f \in \mathscr{F}_1$. Thus f has the representation $f=e^S+\beta e^k-e^{S^c}$, where $\phi \neq S \subseteq \{1, \cdots, n\}$, $k \in S$, $-2<\beta \leqq 0$ and $\sum_{i \in S} u_i+\beta u_k-\sum_{i \in S^c} u_i=0$ and where $\beta=0$ if $u_k=0$. Now, if $u_k=0$, then for $i \in \{1, \cdots, n\}$, $|f_i|=1$, assuring that $x_i=y_i=f_i$. Thus, $x=y=f$. Next assume that $u_k \neq 0$. Then, for $i \in \{1, \cdots, n\} \backslash \{k\}$, $|f_i|=1$, assuring that $f_i=x_i=y_i$. Next observe that, as $\{f, x, y\} \subseteq \mathscr{C}$, we have that $0=\sum_{i=1} u_if_i=\sum_{i=1}^n u_ix_i=\sum_{i=1}^n u_iy_i$. We conclude that $u_kf_k=u_kx_k=u_ky_k$ and there, as $u_k \neq 0$, $f_k=x_k=y_k$. Thus, again, $x=y=f$.

COROLLARY 4.2. *Let $u \in R^n$ and let $\mathscr{C}$ and $\mathscr{F}$ be defined as in Theorem 4.1. Let h be a real valued convex function defined on $\mathscr{C}$. Then*

$$\max_{x \in \mathscr{C}} h(x) = \max_{x \in \mathscr{F}} h(x). \tag{4.4}$$

Proof. The conclusion of our corollary follows directly from Theorem 4.1 and the (well-known) fact that a convex function on a compact convex set attains a maximum at one of the extreme points of the set. □

Let $\mathscr{C}$ be defined through (4.1) with respect to a given vector u. The characterization of the extreme points of $\mathscr{C}$, obtained in Theorem 4.1, suggests a simple method for generating these extreme points. We first introduce some notation. Let $\phi \subseteq S \subseteq T \subseteq \{1, \cdots, n\}$. We define the quantity $u^{(S,T)}$ by

$$u^{(S,T)} \equiv \sum_{i \in S} u_i - \sum_{i \in T \setminus S} u_i.$$

Also, we will use the notation P, N and M, respectively, to denote the subsets of $\{1, \cdots, n\}$ consisting of all indices i with $u_i > 0$, $u_i < 0$ and $u_i = 0$. Finally, let $\mathscr{E}$ be the set of extreme points of $\mathscr{C}$.

We next describe the method for generating the extreme points of $\mathscr{C}$. Consider a set $V \subseteq N$ for which

$$-\sum_{i \in P} u_i \leqq u^{(V,N)} \leqq \sum_{i \in P} u_i. \tag{4.5}$$

(Of course, $V = \phi \subseteq N$ is such a set). Now, if $u^{(V,N)} = -\sum_{i \in P} u_i$, then Theorem 4.1 assures that if $U \subseteq M$ and $S \equiv V \cup P \cup U$, then $e^S - e^{S^c} \in \mathscr{E}$. In the remaining case where $-\sum_{i \in P} u_i < u^{(V,N)} \leqq \sum_{i \in P} u_i$, one can find a set $\phi \neq W \subseteq P$ and an index $k \in W$, such that

$$u^{(W \setminus \{k\},P)} < u^{(V,N)} \leqq u^{(W,P)}. \tag{4.6}$$

In fact, for any enumeration of the elements of P, say $i(1), \cdots, i(p)$, there exists an integer $r \in \{1, \cdots, p\}$ for which $W = \{i(1), \cdots, i(r)\}$ and $k = i(r)$ satisfy (4.6). It now follows from Theorem 4.1 that if $U \subseteq M$, $S \equiv V \cup W \cup U$ and $\beta \equiv -[u^{(V,N)} + u^{(W,P)}]u_k^{-1}$ then $e^S + \beta e^k - e^{S^c} \in \mathscr{E}$. Of course, an alternative method for constructing extreme points of $\mathscr{C}$ is to first select $W \subseteq P$ with

$$\sum_{i \in N} u_i \leqq u^{(W,P)} < -\sum_{i \in N} u_i, \tag{4.7}$$

and then select the corresponding set $\phi \neq V \subseteq N$ and index $k \in V$ such that

$$u^{(V,N)} \leqq u^{(W,P)} < u^{(V \setminus \{k\},N)}. \tag{4.8}$$

It follows, again, from Theorem 4.1, that if $U \subseteq M$, $S \equiv W \cup V \cup U$ and $\beta \equiv -[u^{(W,P)} + u^{(V,N)}]u_k^{-1}$, then $e^S + \beta e^k - e^{\{1,\cdots,n\} \setminus S} \in \mathscr{E}$.

