

Combinatorial
Analysis



PROCEEDINGS OF
SYMPOSIA IN APPLIED MATHEMATICS
VOLUME X

COMBINATORIAL ANALYSIS

AMERICAN MATHEMATICAL SOCIETY

PROVIDENCE, RHODE ISLAND

1960

PROCEEDINGS OF THE
TENTH SYMPOSIUM IN APPLIED MATHEMATICS
OF THE AMERICAN MATHEMATICAL SOCIETY

Held at Columbia University
April 24-26, 1958

COSPONSORED BY
THE OFFICE OF ORDNANCE RESEARCH

EDITED BY
RICHARD BELLMAN
MARSHALL HALL, JR.

Prepared by the American Mathematical Society under Contract
No. DA-19-020-ORD-4545 with the Ordnance Corps, U.S. Army.

International Standard Serial Number 0160-7634
International Standard Book Number 0-8218-1310-2
Library of Congress Catalog Card Number 50-1183

Copyright © 1960 by the American Mathematical Society.
Second printing, 1979

Printed in the United States of America.

All rights reserved except those granted to the United States Government.
This book may not be reproduced in any form without the permission of the publishers.

CONTENTS

PREFACE	v
Current Studies on Combinatorial Designs	1
By MARSHALL HALL, Jr.	
Quadratic Extensions of Cyclic Planes	15
By R. H. BRUCK	
On Homomorphisms of Projective Planes	45
By D. R. HUGHES	
Finite Division Algebras and Finite Planes	53
By A. A. ALBERT	
The Size of the 10×10 Orthogonal Latin Square Problem	71
By L. J. PAIGE and C. B. TOMPKINS	
Some Combinatorial Problems on Partially Ordered Sets	85
By R. P. DILWORTH	
An Enumerative Technique for a Class of Combinatorial Problems	91
By R. J. WALKER	
The Cyclotomic Numbers of Order Ten	95
By A. L. WHITEMAN	
Some Recent Applications of the Theory of Linear Inequalities to Extremal Combinatorial Analysis	113
By A. J. HOFFMAN	
A Combinatorial Equivalence of Matrices	129
By A. W. TUCKER	
Linear Inequalities and the Pauli Principle	141
By H. W. KUHN	
Compound and Induced Matrices in Combinatorial Analysis	149
By H. J. RYSER	
Permanents of Doubly Stochastic Matrices	169
By MARVIN MARCUS and MORRIS NEWMAN	

A Search Problem in the n -cube	175
By A. M. GLEASON	
Teaching Combinatorial Tricks to a Computer.	179
D. H. LEHMER	
Isomorph Rejection in Exhaustive Search Techniques	195
By J. D. SWIFT	
Some Discrete Variable Computations	201
By OLGA TAUSKY and JOHN TODD	
Solving Linear Programming Problems in Integers	211
By R. E. GOMORY	
Combinatorial Processes and Dynamic Programming	217
By RICHARD BELLMAN	
Solution of Large Scale Transportation Problems	251
By MURRAY GERSTENHABER	
On Some Communication Network Problems	261
By ROBERT KALABA	
Directed Graphs and Assembly Schedules	281
By J. D. FOULKES	
A Problem in Binary Encoding	291
By E. N. GILBERT	
An Alternative Proof of a Theorem of König as an Algorithm for the Hitchcock Distribution Problem	299
By M. M. FLOOD	
INDEX	309

PREFACE

Problems in combinatorial analysis range from the study of finite geometries, through algebra and number theory, to the domains of communication theory and transportation networks. Although the questions that arise are all problems of arrangement, they differ enormously in the superficial form in which they arise, and quite often intrinsically, as well.

Perhaps the greatest discrepancy is between the discrete problems involving the construction of designs and the continuous problems of linear inequalities. Nevertheless, in a number of the papers that are presented a basic unity of the whole theory is brought to light. For example, Alan Hoffman has shown that many problems of discrete choice and arrangement may be solved in an elegant fashion by means of recent developments of the theory of linear inequalities, a continuation of work of Dantzig and Fulkerson. Similarly, Robert Kalaba and Richard Bellman have shown that a variety of combinatorial problems arising in the study of scheduling and transportation can be treated by means of functional equation techniques. Marshall Hall has observed that the solution of a problem in arrangements, in particular, the construction of pairs of orthogonal squares, is precisely equivalent to solving a certain equation for a matrix with nonnegative real entries.

A very challenging area of research which is investigated in a number of the papers that follow is that of using a computer to attack combinatorial questions, both by means of theoretical algorithms and by means of sophisticated search techniques. Papers by Paige and Tompkins, Walker, Gerstenhaber, Flood, Gleason, Lehmer, Swift, Todd, and Gomory, discuss versions of this fundamental problem.

Following the manner in which the Symposium was divided into four sessions, the Proceedings are divided into four sections. These are:

- I. Existence and construction of combinatorial designs.
- II. Combinatorial analysis of discrete extremal problems.
- III. Problems of communications, transportation and logistics.
- IV. Numerical analysis of discrete problems.

What is very attractive about this field of research is that it combines both the most abstract and most nonquantitative parts of mathematics with the most arithmetic and numerical aspects. It shows very clearly that the discovery of a feasible solution of a particular problem may necessitate enormous theoretical advances. Perhaps the moral of the tale is that the division into pure and applied mathematics is certainly artificial and to the detriment of the enthusiasts on both sides. Furthermore, the way in which

PREFACE

apparently simple problems require a complex medley of algebraic, geometric, analytic and numerical considerations shows that the traditional subdivisions of mathematics are themselves too rigidly labelled. There is one subject, mathematics, and one type of problem, a mathematical problem.

RICHARD BELLMAN,
The RAND Corporation

MARSHALL HALL, Jr.,
California Institute
of Technology

CURRENT STUDIES ON COMBINATORIAL DESIGNS

BY

MARSHALL HALL, JR.

1. Introduction. Current Studies on Combinatorial Designs have taken us into Number Theory, Group Theory, Matrix Theory, and to a certain extent into the theory of convex bodies. Number Theory and Group Theory have been used almost from the beginning of the theory of Combinatorial Analysis, but the more recent uses have been of a different nature.

The nature of some of the earlier methods may be illustrated in the theory of Steiner triple systems. A Steiner triple system is an arrangement of n objects into triples in such a way that every pair of distinct objects occurs in exactly one triple. It is trivial that n must be of the form $6k + 1$ or $6k + 3$. In 1859, six years after Steiner [21] had posed the problem, Reiss [20] showed that systems exist for every such value. Reiss's method was a recursively constructive method. By a fairly complicated construction he showed how, given a system with $t > 1$ objects he could construct one with $2t + 1$ objects and another with $2t - 5$. Starting with $t = 3$ we may obtain all possible values $6k + 1$ and $6k + 3$ recursively. A more general recursive method of constructing Steiner triple systems is due to E. H. Moore [18] who proved the theorem.

THEOREM. *If there is a Steiner triple system of order t_2 containing a subsystem of order t_3 , and if there is also a Steiner triple system of order $t_1 > 1$, then we can construct a system of order $t = t_3 + t_1(t_2 - t_3)$.*

If $6k + 1 = p$ is a prime and r is a primitive root of p , then the sets of residues mod p $i, i + r^a, i + r^{a+k}, a = 0, \dots, k-1, i = 0, \dots, 6k$ may be shown to be a Steiner triple system. Steiner triple systems also admit a composition, since if there are systems of orders t_1 and t_2 there is also a system of order t_1t_2 . This is easy to see in the following way: Given a Steiner triple system S , let us construct a quasi-group Q from the elements of S by the rules (1) $a^2 = a$ and (2) if $b \neq a$ and a, b, c is the triple of S containing a, b , put $ab = c$. Q may be characterized by the properties $a^2 = a$, $(ab)b = a$. Then the direct product of two such quasi-groups has the same property and this yields the composition rule.

In problems of enumeration, the theory of generating functions has been used from the beginning, and is still of great value, particularly in the study of problems of partitions. But I shall not concern myself here with this branch of Combinatorial Analysis. The symbolic calculus so extensively developed by MacMahon has been successful in giving formal algebraic equivalents of many combinatorial problems, but I cannot think of any

recent instance in which this approach has given either practical methods for constructing designs or for proving theorems about them.

For the greatest part, I shall speak of block designs. A *block design* is an arrangement of v objects into b blocks, each consisting of k distinct objects such that each object occurs in exactly r blocks and each pair of objects occurs in exactly λ blocks. The two following conditions on the five parameters are elementary :

$$(1.1) \quad bk = vr, \quad r(k - 1) = \lambda(v - 1).$$

A Steiner triple system is a block design with $k = 3$, $\lambda = 1$. A *symmetric block design* satisfies the further condition $v = b$, whence also $k = r$. In a symmetric block design the value $k - \lambda = n$ plays a central arithmetical role. A symmetric block design with $\lambda = 1$ is a *finite projective plane*, its parameters being

$$(1.2) \quad \begin{aligned} v &= b = n^2 + n + 1, \\ r &= k = n + 1, \\ \lambda &= 1. \end{aligned}$$

Here n is said to be the *order* of the plane.

2. Matrices and quadratic forms. Let a_1, \dots, a_v be the objects of a block design D and B_1, \dots, B_b the blocks. Let us define incidence numbers a_{ij} , $i = 1, \dots, v$, $j = 1, \dots, b$, where $a_{ij} = 1$ if $a_i \in B_j$ and $a_{ij} = 0$ if $a_i \notin B_j$. Then the design is fully described by the $v \times b$ *incidence matrix*,

$$(2.1) \quad A = (a_{ij}), \quad i = 1, \dots, v, \quad j = 1, \dots, b.$$

If A^T is the transpose of A , then the defining properties of the design imply

$$(2.2) \quad AA^T = B = (r - \lambda)I + \lambda S,$$

where I is the $v \times v$ identity matrix and S is a $v \times v$ matrix consisting entirely of 1's. It is easy to evaluate the determinant of B .

$$(2.3) \quad \det B = (r - \lambda)^{v-1}(r + (v - 1)\lambda).$$

If $r = \lambda$ then the design is the trivial one in which every block contains all the objects. Otherwise $r > \lambda$ and B is non-singular. Since the rank of B cannot exceed the rank of A , we have

$$(2.4) \quad b \geq v,$$

an inequality first proved by R. A. Fisher [8] by other means. Furthermore if D is a symmetric design $b = v$, $k = r$ and conditions (1.1) reduce to

$$(2.5) \quad k(k - 1) = \lambda(v - 1).$$

In this case $k + (v - 1)\lambda = k^2$, and of course

$$(2.6) \quad \det B = (\det A)^2.$$

Here (2.3) and (2.6) show immediately that the following theorem holds:

THEOREM 2.1. *In a symmetric block design, if v is even then $n = k - \lambda$ must be a square.*

For a symmetric design we find that the incidence matrix A is *normal*, i.e.

$$(2.7) \quad A^T A = A A^T = B = nI + \lambda S.$$

This expresses a duality in the design in particular the fact that in a symmetric design any two distinct blocks have exactly λ objects in common.

The matrix equation (2.2) has an alternate representation in terms of quadratic forms. Let x_1, \dots, x_v be indeterminates and let us use the incidence numbers a_{ij} to define linear forms

$$(2.8) \quad L_j = \sum_{i=1}^v a_{ij} x_i.$$

Then (2.3) is equivalent to

$$(2.9) \quad L_1^2 + L_2^2 + \cdots + L_v^2 = (r - \lambda)(x_1^2 + \cdots + x_v^2) + \lambda(x_1 + \cdots + x_v)^2 \\ = Q(x_1, \dots, x_v).$$

For a symmetric design (2.9) takes the form

$$(2.10) \quad L_1^2 + L_2^2 + \cdots + L_v^2 = n(x_1^2 + \cdots + x_v^2) + \lambda(x_1 + \cdots + x_v)^2 \\ = Q(x_1, \dots, x_v).$$

A major step in the study of Combinatorial Analysis was taken by Bruck and Ryser [4] in 1949 when they introduced the matrix notation given here and reasoned as follows: In (2.10) the linear forms L_j have rational coefficients (indeed the integers 0 and 1) and hence (2.10) gives a rational representation of the quadratic form Q by the form $L_1^2 + \cdots + L_v^2$. Thus the deep Hasse-Minkowski criteria for the rational equivalence of quadratic forms are applicable. Using this technique for finite projective planes, they proved the following result:

THEOREM 2.2. *A necessary condition for the existence of a finite projective plane with $n + 1$ points on a line is that if $n \equiv 1, 2 \pmod{4}$ then $n = a^2 + b^2$ for appropriate integers a and b .*

This shows that finite projective planes do not exist for an infinite set of values of n beginning with $n = 6, 14, 21, 22, \dots$. This result is in sharp contrast to the results on Steiner triple systems, where every value of the parameters consistent with the basic relations (1.1) is possible. The value $n = 6$ had previously been shown impossible by Tarry [22] by straightforward enumeration.

Theorem 2.2 was generalized by Chowla and Ryser [6] to symmetric block designs.

THEOREM 2.3. *If a symmetric design exists with parameters v , k , λ and $n = k - \lambda$, then (1) for v even, n is a square and (2) for v odd, the equation*

$$z^2 = nx^2 + (-1)^{(v-1)/2}\lambda y^2$$

is solvable in integers not all zero.

A different use was made of equation (2.9) by W. S. Connor [7]. If we specify the first t blocks of a design we have determined the first t forms L_1, \dots, L_t of (2.9). Then we have

$$(2.11) \quad L_{t+1}^2 + \dots + L_b^2 = Q(x_1, \dots, x_v) - L_1^2 - \dots - L_t^2 = Q^*.$$

If there exists a design with these initial blocks, then L_{t+1}, \dots, L_b exist and in particular the form Q^* must be positive semi-definite. Hence a necessary condition for the existence of a design with t specified blocks is that Q^* be positive semi-definite. Connor gives a test from this property in terms of a determinant. Let the t given blocks be B_1, \dots, B_t and let s_{ij} be the number of objects common to B_i and B_j for $i, j = 1, \dots, t$. Form the matrix

$$(2.12) \quad C_t = (c_{ij}), \quad i, j = 1, \dots, t \\ c_{ii} = (r - k)(r - \lambda), \quad c_{ij} = \lambda k - rs_{ij}, \quad i \neq j.$$

Then the determinant of C_t must satisfy

$$(2.13) \quad \begin{aligned} \text{(i)} \quad & \det |C_t| \geq 0 \quad \text{if } t < b - v, \\ \text{(ii)} \quad & \det |C_t| = 0 \quad \text{if } t > b - v, \\ \text{(iii)} \quad & k(r)^{-b+v+1}(r - \lambda)^{2v-b-1} \det C_{b-v} \end{aligned}$$

is a perfect integral square if there is to exist a design with blocks B_1, \dots, B_t . As one consequence he finds inequalities for the s_{ij} . He shows

$$(2.14) \quad \frac{1}{r} [2\lambda k + r(r - \lambda - k)] \geq s_{ij} \geq -r + k + \lambda.$$

For the symmetric designs we have $r = k$ and (2.14) gives $s_{ij} = \lambda$, the result previously noted, being equivalent to the duality of the symmetric designs and the normality of the incidence matrix A in (2.7). One of the applications of this method was the proof of the following embedding theorem by Hall and Connor [11].

THEOREM 2.4. *A block design with parameters $v = 2^{-1}t(t + 1)$, $b = 2^{-1}(t + 1)(t + 2)$, $r = t + 2$, $k = t$, $\lambda = 2$ can be embedded in a symmetric block design with $v = b = 2^{-1}(t^2 + 3t + 4)$, $r = k = t + 2$, $\lambda = 2$.*

This is the analogue, with $\lambda = 2$ instead of $\lambda = 1$ of the well known result that an affine plane can be embedded in a projective plane by adjoining a line at infinity. An example due to Bhattacharya shows the corresponding theorem for $\lambda = 3$ to be false. Connor and others have applied his method

to the study of partially balanced designs and other generalizations of block designs.