The above method for constructing the extreme points of $\mathscr{C}$ simplifies when $u \geqq 0$. For a set $S \subseteq \{1, \cdots, n\}$, let $u^{(S)} \equiv u^{(S,\{1,\cdots,n\})}$. We observe that when $u \geqq 0$, the only set $V \subseteq N$ is the empty set, for which (4.5) is straightforward. In particular, (4.6) becomes

$$u^{(W \setminus \{k\})} < 0 \leqq u^{(W)}. \tag{4.9}$$

Moreover, there is clearly no need to select first $W \subseteq P$ and then $V \subseteq N$. A corresponding simplification of the method holds when $u \leqq 0$.

We will next obtain explicit solutions to linear programming problems whose feasible set is given by the set $\mathscr{C}$ defined through (4.1) with respect to some vector

$u \in R^n$. Specifically, we consider linear programs having the form:

$$\max_{\substack{\|x\|_\infty \leqq 1 \\ x^T u = 0 \\ x \in R^n}} x^T a, \tag{4.10}$$

where $a \in R^n$ and $u \in R^n$ are arbitrary given vectors. We next observe that by making the change of variables $y_i = x_i$ if $u_i \geqq 0$ and $y_i = -x_i$ if $u_i < 0$, we convert (4.10) to the problem:

$$\max_{\substack{\|y\|_\infty \leqq 1 \\ y^T v = 0 \\ y \in R^n}} y^T a, \tag{4.11}$$

where $v_i = u_i$ if $u_i \geqq 0$ and $v_i = -u_i$ if $u_i < 0$. In particular $v \geqq 0$. We conclude that, without loss of generality, one can consider (4.10) under the assumption that $u \geqq 0$. We also observe that the solution of (4.10) when $u = 0$ is trivial as the set of the optimal solutions in this case is the set $\{x^* \in R^n \colon x_i^* = 1 \text{ if } a_i > 0,\ x_i^* = -1 \text{ if } a_i < 0 \text{ and } -1 \leqq x^* \leqq 1 \text{ if } a_i = 0\}$. Thus we will consider (4.10) under the assumption that $u > 0$.

THEOREM 4.3. *Let $u \in R^n$ be a semipositive vector and let a be an arbitrary vector in R^n with*

$$a_1/u_1 \geqq a_2/u_2 \geqq \cdots \geqq a_n/u_n, \tag{4.12}$$

where $\alpha/0$ is defined to be $+\infty$ if $\alpha > 0$ and $-\infty$ if $\alpha \leqq 0$. Then an optimal solution to (4.10) *is the vector*

$$x^* = e^{\{1,\cdots,k-1\}} + \gamma e^k - e^{\{k+1,\cdots,n\}} \tag{4.13}$$

where $k \in \{1, \cdots, n\}$ is the smallest integer with $2\sum_{i=1}^k u_i > \sum_{i=1}^n u_i$ and where

$$\gamma \equiv 1 + \left(\sum_{i=1}^n u_i - 2 \sum_{i=1}^k u_i \right) u_k^{-1}. \tag{4.14}$$

Moreover, if $\{i = 1, \cdots, n \colon u_i = a_i = 0\} = \phi$ and if the inequalities in (4.12) *are strict whenever the corresponding terms are finite, then the vector x^* defined above is the unique optimal solution of* (4.10).