If we have t blocks B_1, \dots, B_t as the initial blocks of a symmetric design and if the obviously necessary condition $s_{ij} = \lambda$ holds, then the matrix C_t of (2.12) is identically zero and the Connor method gives no further information. This seemed a little strange and Hall and Ryser [13] endeavored to find out more about this situation. The results obtained indicate that indeed in the real field and more strongly, even in the rational field, no further information is available beyond that of Theorem 2.3 and the condition $s_{ij} = \lambda$. The precise state of affairs is given by the following theorem:

THEOREM 2.5. (NORMAL COMPLETION THEOREM). *Let v, k, λ be integers such that $k(k - 1) = \lambda(v - 1)$ and such that the conditions of Theorem 2.3 are satisfied. Let A_t be a rational $v \times t$ matrix whose columns all sum to k , and such that $A_t^T A_t = (k - \lambda)I_t + \lambda S_t$, I_t and S_t being $t \times t$ matrices such that I_t is the identity and S_t consisting entirely of 1's. Then there exists a rational $v \times v$ matrix A which has A_t for its first t columns such that $A^T A = A A^T = (k - \lambda)I + \lambda S$.*

Note that the hypothesis on A_t is certainly satisfied if this is the matrix of blocks B_1, \dots, B_t for which $s_{ij} = \lambda$. This says that not only is there a rational matrix A completing A_t to give a solution of the quadratic condition (2.9) but even a normal matrix A satisfying $A^T A = A A^T$. One corollary of this theorem is that the existence of a rational matrix X satisfying

$$X^T X = (k - \lambda)I + \lambda S$$

implies the existence of a rational A satisfying

$$A^T A = A A^T = (k - \lambda)I + \lambda S.$$

A special case of this last result had been proved previously by Albert [1].

3. A problem in convex spaces. Suppose we are given two superposed $n \times n$ orthogonal squares. For $n = 4$ an example is:

$$\begin{array}{cccc} 11 & 22 & 33 & 44 \\ 23 & 14 & 41 & 32 \\ 34 & 43 & 12 & 21 \\ 42 & 31 & 24 & 13 \end{array}$$

In general we have an $n \times n$ square and each cell contains a first and a second digit, chosen from 1 to n .

The first digits and second digits separately are Latin squares, i.e. each of 1 to n occurs exactly once in each row and once in each column. Furthermore the squares are *orthogonal*, i.e. in the superposed square the pairs of first and second digits occurring are all the combinations 11, 12, ..., nn. A pair of orthogonal squares may be regarded as representing four parallel

pencils in a finite affine plane with n^2 points. For let each of the n^2 cells be associated with a point.

In the first pencil let there be n lines, the i th containing the points of the i th row of the square. In the second let there be n lines each containing the points, the j th containing the points of the j th column. For the third pencil let the k th line, $k = 1, \dots, n$ consist of those points whose cells have k as their first digit. For the fourth pencil let the t th line, $t = 1, \dots, n$ consist of those points whose cells have t as their second digit. Then geometrically we have n^2 points and four sets of n lines such that

- (1) Each line contains n points.
- (2) Two lines of the same family are parallel.
- (3) Two lines of different families have exactly one point in common.
- (4) Through each of the n^2 points there is exactly one line of each family.

For property (3) as respects the third and fourth pencils, this is the orthogonality condition. Our construction may easily be reversed so that from four parallel pencils satisfying (1), (2), (3), (4) we may construct a pair of orthogonal $n \times n$ Latin squares. Thus orthogonal squares are not only interesting in themselves, but their existence is a necessary condition for the existence of an affine plane (and so also projective plane) of order n whenever $n \geq 3$. Euler conjectured that orthogonal squares do not exist if $n \equiv 2 \pmod{4}$. Tarry [22] verified this for $n = 6$ by trial, but up to the present no theorem on this exists¹ and the attempt to test $n = 10$ will be discussed by Tompkins and Paige at this Symposium. Mann [17] has shown that for $n \not\equiv 2 \pmod{4}$ two orthogonal squares exist. Thus there exist two orthogonal 21×21 squares although there is no plane of order 21.

Here I formulate the existence problem in terms of real quadratic forms in a way that leads to a number of problems on convex spaces.

Let us take $4n$ variables associating n with each pencil $x_i, i = 1, \dots, n$ with rows; $y_j, j = 1, \dots, n$ with columns; $z_k, k = 1, \dots, n$ with first digits and $w_t, t = 1, \dots, n$ with second digits. With the r th point $P_r, r = 1, \dots, n^2$ associate the linear form

$$(3.1) \quad L_r = x_i + y_j + z_k + w_t$$

if P_r is on the i th, j th, k th, t th lines of the respective pencils. Then

$$(3.2) \quad \begin{aligned} L_1^2 + L_2^2 + \cdots + L_{n^2}^2 &= Q \\ &= n(x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 + z_1^2 + \cdots + z_n^2 + w_1^2 + \cdots + w_n^2) \\ &\quad + 2 \sum_{i,j} x_i y_j + 2 \sum_{i,k} x_i z_k + 2 \sum_{i,t} x_i w_t + 2 \sum_{j,k} y_j z_k + 2 \sum_{j,t} y_j w_t \\ &\quad + 2 \sum_{k,t} z_k w_t. \end{aligned}$$

¹ Note added in proof. R. C. Bose, S. S. Shrikhande and E. T. Parker have succeeded in constructing pairs of $n \times n$ orthogonal squares for all $n \equiv 2 \pmod{4}, n \geq 10$.

The value of Q is easily determined from the defining properties of the pencils. Conversely if L_1, \dots, L_{n^2} are a selection of n^2 of the n^4 forms of (3.1) satisfying (3.2) then they determine a pair of $n \times n$ orthogonal squares. (For $n = 10$ we have only to choose 100 of 10,000 linear forms.)

THEOREM 3.1. *If for the Q of (3.2) we have*

$$(3.3) \quad Q = U_1^2 + \cdots + U_M^2$$

where the U_m , $m = 1, \dots, M$ are linear forms in the x, y, z, w with non-negative coefficients, then each of U_1, \dots, U_m is a scalar multiple of one of the n^4 forms of (3.1).

Proof. In (3.3) each U_m is of the form (3.4) $U_m = a_m x_{i_m} + b_m y_{j_m} + c_m z_{k_m} + d_m w_{l_m}$, where a_m, b_m, c_m, d_m are non-negative, since if a U_m contained as many as two x 's (or y 's, z 's, w 's) with positive coefficients, this would give a positive cross product in U_m^2 involving say $x_r x_s$, $r \neq s$ which cannot be canceled by the remaining U^2 's and yet is not present in Q . Of course a U_m might conceivably contain no x with a positive coefficient. Now consider those m 's for which $i_m = r$, $j_m = s$ and $a_m > 0$, $b_m > 0$. Call $T_{r,s}$ this set of m 's. Then, these being precisely those U_m 's for which U_m^2 gives a positive term in x_r, y_s , we have

$$(3.5) \quad \begin{aligned} \sum a_m b_m &= 1, \quad m \in T_{r,s}, \\ \sum a_m^2 &= A_{r,s}, \quad m \in T_{r,s}, \\ \sum b_m^2 &= B_{r,s}, \quad m \in T_{r,s}. \end{aligned}$$

Now as

$$(3.6) \quad \sum (a_m - b_m)^2 \geq 0 \quad m \in T_{r,s}$$

we have

$$(3.7) \quad A_{r,s} - 2 + B_{r,s} \geq 0$$

or

$$(3.8) \quad A_{r,s} + B_{r,s} \geq 2$$

with a strict inequality unless every term in (3.6) is 0.

$$(3.9) \quad A_r = \sum_{s=1}^n A_{r,s} \leq n$$

since the left hand side is $\sum a_m^2$ for all values of m for which $x_{i_m} = x_r$ and there is a y term. Further there is a strict inequality if for any m we have $x_{i_m} = x_r$ but no y term. Similarly

$$(3.10) \quad B_s = \sum_{r=1}^n B_{r,s} \leq n.$$

Combining (3.9) and (3.10)

$$(3.11) \quad \sum_{r,s} (A_{r,s} + B_{r,s}) \leq 2n^2.$$

But from (3.8)

$$(3.12) \quad \sum_{r,s} A_{r,s} + B_{r,s} \geq 2n^2.$$

Hence in all of (3.6) ··· (3.12) we must have strict equalities. In particular $a_m = b_m$ for $m \in T_{r,s}$ and for every U with a positive x term there is a positive y term. Continuing we conclude that

$$(3.13) \quad a_m = b_m = c_m = d_m, \quad m = 1, \dots, M$$

proving the theorem.

We may say even more about (3.3).

THEOREM 3.2. *In (3.3) $M \geq n^2$, and if $M = n^2$ then U_1, \dots, U_M determine a pair of orthogonal $n \times n$ Latin squares.*

Proof. Let $U_m = a_m(x_{i_m} + y_{j_m} + z_{k_m} + w_{l_m})$. Each U gives exactly 6 non-zero cross products xy etc. As Q has $6n^2$ non-zero cross products we must have $M \geq n^2$. If $M = n^2$ then each of the $6n^2$ cross products xy etc. must occur exactly once. Here from U_m we have the cross product $2a_m^2 x_{i_m} y_{j_m} = 2x_{i_m} y_{j_m}$ whence $a_m = 1$ in every case, and the U 's are L 's and so yield orthogonal squares.

These theorems can be given a matrix formulation:

THEOREM 3.3. *Let A be a $4n \times n^2$ matrix satisfying*

$$AA^T = \begin{pmatrix} nI & S & S & S \\ S & nI & S & S \\ S & S & nI & S \\ S & S & S & nI \end{pmatrix}.$$

If A is non-negative then A is the incidence matrix for a pair of orthogonal Latin squares.

The existence problem considered here has aspects relevant to general problems in the theory of convex spaces. The space S of semi-definite quadratic forms is a convex cone. So also is the space P of non-negative quadratic forms. The quadratic forms arising in our combinatorial problems are sums of squares of non-negative linear forms and this is again a convex cone D . Clearly $D \subseteq S \cap P$, but it has been shown by Horn that for 5 or more variables D is a proper subspace of $S \cap P$. Each of S and P is self-adjoint and the adjoint space of $S \cap P$ is $S \cup P$. The adjoint space of D is the space N of quadratic forms non-negative for non-negative arguments. Here $N \supseteq S \cup P$ and the inclusion is proper for forms with 5 or more variables by Horn's result. Let us note that $Q = n^{-2} \sum L^2$, L ranging

over the n^4 forms of (3.1). Theorem 3.1 says that a representative of Q as a linear combination of extreme points of D is in fact a linear combination of extreme points of a polyhedron, whose vertices are the squares of the n^4 forms of (3.1). Q/n^2 is indeed the center of gravity of these points. Theorem 3.2 points out that the number of extreme points needed in the representation of Q is vital to our problem. What is the nature of the points given as combination of a limited number of extreme points? In a square the combinations of two extreme points are the edges and diagonals.

4. Group theory and designs. From the beginning one way of constructing designs has been to take a group and find a design which has this as an automorphism group (or collineation group using geometric terminology). Occasionally the elements of the group itself form a design. Thus, given an elementary Abelian group G of order 2^r , let us delete from G the identity. Then the triples of elements of the form a, b, ab (i.e. a subgroup of order 4 with the identity deleted) form a Steiner triple S system on $2^r - 1$ elements. Here S has as its automorphism group not G itself, but $A(G)$ the group of automorphisms of G . As is well known, $A(G)$ is doubly transitive on S and is of order $(2^r - 1)(2^r - 2)\dots(2^r - 2^{r-1})$.

A result which may now be regarded as classical in the theory of projective planes states that the existence of certain families of configurations is equivalent to the existence of certain collineations. Specifically the validity of the theorem of Desargues for all configurations with a given center and axis is equivalent to the existence of all perspective collineations with that center and axis. The analogue of this state of affairs has not been sufficiently developed for designs in general. Let me however give one theorem of this kind.

THEOREM 4.1. *The following two conditions are equivalent in a Steiner triple system S :*

(1) *For every object $a \in S$ there is an involution α_a of S which fixes a and interchanges an object b with c if a, b, c are a triple of S .*

(2) *Every pair of intersecting triples of S lies in a subsystem which is an S_9 , i.e. a Steiner triple system with 9 objects.*

Proof. Let us show that (2) implies (1). Let us designate an object of S as 1. We wish to show that α_1 , the permutation fixing 1 and interchanging i with j if $1, i, j$ are a triple of S is a collineation of S . Clearly, α_1 maps onto themselves all triples including 1. It remains only to show that a triple not including 1 is also mapped onto a triple of S by α_1 . Let such a triple be say 2, 4, 6 and let 1, 2, 3 be the triple including 1 and 2. Then by our hypothesis 1, 2, 3 and 2, 4, 6 lie in an S_9 . We readily see that this must include

$$(4.1) \quad \begin{array}{ccc} 1 & 2 & 3 \\ 1 & 4 & 5 \\ 1 & 6 & 7 \end{array} \qquad \begin{array}{ccc} 2 & 4 & 6 \end{array}$$

The S_9 is now easily completed and seen to be of the form

$$(4.2) \quad \begin{array}{ccccc} 1 & 2 & 3 & 2 & 4 & 6 & 3 & 4 & 9 & 4 & 7 & 8 \\ 1 & 4 & 5 & 2 & 5 & 8 & 3 & 5 & 7 & 5 & 6 & 9 \\ 1 & 6 & 7 & 2 & 7 & 9 & 3 & 6 & 8 \\ 1 & 8 & 9 \end{array}$$

But then $\alpha_1 = (1)(2,3)(4,5)(6,7)(8,9)$ and $(2,4,6)\alpha_1 = 3,5,7$ which is a triple. Thus property (2) implies property (1).

On the other hand let us assume property (1) and let $1,2,3 ; 1,4,5$ be two intersecting triples of S . Then we certainly have triples

$$(4.3) \quad \begin{array}{ccccc} 1 & 2 & 3 & 2 & 4 & 6 \\ 1 & 4 & 5 \\ 1 & 6 & 7 \end{array}$$

and $\alpha_1 = (1)(2,3)(4,5)(6,7)$. Here $(2,4,6)\alpha_1 = 3,5,7$ must be a triple of S . But then the third element x of $2,5,x$ must be different from $1, \dots, 7$, say 8, and we have $2,5,8$ and also a triple $1,8,9$. Here $\alpha_1 = (1)(2,3)(4,5)(6,7)(8,9)$ and $(2,5,8)\alpha_1 = 3,4,9$. Hence we have triples

$$(4.4) \quad \begin{array}{ccccc} 1 & 2 & 3 & 2 & 4 & 6 & 3 & 5 & 7 \\ 1 & 4 & 5 & 2 & 5 & 8 & 3 & 4 & 9 \\ 1 & 6 & 7 \\ 1 & 8 & 9 \end{array}$$

and collineations

$$(4.5) \quad \begin{aligned} \alpha_1 &= (1)(2,3)(4,5)(6,7)(8,9) \\ \alpha_2 &= (2)(1,3)(4,6)(5,8) \\ \alpha_3 &= (3)(1,2)(4,9)(5,7) \\ \alpha_4 &= (4)(1,5)(2,6)(3,9) \\ \alpha_5 &= (5)(1,4)(2,8)(3,7) \end{aligned}$$

here

$$(4.6) \quad \begin{aligned} (1,4,5)\alpha_2 &= 3,6,8 \\ (1,4,5)\alpha_3 &= 2,9,7 \\ (1,2,3)\alpha_4 &= 5,6,9 \\ (1,2,3)\alpha_5 &= 4,8,7 \end{aligned}$$

giving us the complete S_9 containing $1,2,3$ and $1,4,5$ as above in (4.2). Thus property (1) implies (2).

A number of recent results are to the effect that certain hypotheses imply that a finite plane is Desarguesian. I shall content myself with listing several of these :

Gleason [9]. *Every finite Fano plane is Desarguesian.*

Here a Fano plane is a plane in which the diagonal points of a complete quadrilateral are collinear.

Gleason [9]. *A finite plane is Desarguesian if for every pair P,l where P is a point lying on the line l there is a non-identical elation with center P and axis l .*

André [2]. A finite plane is Desarguesian if for every pair P, l where P is a point not on the line l , there is a non-identical homology with center P and axis l .

Ostrom-Wagner [19]. A finite plane is Desarguesian if its collineation group is doubly transitive on the points of the plane.

The following is a conjecture :

CONJECTURE. A finite plane is Desarguesian if its collineation group is transitive on the points of the plane.

Approximating the conjecture is a result of Wagner :

A finite plane is Desarguesian if its collineation group is transitive and also contains a central collineation.