Proof. For each feasible solution x of (4.10) we define two quantities, $k(x)$ and $m(x)$. The quantity $k(x)$ is defined to be the smallest index $k \in \{1, \cdots, n\}$ with $x_k < 1$. Of course, as feasibility of x for (4.10) requires that $x^T u = 0$, we conclude from the semipositivity of u that $x_k < 1$ for some $k \in \{1, \cdots, n\}$, i.e., $k(x)$ is well defined. The quantity $m(x)$ is defined to be the smallest integer $m \in \{k(x)+1, \cdots, n\}$ for which $x_m > -1$, if such an integer exists, and $m(x) = n+1$ if $x_i \leqq -1$ for all $i \in \{k(x)+1, \cdots, n\}$. We observe that, as feasibility of a vector x for (4.10) assures that $-1 \leqq x_i \leqq 1$ for $i = 1, \cdots, n$ we have that $x_i = 1$ for $i \in \{1, \cdots, k(x)-1\}$, and $x_i = -1$ for $i \in \{k(x)+1, \cdots, m(x)-1\}$.

Compactness and continuity arguments assure that (4.10) has an optimal solution. Let x^* be an optimal solution of (4.10) which maximizes lexicographically the pair $(k(\cdot), m(\cdot))$ among all optimal solutions to (4.10). Thus, for every optimal solution x of (4.10), $k(x) \leqq k(x^*)$ and if $k(x) = k(x^*)$ then $m(x) \leqq m(x^*)$. To abbreviate notation let, henceforth, $k \equiv k(x^*)$ and $m \equiv m(x^*)$. We will next show that $m = n+1$.

Assume that $m \leqq n$. We first observe that (4.12) implies that there exists integers $1 \leqq p < q \leqq n$ with

$$i \in \{1, \cdots, p\} \Leftrightarrow u_i = 0 \text{ and } a_i > 0, \tag{4.15}$$

$$i \in \{p+1, \cdots, q\} \Leftrightarrow u_i > 0, \tag{4.16}$$

and

$$i \in \{q+1, \cdots, n\} \Leftrightarrow u_i = 0 \text{ and } a_i \leqq 0. \tag{4.17}$$

It is immediate to see that $k > p$ and, as $m \leqq n$, $m \leqq q$ (see the discussion preceding this Theorem 4.3 on solving (4.10) when $u = 0$). In particular, $u_k > 0$ and $u_m > 0$. We also have that $x_k^* < 1$ and, as $m \leqq n$, $x_m^* > -1$. Let $\tilde{x} \equiv x^* + \delta(u_k^{-1} e^k - u_m^{-1} e^m)$ where $\delta = \min\{(1 - x_k^*)u_k, \ (x_m^* + 1)u_m\}$. Evidently, $\delta > 0$. Also, $\tilde{x}_1 = x_i^*$ for $i \in \{1, \cdots, n\} \backslash \{k, m\}$, $-1 \leqq x_k^* \leqq \tilde{x}_k = x_k^* + \delta u_k^{-1} \leqq x_k^* + (1 - x_k^*) = 1$ and $1 \geqq x_m^* \geqq \tilde{x}_m = x_m^* - \delta u_m^{-1} \geqq x_m^* - (x_m^* + 1) = -1$. Hence, $\|\tilde{x}\|_\infty \leqq 1$. Also, $(\tilde{x})^T u = (x^*)^T u = 0$. We conclude that $\tilde{x}$ is feasible for (4.10). Next observe that (4.12) implies that $((\tilde{x})^T a = (x^*)^T a + \delta(u_k^{-1} a_k - u_j^{-1} a_j) \leqq (x^*)^T a$. Since x^* is optimal for (4.10), it follows that $(\tilde{x})^T a = (x^*)^T a$ and therefore $\tilde{x}$ is also optimal for (4.10). Next observe that if $\delta = (1 - x_k^*)u_k$ then $\tilde{x}_k = x_k^* + (1 - x_k^*) = 1$, assuring that $k(\tilde{x}) \geqq k + 1 > k(x^*)$. Alternatively, if $\delta < (1 - x_k^*)u_k$ then $\delta = (x_m^* + 1)u_m$, and therefore $\tilde{x}_k < x_k^* + (1 - x_k^*) = 1$ and $\tilde{x}_m = x_m^* - (x_m^* + 1) = -1$, implying that $k(\tilde{x}) = k(x^*)$ and $m(\tilde{x}) > m(x^*)$. In either case we obtain a contradiction to the lexicographic maximality of $(k(x^*), m(x^*))$ among all optimal solutions of (4.10). This contradiction proves that $m \equiv n + 1$.