Most of the above results can be described roughly by saying : If a collineation group G of a finite plane is generated by certain collineations, then G in fact contains every possible elation. Results of a different, and perhaps more surprising type say : If a design has a certain collineation group G , then it has a larger collineation group G_1 .

Suppose that K is a $v - k - \lambda$ design and that G is a collineation group of K transitive and regular on the points of K . We may choose an arbitrary point P_0 as a base point. Then if P_i is any other point there is a unique $x \in G$ such that $P_0 x = P_i$. In this way there is a 1-1 correspondence between the points of K and the elements of G . Hence we may think of the points as being elements of G . Let D be the sets of points in a particular line (i.e. block) of K . Then the lines of K are precisely the sets Dx , $x \in G$. D is called a difference set and has the properties:

(i) If $x \in G$, $x \neq 1$, then there are exactly λ distinct ordered pairs (d_1, d_2) of elements of D such that $x = d_1^{-1}d_2$.

(ii) If $x \in G$, $x \neq 1$, then there are exactly λ distinct ordered pairs (d_3, d_4) of elements of D such that $x = d_3 d_4^{-1}$. Conversely k elements of a group G of order v , where $k(k - 1) = \lambda(v - 1)$ satisfying either (i) or (ii) will satisfy the other and determine a $v - k - \lambda$ design K whose blocks are the sets D_x , $x \in G$.

In the group ring of G over the rationals let

$$(4.7) \quad \begin{aligned} \Delta &= \sum_{x \in D} x, & S &= \sum_{x \in G} x, \\ \Delta^* &= \sum_{x \in D} x^{-1}. \end{aligned}$$

Then the basic properties of the design imply

$$(4.8) \quad \Delta\Delta^* = \Delta^*\Delta = (k - \lambda) \cdot 1 + \lambda S.$$

This relation is analogous to the basic relation (2.7) on incidence matrices. The major theorem on further collineations of K is the following :

THEOREM 4.2. If G is Abelian and p is a prime dividing $n = k - \lambda$, $(p, v) = 1$, and $p > \lambda$ then the mapping $x \rightarrow x^p$ is also a collineation of K .

This result was first proved by Hall [10] for projective planes when G is cyclic. With G cyclic it was proved for general $v - k - \lambda$ designs by Hall and Ryser [12]. The general formulation given here is due to Bruck [4]. In every known example the condition $p > \lambda$ is superfluous but the known proofs require this. Naturally if $\lambda = 1$ the condition $p > \lambda$ is automatically satisfied. Theorem 4.2 is valid in certain other cases, in particular for projective planes ($\lambda = 1$) if G fixes a line and a point not on the line, or the vertices and edges of a triangle providing G is Abelian, and is transitive and regular on the remaining points. These results are due to A. J. Hoffman [14] and D. R. Hughes [15].

If a design D has a collineation group G then application of the Hasse-Minkowski techniques yields still further existence conditions over and above the Bruck-Ryser condition for the existence of the design. D. R. Hughes [16] obtains the following theorem:

THEOREM 4.3. *Let G be a collineation group of a $v - k - \lambda$ design D and suppose that every element of G not the identity fixes the same N points. If G is of order m , $t = (v - N)/m$, $e = (t + N - 1)/2$, then the equation*

$$x^2 = (k - \lambda)y^2 + (-1)^e m^{N-1} \lambda z^2$$

possesses a nontrivial solution in integers.

In a converse sense certain groups lead to combinatorial designs. For example the Mathieu groups are associated with certain complicated designs. Let me discuss another case in which the combinatorial side arose unexpectedly. We find the following theorem in Burnside [3, p. 207].

THEOREM 4.4. *A primitive permutation group G of degree n , which has a subgroup H that keeps $n - m$ symbols fixed and is transitive on the remaining m symbols is at least doubly transitive. If H is primitive then G is $(n - m + 1)$ -ply transitive.*

Let us investigate the possibility that G is not $(n - m + 1)$ -ply transitive. We may suppose that G is r -ply transitive with $2 \leq r \leq n - m - 1$. Note that if G is $(n - m)$ -ply transitive it is also $(n - m + 1)$ -ply transitive since H , fixing $n - m$ letters is transitive on the remaining m letters. Then G has a subgroup G_1 fixing $r - 2$ letters which is doubly but not triply transitive. G is a transitive extension of G_1 , i.e. given G_1 by adjoining $r - 2$ letters to the letters moved by G_1 , G is an r -ply transitive group in which G_1 is the subgroup fixing the specified $r - 2$. We investigate G_1 of degree $n_1 = n - r + 2$. We may suppose that m is taken as large as possible and that H consists of all permutations fixing the $n_1 - m$ letters. Now as $2 \leq r \leq n - m - 1$ it follows that $m \leq n - r - 1 = n_1 - 3$. Here since m is as large as possible it also follows that $m > n_1/2$. For if $m \leq n_1/2$, by the primitivity of G , it follows that there is a conjugate H_1 of H such that H and H_1 have some but not all their displaced letters in common. $H_1 \cup H_2$

is then transitive on the letters displaced by either of them, this being on $2m - s$ letters where there are s letters displaced by both of them. Here $m < 2m - s$ but as m was chosen as large as possible we must have $2m - s \geq n_1 - 1$. If $m \leq n_1/2$ this is possible only if $m = n_1/2$ and $s = 1$. But then a third conjugate of H , say H_2 , displaces at least two letters in common with either H or H_1 and either $H \cup H_2$ or $H_1 \cup H_2$ will give a larger m contrary to the choice of m_1 . Hence $m > n_1/2$ and so any two conjugates of H have a letter in common, and by the maximality of m the union of two conjugates of H displaces $n_1 - 1$ or n_1 letters. Let us write $k = n_1 - m$, and consider all the sets of k letters fixed by different conjugates of H . Any two distinct k -sets have at most one letter in common since the union of distinct conjugates of H displaces $n_1 - 1$ or n_1 letters. Furthermore since G is doubly transitive and $k \geq 3$, there is one k -set (and so only one) containing a specified pair of distinct letters. There are as many k -sets containing a letter i as there are conjugates of H in the subgroup of G_{14} and G_1 fixing the letter i . Hence this number r is the same for every i . But these results show that the k -sets form a block design D with $\lambda = 1$. Furthermore G_1 may be regarded as a collineation group of D with the property that the subgroup H of G_1 fixing a block of D pointwise is transitive on the remaining points. We state our conclusions as a theorem.

THEOREM 4.5. *A group G_1 which is doubly but not triply transitive on n_1 letters and contains a subgroup H fixing $n - m$ letters and transitive on the remaining m is the collineation group of a block design D with $\lambda = 1$. Conversely a block design D with $\lambda = 1$ and a collineation group G_1 which is doubly transitive on the points of D and has a subgroup fixing the points of a line and transitive on the remaining points gives such a group. Every primitive group G of degree n with a subgroup H fixing $n - m$ letters and transitive on the remaining letters is either (i) $(n - m + 1)$ -ply transitive or (ii) is a group G_1 as above or a transitive extension of such a G_1 .*

BIBLIOGRAPHY

1. A. A. Albert, *Rational normal matrices satisfying the incidence equation*, Proc. Amer. Math. Soc. vol. 4 (1953) pp. 554–559.
2. J. André, *Über Perspektivitäten in endlichen projektiven Ebenen*, Arch. Math. vol. 6 (1954) pp. 29–32.
3. W. Burnside, *The theory of groups*, Cambridge University Press, 2d ed., 1911.
4. R. H. Bruck, *Difference sets in a finite group*, Trans. Amer. Math. Soc. vol. 78 (1955) pp. 464–481.
5. R. H. Bruck and H. J. Ryser, *The nonexistence of certain finite projective planes*, Canad. J. Math. vol. 1 (1949) pp. 88–93.
6. S. Chowla and H. J. Ryser, *Combinatorial problems*, Canad. J. Math. vol. 2 (1950) pp. 93–99.
7. W. S. Connor, *On the structure of balanced incomplete block designs*, Ann. Math. Statist. vol. 23 (1952) pp. 57–71.

8. R. A. Fisher, *An examination of the different possible solutions of a problem in incomplete blocks*, Ann. of Eugenics vol. 10 (1940) pp. 52–75.
9. A. M. Gleason, *Finite Fano planes*, Amer. J. Math. vol. 78 (1956) pp. 797–807.
10. Marshall Hall, Jr., *Cyclic projective planes*, Duke Math. J. vol. 14 (1947) pp. 1079–1090.
11. Marshall Hall, Jr. and W. S. Connor, *An embedding theorem for balanced incomplete block designs*, Canad. J. Math. vol. 6 (1953) pp. 35–41.
12. Marshall Hall, Jr. and H. J. Ryser, *Cyclic incidence matrices*, Canad. J. Math. vol. 3 (1951) pp. 495–502.
13. ———, *Normal completion of incidence matrices*, Amer. J. Math. vol. 76 (1954) pp. 581–589.
14. A. J. Hoffman, *Cyclic affine planes*, Canad. J. Math. vol. 4 (1952) pp. 295–301.
15. D. R. Hughes, *Partial difference sets*, Amer. J. Math. vol. 78 (1956) pp. 650–674.
16. ———, *Collineations and generalized incidence matrices*, Trans. Amer. Math. Soc. vol. 86 (1957) pp. 284–296.
17. H. B. Mann, *On the construction of sets of orthogonal Latin squares*, Ann. Math. Statist. vol. 14 (1943) pp. 401–414.
18. E. H. Moore, *Concerning triple systems*, Math. Ann. vol. 43 (1893) pp. 271–285.
19. T. G. Ostrom, *Double transitivity in finite projective planes*, Canad. J. Math. vol. 8 (1956) pp. 563–567.
20. M. Reiss, *Über eine Steinersche kombinatorische Aufgabe welche in 45 sten Bande dieses Journals, Seite 181, gestellt worden ist*, Crelle's Journal vol. 56 (1859) pp. 326–344.
21. J. Steiner, *Combinatorische Aufgabe*, Crelle's Journal vol. 45 (1853) pp. 181–182.
22. G. Tarry, *Le problème des 36 officiers*, Compte Rendu de l'Ass. Fr. pour l'Av. de Sci. Nat. 29 (1900) Part I pp. 122–123; Part II pp. 170–203.

OHIO STATE UNIVERSITY,
COLUMBUS, OHIO

CALIFORNIA INSTITUTE OF TECHNOLOGY,
PASADENA, CALIFORNIA

QUADRATIC EXTENSIONS OF CYCLIC PLANES

BY

R. H. BRUCK¹

1. Introduction. The specific facts to be proved in this paper may be indicated in the following table:

m :	2	3	4	5	6	7	8	9	10
n :	4	9	16	25	—	49	64	81	—
Section:	4	8	5	9	—	10	7	11	—
Comment:	K	K	K	K	—	N	N	N	—

Briefly put: for each of the values listed, we prove that a cyclic projective plane of square order $n = m^2$ is unique and hence Desarguesian. (The dashes correspond to cases where the planes do not exist.) Since the natural order for m is not the most convenient order for the proofs, we list the appropriate sections of the paper for ready reference. The line of "Comment" indicates whether the result is known (K) or new (N).

In the paragraphs which follow we shall touch on the theoretical background of the problem and the methods of the present paper. Let us note here that although a great deal of exploratory (hand) calculation lies behind the paper, the "proofs" presented are proofs in the accepted mathematical sense. It appears to the author that the present results could be extended considerably by using high speed computing machines to find what might be called "counter-examples"; the "counter-examples", once known, are easily verified directly. In this connection, see, for example, the method of proof in §§7, 10, and 11.

By a finite projective plane π of order n (where n is a positive integer not less than two) we mean a non-empty system of undefined objects called points and lines together with an incidence relation such that every two distinct points (lines) are together incident with just one line (point) and every point (line) is incident with precisely $n + 1$ distinct lines (points). The total number of points (lines) is $n^2 + n + 1$.

A collineation of a projective plane π is a one-to-one mapping of the points upon the points and the lines upon the lines which preserves incidence. A projective plane π is cyclic if π possesses a cyclic group of collineations which is both transitive and regular on the points and on the lines of π . Every finite Desarguesian projective plane is known [1] to be cyclic. There exist infinite cyclic planes which are not Desarguesian [2] but the existence of finite cyclic non-Desarguesian planes is still undecided. The statement that

¹ This research was supported by the National Science Foundation.

every finite cyclic plane is Desarguesian is equivalent to the following pair of statements :

- (i) *Every finite cyclic projective plane has prime-power order.*
- (ii) *The cyclic projective planes of a given prime-power order are all isomorphic.*

Statement (i) has been proved for all orders up to 1600 [3] as well as for various infinite classes of orders [4; 5; 6; 7]. The information about (ii), on the other hand, seems to go no further than order 49 [8]. In addition, some of the facts about (ii) must be regarded as experimental in nature, inasmuch as they rest on the assumption that a search by machine has been carried out completely and correctly.

In §§2, 3 we exploit the known correspondence between cyclic planes and difference sets. The results, in essence, are as follows: Every cyclic plane π of order $n = m^2$ possesses a cyclic projective subplane π' of order m uniquely determined by the fact that it is mapped into itself by the given cyclic group of collineations of π . Conversely, π can be constructed from π' in terms of a so-called “fixed” difference set D_m for π' together with a single-valued mapping f from the integers (incongruent to zero) modulo

$$s = m^2 - m + 1$$

into a well-defined subset $K = K(D_m)$ of the integers modulo

$$t = m^2 + m + 1.$$

The function f satisfies certain conditions (see §2) the simplest of which is this :

$$f(ax) \equiv af(x) \pmod{t}$$

for every multiplier (a) of D_m . The conditions do not determine f uniquely (except when $m = 2, 3, 4$). However, if (a) is a multiplier of D_m and if b is prime to s , f may always be replaced by an “equivalent” function g , defined by

$$g(x) \equiv af(bx) \pmod{t}.$$

The rest of the paper consists in showing, for the values of m listed above, that D_m determines f uniquely aside from an equivalence. Since we use Desarguesian planes as our initial building blocks, we always end up with a Desarguesian plane.

2. Cyclic planes of square order. Let n be a positive integer, $n \geq 2$, and set $N = n^2 + n + 1$. By a *difference set*, $D = D_n$, of order n , we mean² a set of $n + 1$ integers, incongruent modulo N , such that to each integer x , incongruent to zero modulo N , there corresponds a unique ordered pair d ,

² The only difference sets considered here are those appropriate for cyclic projective planes. That is, the parameter usually called λ has the value 1.

d' of elements of D satisfying the congruence $d - d' \equiv x \pmod{N}$. Each translate $D + a$ of D (where a is an integer) is also a difference set of order n if D is. Every cyclic plane π of order n can be represented [2] as follows: The N points of π are the integers modulo N ; the N lines of π are the translates modulo N of a difference set D of order n ; and point x is incident with line $D + y$ if and only if $x - y \equiv d \pmod{N}$ for some d in D . The mapping

$$x \rightarrow x + 1, \quad D + y \rightarrow D + y + 1$$

generates the natural cyclic collineation group of π . If k is an integer, the mapping

$$(k): x \rightarrow kx, \quad D + y \rightarrow kD + ky$$

is an isomorphism of π upon a cyclic plane provided

(a) k is prime to N .

And (k) will be a collineation of π if also

(b) kD is a translate of D .

When (a), (b) hold, (k) is called a *multiplier* of π (or D). It is known [3; 8] that if k is congruent modulo N to a product of prime divisors of n , then (k) is a multiplier of π . Moreover, D can be replaced by a suitable translate (in at least one and at most three ways modulo N) so as to ensure that D is left fixed by every multiplier; that is, $kD \equiv D \pmod{N}$ for every multiplier (k) of π .

We assume henceforth that the order n of the cyclic plane π is a square; specifically, $n = m^2$, $m \geq 2$. If

$$(2.1) \quad s = m^2 - m + 1, \quad t = m^2 + m + 1,$$

then

$$(2.2) \quad N = n^2 + n + 1 = st.$$

We also assume that $D = D_n$ is left fixed by every multiplier. First we consider the multiplier (m^3) and note that $m^3x \equiv x \pmod{N}$ if and only if $x \equiv 0 \pmod{s}$. Thus the t points of form sy are precisely those left fixed by (m^3) . In view of [2], the multiplier (m^3) also leaves fixed precisely t lines. One of these lines is D by assumption. Therefore (m^3) leaves fixed precisely the t lines of form $D + sz$. Since $t > 3$, the t fixed points and t fixed lines constitute a projective subplane which is cyclic with respect to the collineation group of π generated by

$$x \rightarrow x + s, \quad D + y \rightarrow D + y + s.$$

It now follows readily that $D = D_n$ has the form

$$(2.3) \quad D_n = sD_m \cup E$$

where sD_m , E are disjoint sets of integers and D_m is a difference set of order m .