As $m = n + 1$, we have that $x_i^* = -1$ for $i = k + 1, \cdots, n$ and therefore $x^* = e^{\{1, \cdots, k-1\}} + x_k^* e^k - e^{\{k+1, \cdots, n\}}$. Now as $(x^*)^T u = 0$ we have that $\sum_{i=1}^{k-1} u_i + x_k^* u_k - \sum_{i=k+1}^n u_i = 0$, i.e., $x_k^* = (\sum_{i=k+1}^n u_i - \sum_{i=1}^{k-1} u_i) u_k^{-1} = (\sum_{i=1}^n u_i + u_k - 2\sum_{i=1}^k u_i) u_k^{-1} = 1 + (\sum_{i=1}^n u_i - 2\sum_{i=1}^k u_i) u_k^{-1}$. Now, as $x_k^* < 1$ we have that $\sum_{i=1}^n u_i < 2\sum_{i=1}^k u_i$. Also, as $x_k^* \geqq -1$, we have that $(\sum_{i=1}^n u_i - 2\sum_{i=1}^k u_i) u_k^{-1} \geqq -2$, i.e., $\sum_{i=1}^n u_i - 2\sum_{i=1}^k u_i \geqq -2u_k$, or equivalently, $2\sum_{i=1}^{k-1} u_i \leqq \sum_{i=1}^n u_i$. Thus, k is the smallest integer with $2\sum_{i=1}^k u_i > \sum_{i=1}^n u_i$. We conclude that (4.13) holds with γ given by (4.14).

Next assume that for no $i \in \{1, \cdots, n\}$, $u_i = a_i = 0$ and that the inequalities in (4.12) are strict whenever the corresponding terms are finite. Let x^* be any optimal solution of (4.10) and let $k \equiv k(x^*)$ and $m \equiv m(x^*)$. The above arguments show that $k \geqq p$, and if $m \leqq n$ then $m \leqq q$, where p and q are determined by (4.15)–(4.17). Now, if $m \leqq n$, then $\tilde{x} \equiv x^* + \delta(u_k^{-1} e^k - u_m^{-1} e^m)$, where $\delta \equiv \min\{(1 - x_k^*)u_k^{-1}, (x_m^* + 1)u_m^{-1}\} > 0$, is feasible for (4.10). Next observe that the corresponding strict inequalities in (4.12) imply that $(\tilde{x})^T a = (x^*)^T a + \delta(u_k^{-1} a_k - u_m^{-1} a_m) > (x^*)^T a$, contradicting the optimality of x^*. This contradiction proves that $m(x^*) = n + 1$ and therefore $x^* = e^{\{1, \cdots, k-1\}} + x_k^* e^k - e^{\{k+1, \cdots, n\}}$. The fact that x^* is necessarily the vector given by (4.13) now follows from the arguments used in the previous paragraphs.

We remark that an alternative proof to Theorem 4.1 can be obtained from Theorem 4.3 and Corollary 4.4 (which were obtained independently of Theorem 4.1) by using standard facts about the relation of extreme points of bounded, convex polyhedral sets and maximizers of linear functions over such sets. (Specifically, if $\mathscr{C}$ is a bounded, convex, polyhedral set, then every linear function on $\mathscr{C}$ obtains a maximum at some extreme point of $\mathscr{C}$. Also, for every extreme point of $\mathscr{C}$ there exists a linear function having a unique maximizer over $\mathscr{C}$ which is the given extreme point.)

Theorem 4.3 allows us to solve optimization problem (4.10) even when (4.12) is not satisfied, by simply permuting the indices $1, \cdots, n$ so that the permuted indices satisfy (4.12). Specifically, we get the following immediate corollary of Theorem 4.3.