The number of elements of E is $(n + 1) - (m + 1) = m^2 - m = s - 1$. If $x \not\equiv 0 \pmod{t}$, then there exists a unique ordered pair d, d' of elements of

D_m such that $x \equiv d - d' \pmod{t}$ and hence $sx \equiv sd - sd' \pmod{N}$. Consequently if a, b are distinct elements of D_n such that $a - b \equiv 0 \pmod{s}$ but $a - b \not\equiv 0 \pmod{N}$, then $a \equiv sd, b \equiv sd' \pmod{N}$ for a unique ordered pair d, d' of elements of D_m . Thus we see the following:

- (I) If e is in E , then $e \not\equiv 0 \pmod{s}$.
- (II) If e, e' are in E and if $e \not\equiv e' \pmod{N}$, then $e \not\equiv e' \pmod{s}$. In view of (I), (II), the $s - 1$ elements of E , together with 0, constitute a complete set of residues modulo s . Therefore (modulo N)

$$(2.4) \quad E = \{xt + f(x)s \mid x \not\equiv 0 \pmod{s}\}$$

where f is a suitable single-valued function from the nonzero integers mod s into the integers mod t .

We shall need further properties of f . Let (k) be a multiplier of π . Since (k) leaves D_n fixed and since sD_m , E consist of integers divisible by s and not divisible by s , respectively, it is clear from (2.3) that $kE \equiv E \pmod{N}$. Therefore, by (2.4) for each $x \not\equiv 0 \pmod{s}$,

$$k(xt + f(x)s) \equiv kxt + f(kx)s \pmod{N}.$$

That is:

$$(2.5) \quad f(kx) \equiv kf(x) \pmod{t} \quad \text{if } x \not\equiv 0 \pmod{s}$$

for every multiplier (k) of π . In particular we take $k = m^3$. Since $m^3 \equiv -1 \pmod{s}$ and $m^3 \equiv 1 \pmod{t}$, we deduce that

$$(2.6) \quad f(-x) \equiv f(x) \pmod{t} \quad \text{if } x \not\equiv 0 \pmod{s}.$$

And (2.6) shows that f takes on at most $(s - 1)/2$ incongruent values modulo t .

Next let us consider, for a fixed $x \not\equiv 0 \pmod{s}$, the t incongruent integers $z \pmod{N}$ such that $z \equiv xt \pmod{s}$. In view of the form of D_n , these must be congruent mod N to the integers of the following forms:

- (a) The $m + 1$ integers $xt + (f(x) - d)s$, d in D_m .
- (b) The $m + 1$ integers $xt + (d - f(-x))s$, d in D_m .
- (c) The $s - 2$ integers $xt + (f(x + y) - f(y))s$, $y \not\equiv 0, -x \pmod{s}$.

Since $s - 2 + 2(m + 1) = t$, we see using (2.6) that f must satisfy the following condition (C):

(C) For each $x \not\equiv 0 \pmod{s}$, the following three sets of integers make up a single complete set of residues mod t :

- (i) The $m + 1$ integers $f(x) - d$, d in D_m .
- (ii) The $m + 1$ integers $d - f(x)$, d in D_m .
- (iii) The $s - 2$ integers $f(x + y) - f(y)$, $y \not\equiv 0, -x \pmod{s}$.

Our first use of (C) is to prove:

$$(2.7) \quad f \text{ takes on exactly } (s - 1)/2 \text{ incongruent values modulo } t.$$

To prove (2.7), consider two integers $a, b \pmod s$ such that $a \not\equiv 0, b \not\equiv 0, a \not\equiv \pm b \pmod s$. Since s is odd, the congruences $2x \equiv a + b, 2y \equiv a - b \pmod s$ have unique solutions x, y . Moreover, $x \not\equiv 0, y \not\equiv 0 \pmod s$ and, in addition, $x + y \equiv a, x - y \equiv b \pmod s$. Clearly $y \not\equiv \pm x \pmod s$. Consequently, by (iii) of (C),

$$f(x + y) - f(y) \not\equiv f(x - y) - f(-y) \pmod t.$$

Since $f(y) \equiv f(-y) \pmod t$ by (2.6), we see that $f(x + y) \not\equiv f(x - y) \pmod t$. That is, $f(a) \not\equiv f(b) \pmod t$. By this and (2.6): if $x, y \not\equiv 0 \pmod s$, then $f(x) \equiv f(y) \pmod t$ if and only if $x \equiv \pm y \pmod s$. Hence we have (2.7).

Before determining the set of values of f explicitly it will be convenient to introduce a certain set $K = K(D)$ as follows: If $m \geq 2$ is an integer, if D is a set of $m + 1$ integers incongruent modulo $t = m^2 + m + 1$, then $K = K(D)$ is the set of integers $k \pmod t$ such that

$$(2.8) \quad 2k \not\equiv d + d' \pmod t$$

for every choice of d, d' in D . We now prove a simple lemma:

LEMMA 2.1. *Let s, t be given by (2.1), where $m \geq 2$ is an integer. A necessary and sufficient condition that a set D of $m + 1$ integers incongruent modulo t be a difference set of order m is that $K(D)$ consist of exactly $(s - 1)/2$ incongruent integers modulo t .*

Proof. There are exactly $(m + 1)(m + 2)/2$ unordered pairs (d, d') of elements of D and each contributes exactly one unordered sum $d + d'$. Since t is odd, K will have exactly

$$t - (m + 1)(m + 2)/2 = (s - 1)/2$$

incongruent elements modulo t in case distinct unordered pairs contribute sums incongruent modulo t ; and will have more elements otherwise. If D is a difference set the congruence $d_1 + d_2 \equiv d_3 + d_4 \pmod t$, or $d_1 - d_3 \equiv d_4 - d_2 \pmod t$, can hold for d_i in D if and only if either $d_1 \equiv d_3, d_2 \equiv d_4 \pmod t$ or $d_1 \equiv d_4, d_2 \equiv d_3 \pmod t$; that is, if and only if the unordered pairs $(d_1, d_2), (d_3, d_4)$ are equal. In this case, K has exactly $(s - 1)/2$ distinct elements modulo t . Conversely, if K has exactly $(s - 1)/2$ elements modulo t and if $d_1 - d_2 \equiv d_3 - d_4 \pmod t$ for d_i in D , then $d_1 + d_4 \equiv d_2 + d_3 \pmod t$, so that the unordered pairs $(d_1, d_4), (d_2, d_3)$ must be equal. This implies that D is a difference set.

Returning to our earlier discussion, we now can prove another property of f : *f is a single-valued mapping of the integers incongruent to zero mod s upon the set $K(D_m)$.* To see this, we first observe that the number, $(s - 1)/2$, of distinct elements of $K(D_m)$ is the same as the number of distinct values of f . Next we observe, by condition (C), that, for each $x \not\equiv 0 \pmod s$,

$$f(x) - d \not\equiv d' - f(x) \pmod t$$

for all d, d' in D_m . Therefore $f(x)$ is in $K(D_m)$ and the proof is complete.

We note finally that the assumption that D_n was left fixed by every multiplier of the cyclic plane π of order $n = m^2$ can be relaxed somewhat at very little expense. Suppose that we replace D_m by a translate $D_m + c$. Then (2.3) suggests that D_n be replaced by the translate

$$D_n + sc = s(D_m + c) \cup (E + cs).$$

In this case f should be replaced by g where

$$g(x) \equiv f(x) + c \pmod{t}.$$

Moreover, we observe that (2.6) and condition (C) still hold for g , though (2.5) may not, and that

$$K(D_m + c) \equiv K(D_m) + c \pmod{t}.$$

In the next section we shall invert our point of view, regarding D_m as given and considering the problem of constructing D_n .