COROLLARY 4.4. *Let $u \in R^n$ be a semipositive vector and let a be an arbitrary vector in R^n. Let $i(1), \cdots, i(n)$ be an enumeration of the indices $1, \cdots, n$ such that*

$$a_{i(1)}/u_{i(1)} \geqq a_{i(2)}/u_{i(2)} \geqq \cdots \geqq a_{i(n)}/u_{i(n)}, \tag{4.18}$$

where division by zero is defined as in Theorem 4.3. *Then, an optimal solution to* (4.10) *is the vector*

$$x^* \equiv e^{\{i(1),\cdots,i(k-1)\}} + \gamma e^{i(k)} - e^{\{i(k+1),\cdots,i(n)\}}, \tag{4.19}$$

where $k \in \{1, \cdots n\}$ is the smallest integer with $2\sum_{p=1}^{k} u_{i(p)} > \sum_{i=1}^{n} u_i$ and where

$$\gamma \equiv 1 + \left(\sum_{i=1}^{n} u_i - 2 \sum_{p=1}^{k} u_{i(p)} \right) u_k^{-1}. \tag{4.20}$$

Moreover, if $\{1 = 1, \cdots, n\colon u_i = a_i = 0\} = \phi$ and if the inequalities in (4.18) *are strict whenever the corresponding terms are finite, then the vector x^* defined above is the unique optimal solution of* (4.10).

Proof. The conclusion follows directly from Theorem 4.3. □

Explicit methods that use Corollary 4.4 to solve the optimization problem given by (4.10) are discussed in the Appendix.

We next specialize Corollary 4.4 to the case where $u = e \equiv (1, \cdots, 1)^T \in R^n$. It turns out that the corresponding expressions are easier to compute in this special case.

COROLLARY 4.5. *Let a be an arbitrary vector in R^n and let $i(1), \cdots, i(n)$ be an enumeration of the indices $1, \cdots, n$ such that*

$$a_{i(1)} \geqq a_{i(2)} \geqq \cdots \geqq a_{i(n)}. \tag{4.21}$$

Let $e \equiv (1, \cdots, 1)^T \in R^n$ and consider the optimization problem

$$\max_{\substack{\|x\|_\infty \leqq 1 \\ x^T a = 0 \\ x \in R^n}} x^T a. \tag{4.22}$$

Let $\lceil n/2 \rceil$ be the smallest integer larger than $n/2$. Then an optimal solution to (4.22) *is the vector*

$$x^* \equiv e^{\{i(1),\cdots,i(\lceil n/2\rceil - 1)\}} + \gamma e^{\lceil n/2 \rceil} - e^{\{\lceil n/2\rceil + 1,\cdots,n\}}, \tag{4.23}$$

where $\gamma \equiv 0$ if n is odd and $\gamma \equiv -1$ if n is even. Also the optimal value of the objective of (4.22) *is given by*

$$(x^*)^T a = \sum_{i=1}^{n} |a_i - \mu|, \tag{4.24}$$

where $\mu \equiv a_{\lceil n/2 \rceil}$ (i.e., *μ is a median of $a_1, \cdots, a_n$*).

Moreover, if (4.21) *holds with strict inequalities, then the vector x^* given by* (4.23) *is the unique optimal solution of* (4.22).

Proof. It is easily seen that when $u = e$, then the smallest integer k with $2\sum_{i=1}^{n} u_i > \sum_{i=1}^{n} u_i$ is the integer $\lceil n/2 \rceil$, and $\sum_{i=1}^{n} u_i - 2\sum_{i=1}^{\lceil n/2 \rceil} u_i$ equals -1 or 0 depending whether n is odd or even. These observations together with Corollary 4.4 imply that x^* defined by (4.23) with the corresponding γ is an optimal solution to (4.22) and it is a unique optimal solution when (4.21) holds with strict inequalities.

We finally establish (4.24). First consider the case where n is odd. Let $m = \lceil n/2 \rceil$ and observe the definition of x^* assures that

$$(x^*)^T a = \sum_{i=1}^{m-1} a_i - \sum_{i=m+1}^{n} a_i = \sum_{i=1}^{m-1} (a_i - a_m) - \sum_{i=m+1}^{n} (a_i - a_m)$$

$$= \sum_{i=1}^{m} |a_i - a_m| + \sum_{i=m+1}^{n} |a_i - a_m| = \sum_{i=1}^{n} |a_i - a_m|,$$

establishing (4.24). Similar arguments can be used to establish (4.24) when n is even.

We remark that Haviv (1983) and Haviv and van der Heyden (1983) have obtained (4.24) for a class of optimization problems that determine bounds on approximations of stationary vectors of stochastic matrices.

Appendix. The purpose of this appendix is to describe explicit procedures for solving the optimization problem (4.10) and their computational complexity. Throughout this appendix, let a be an arbitrary vector in R^n and let u be a semipositive vector in R^n. Also, we define the division of a real number by zero as in Theorem 4.3.

Corollary 4.4 suggests the following method for solving the optimization problem given by (4.10). First sort the n numbers $\{a_i/u_i \colon i = 1, \cdots n\}$ and then compute the quantities $i(k)$, γ, x^* and $(x^*)^T a$ defined in Corollary 4.4 (by using (4.19) and (4.20)). Since sorting can be accomplished by $O(n \lg n)$ comparisons (e.g., Knuth (1973, Vol. 3)) and the remaining calculations require $O(n)$ additions and multiplications, the computational effort of the resulting method is of the order of $O(n \lg n)$.

We next describe an $O(n)$ method for solving the optimization problem given by (4.10). Assume that one has identified an index m and a partition of $\{1, \cdots, n\}$ into two sets I_- and I_+ where $m \in I_-$, $I_- \subseteq \{i = 1, \cdots, n \colon a_i/u_i \leqq a_m/u_m\}$, $I_+ \subseteq \{i = 1, \cdots, n \colon a_i/u_i \geqq a_m/u_m\}$ and $\sum_{i \in I_- \setminus \{m\}} u_i \leqq 2^{-1} \sum_{i=1}^{n} u_i < \sum_{i \in I_-} u_i$. Now let

$$\gamma \equiv 1 + \left(\sum_{i=1}^{n} u_i - 2 \sum_{i \in I_-} u_i \right) u_m^{-1}$$

and let $x^* \in R^n$ be defined by

$$x_i^* \equiv \begin{cases} 1 & \text{if } i \in I_- \setminus \{m\}, \\ \gamma & \text{if } i = m, \\ -1 & \text{if } i \in I_+. \end{cases}$$

It is easy to see that there exists an enumeration $i(1), \cdots, i(n)$ of the indices $1, \cdots, n$ so that (4.18) holds, $I_- = \{i(1), \cdots, i(k)\}$, $I_+ = \{i(k+1), \cdots, i(n)\}$ and $i(k) = m$. It now follows directly from Corollary 4.4 that x^* is optimal for the optimization problem given by (4.10). Moreover, x^* is computable by $O(n)$ additions and multiplications (once m, I_- and I_+ are given).

We will next demonstrate that an index m and the corresponding sets I_- and I_+ can be identified by $O(n)$ comparisons, additions and multiplications. In fact, we will show that the (more general) "weighted median problem" (defined in the next paragraph) is solvable by $O(n)$ comparisons, additions and multiplications. For a finite real number x let $\lceil x \rceil$ be the smallest integer larger than or equal to x. We recall that an element α of a set $\{\alpha_1, \cdots, \alpha_n\}$ with $-\infty \leqq \alpha_i \leqq \infty$ for $i = 1, \cdots, n$ is called a *median* if there exist $\lceil n/2 \rceil$ elements in $\{\alpha_1, \cdots, \alpha_n\}$ which are smaller than or equal to α and the remaining elements are all larger than or equal to α. It is shown in Knuth (1973, Vol. 3, p. 216) that a median of n numbers can be identified by $O(n)$ comparisons, i.e., for some positive integer K a median or any n numbers, $n = 1$,

$2, \cdots$, can be determined by at most Kn comparisons. Moreover, Knuth (1973, Vol. 3, p. 219) assures that when a median is found, the relationship (namely, being smaller than, being equal to or being larger than) of each of the n numbers to this median must have been determined.

We will next describe the *weighted median problem.* Let $\alpha_1, \cdots, \alpha_n$ be given numbers, where $-\infty \leqq \alpha_i \leqq +\infty$ for $i=1, \cdots, n$. Also, let $w_1, \cdots, w_n$, w be additional numbers (called *weights*), where $0 \leqq w_i < \infty$ for $i=1, \cdots, n$ and $0 \leqq w < \sum_{i=1}^n w_i$. The corresponding weighted median problem is the problem of identifying an index $m \in \{1, \cdots, n\}$ and a partition of $\{1, \cdots n\}$ into two sets I_- and I_+ such that $m \in I_-$, $I_- \subseteq \{i=1, \cdots, n\colon \alpha_i \leqq \alpha_m\}$, $I_+ \subseteq \{i=1, \cdots, n\colon \alpha_i \geqq \alpha_m\}$ and $(\sum_{i \in I_-} w_i) - w_m \leqq w < \sum_{i \in I_-} w_i$. We will next establish the (possibly known) result that the weighted median problem is always solvable by $2(K+2)n$ comparisons, additions and multiplications. The following proof was communicated to us by Dan Gusfield. The proof follows by induction on n. The case where $n=1$ is trivial. Next assume that for some positive integer n^*, our assertion holds whenever $n=n^*$ and consider $n=n^*+1$. It follows from the discussion in the previous paragraph that, by at most Kn comparisons, one can determine a median of $\alpha_1, \cdots, \alpha_n$, say α_k, and a partition of $\{1, \cdots, n\}$ into two sets, say J_- and J_+, such that $k \in J_-$, $J_- \subseteq \{i=1, \cdots, n\colon \alpha_i \leqq \alpha_k\}$ $J_+ \subseteq \{i=1, \cdots, n\colon \alpha_i \geqq \alpha_k\}$ where J_- consists of $\lceil n/2 \rceil$ elements and J_+ consists of (the remaining) $n - \lceil n/2 \rceil$ elements. Next, $n-1$ additions and two comparisons can be executed to compute $w_- = \sum_{i \in I_-} w_i$, $w_+ = \sum_{i \in I_+}$ and determine whether $w_- - w_m \leqq w < w_-$, or $w < w_- - w_m$, or $w \geqq w_-$. In the former case, our weighted median problem has been solved by $Kn + n + 1 \leqq 2(k+2)n$ comparisons, additions and multiplications. In the two remaining cases, the problem is easily seen to be reduced to the weighted median problem of $\{\alpha_i\colon i \in J_- \backslash \{k\}\}$ with the corresponding weights and w unchanged, or to the weighted median problem of $\{\alpha_i\colon i \in J_+\}$ with corresponding weights and w being replaced by $w - w_-$ (the latter case requiring an additional subtraction). In either of the two cases, the number of corresponding elements does not exceed $n/2$ and therefore, by the induction hypothesis, is solvable by at most $2(K+2)(n/2) = (K+2)n$ comparisons, additions and multiplications. In particular, the original weighted median problem is solvable by $Kn + n + 2 + (K+2)n \leqq 2(K+2)n$ comparisons, additions and multiplications.

Acknowledgment. The author wishes to acknowledge the helpful comments of Ludo van der Heyden.

REFERENCES

F. R. Gantmacher (1959), *The Theory of Matrices*, translated from Russian in 1971 by K. A. Hirsh, Chelsea, New York.

D. E. Knuth (1973), *The Art of Computer Programming*, Addison-Wesley, Reading, MA.

M. Haviv (1982), *The stationary vector of a nearly completely-decomposable stochastic matrix: approximations and error bounds*, unpublished manuscript.

M. Haviv and L. van der Heyden (1983), *Perturbation bounds for the stationary probabilities of a finite Markov chain*, unpublished manuscript.

U. G. Rothblum and H. Schneider (1980), *Characterizations of optimal scalings of matrices*' Math. Programming 19, pp. 121–136.

U. G. Rothblum and C. P. Tan (1983), *Upper bounds on the maximum modulus of subdominant eigenvalues of nonnegative matrices*, Linear Algebra and Appl., to appear.

E. Seneta (1979), *Coefficients of ergodicity: structure and applications*, Adv. Appl. Prob., 11, pp. 576–590.

——— (1983a), *Spectrum localization by ergodicity coefficients for stochastic matrices*, Linear and Multilinear Algebra, to appear.

E. SENETA (1983b), *Explicit forms for ergodicity coefficients and spectrum localization*, Linear Algebra and Appl., to appear.

E. SENETA AND C. P. TAN (1983), *The Euclidean and Frobenius ergodicity coefficients and spectrum localization*, unpublished manuscript.

C. P. TAN (1982), *A functional form for a particular coefficient of ergodicity*, J. Appl. Prob. 4, pp. 858–863.

——— (1983), *Coefficients of ergodicity with respect to vector norms*, J. Appl. Prob., to appear.

——— (1984), *Spectrum localization of an ergodic stochastic matrix*, Bull. Inst. Math., Acad. Sinica, to appear.