Before leaving the present discussion we shall add a remark, unnecessary for the sequel, which indicates an interesting pattern for certain projective planes of square order: *Every cyclic projective plane π of order $n = m^2$, $m \geq 2$, can be partitioned into $s = m^2 - m + 1$ projective subplanes $\pi_1, \pi_2, \dots, \pi_s$, each of order m , such that every point and every line of π belongs to exactly one of the π_i .* Indeed, for each integer $i = 1, 2, \dots, s$, let the points of π_i be those of form $zs + i - 1$ and the lines of π_i be those of form $D_n + zs + i - 1$ where z ranges over the integers mod t .—It was through the observation that the cyclic planes of square order could so be partitioned that the author was led to the present paper.

3. Quadratic extensions. We now suppose that a cyclic projective plane π' of order m , $m \geq 2$, is given, and we consider the problem of constructing a cyclic projective plane π of order $n = m^2$ containing a subplane isomorphic to π' . We use the notation of §2. We may assume that π' is represented by a difference set D_m mod t and that π is represented by a difference set D_n mod N , $N = st$, of form (2.3) where E has the form of (2.4). The main theorem may be stated as follows:

THEOREM 3.1. *Let D_m be a difference set mod t and let D_n be a set mod $N = st$ given by (2.3), (2.4). A necessary and sufficient condition that D_n be a difference set mod N is that f be a single-valued function from the integers incongruent to zero mod s upon the set $K = K(D_m)$ such that f satisfies (2.6) and condition (C) of §2. If D_m is left fixed by each of its multipliers then a like fact is true for D_n and, in addition, f satisfies (2.5) for every multiplier (k) of D_n .*

Proof. The necessity of the conditions has been proved in §2. Sufficiency is a straightforward consequence of (2.6) and (C), together with the fact that D_m is a difference set. Now suppose that D_m is left fixed by each of its multipliers and let (k) be a multiplier of D_n . Thus

$$(3.1) \quad kD_n \equiv D_n + a \pmod{N}$$

for some integer a . We observe that kD_n contains precisely $m + 1$ integers congruent to zero mod s , namely the elements of ksD_m . If $a \not\equiv 0 \pmod{s}$ then $D_n + a$ contains exactly one integer congruent to zero mod s , namely one element of $E + a$; but this is a contradiction. Hence $a \equiv bs \pmod{N}$ for some integer b . We now deduce from (3.1) that

$$ksD_m \equiv sD_m + bs \pmod{N},$$

whence

$$kD_m \equiv D_m + b \pmod{t}.$$

Therefore (k) induces a multiplier of D_m . Consequently, by our supposition, $b \equiv 0 \pmod{t}$ and $a \equiv bs \equiv 0 \pmod{N}$. Therefore (k) leaves D_n fixed. Using this fact, we see as in §2 that f satisfies (2.5) for every multiplier (k) of D_n . The proof of Theorem 3.1 is now complete.

In applying Theorem 3.1 to the problem of constructing D_n from a given D_m we are continually embarrassed by a lack of restrictions on the choice of f . The lemma which follows gives partial help in this matter.

LEMMA 3.1. *Let D_m be a difference set mod t which is left fixed by each of its multipliers. Let (a) be any multiplier of D_m and let b be any integer prime to s . If the function f of Theorem 3.1 is replaced by a function g such that*

$$g(x) \equiv af(bx) \pmod{t}$$

for every integer $x \not\equiv 0 \pmod{s}$, then the difference set D_n of Theorem 3.1 is replaced by an equivalent difference set.

Proof. We note the fact that s and t are relatively prime. In view of this fact, there exists exactly one integer $r \pmod{N}$ such that $r \equiv a \pmod{t}$ and $rb \equiv 1 \pmod{s}$. Since, necessarily, a is prime to t , we see that r is prime to $N = st$. Therefore the mapping

$$x \rightarrow rx, \quad D_n + y \rightarrow rD_n + ry \pmod{N}$$

replaces the cyclic plane π defined by D_n by an equivalent plane defined by

$$rD_n = rsD_m \cup rE.$$

Since $r \equiv a \pmod{t}$ and since $aD_m \equiv D_m \pmod{t}$ we see that

$$rsD_m \equiv sD_m \pmod{N}.$$

Since E consists of elements $xt + f(x)s$ where x ranges over all integers $x \not\equiv 0 \pmod{s}$, and since

$$r(bxt + f(bx)s) \equiv xt + af(bx)s \pmod{N},$$

we see that

$$rD_n \equiv sD_m \cup E' \pmod{N}$$

where E' is defined in terms of g instead of f . This completes the proof of Lemma 3.1.

The following concept is helpful when the integer t is not too large. For each k in $K = K(D)$ (where $D = D_m$) let $C(k)$ be the set of all integers mod t which are not in $(D - k) \cup (k - D)$. By applying condition (C) to f we get the following:

LEMMA 3.2. *If p, q are integers such that $p, q, p - q, p + q$ are incongruent to zero mod s then, modulo t ,*

$$(3.2) \quad f(p) - f(q) \in C(f(p - q)) \cap C(f(p + q)).$$

Proof. We first assume only that $p, q, p - q \not\equiv 0 \pmod{s}$. Then the integer x defined by $q + x \equiv p \pmod{s}$ satisfies the condition $x \not\equiv 0 \pmod{s}$. Set

$$k = f(x) = f(p - q).$$

By applying (C) with this choice of x , we see that

$$f(q + x) - f(q) = f(p) - f(q)$$

is incongruent mod t to any element of form $k - d$ or $d - k$ where d is in $D = D_m$. That is, $f(p) - f(q)$ belongs to $C(k) = C(f(p - q))$. Since $f(-q) \equiv f(q) \pmod{t}$, we need merely add the requirement that $p + q \not\equiv 0 \pmod{s}$ and replace q by $-q$ to see that $f(p) - f(q)$ belongs to $C(f(p + q))$. This completes the proof of Lemma 3.2.

4. $m = 2$. Here $s = 3, t = 7$. It is well known that the only projective plane of order 2 is Desarguesian and hence cyclic. This plane has the powers of (2) as its only multipliers and can be represented by the difference set

$$(4.1) \quad D_2 \equiv \{1, 2, 4\} \pmod{7},$$

which is left fixed by (2). We readily find that $K(D_2) \equiv \{0\} \pmod{7}$. Hence $f(\pm 1) \equiv 0 \pmod{7}$ and (with cycles mod 21 under (2) separated by a semi-colon):

$$(4.2) \quad D_4 \equiv \{3, 6, 12; 7, 14\} \pmod{21}.$$

Since D_2 is left fixed by all of its multipliers, so (by Theorem 3.1) must D_4 be. Consequently, the only multipliers of D_4 are the powers of (2).

5. $m = 4$. Here $s = 13, t = 21$. We may assume that D_4 is given by (4.2). We find that $K(D_4)$ consists of the six distinct powers of 2 mod 21. In view of Lemma 3.1, we may assume that $f(1) \equiv 1 \pmod{21}$. By Theorem 3.1,

$$f(2x) \equiv 2f(x) \pmod{21}$$

for each $x \not\equiv 0 \pmod{13}$. This determines f completely and we find that (with cycles mod 273 under (2) separated by semi-colons):

$$(5.1) \quad D_{16} \equiv \{39, 78, 156; 91, 182; 17 \cdot 2^k, k = 0, 1, \dots, 11\} \pmod{273}.$$

6. $m = 16$. Here $s = 241$, $t = 273$. We may assume that D_{16} is given by (5.1). The present methods seem inadequate to allow the construction of D_{256} , let alone to consider the question of uniqueness, without the use of computing machines. We shall indicate briefly why this is so. Part of the computations can certainly be carried out by hand. For example, since D_{16} is left fixed by the multiplier (2), so must be $K = K(D_{16})$; and this observation allows us to verify that K consists of 10 cycles of length 12 each, under multiplication by 2, namely those cycles containing 3, 5, 11, 21, 25, 29, 41, 49, 57 and 97 respectively. The $C(k)$ can also be calculated by first giving k the 10 stated values and then using the fact that $C(2k) \equiv 2C(k) \pmod{273}$. Moreover the integers incongruent to zero mod 241 break up into 10 cycles of length 24 each, under multiplication by 2; and these cycles are easily determined. However, Lemma 3.1 only allows us to restrict f by the requirement that

$$f(2^i) \equiv 3 \cdot 2^i \pmod{273}, \quad i = 0, 1, \dots, 23.$$

We still have to decide how to map 9 cycles into 9 cycles. Even with the help of Lemma 3.2 the details are formidable.³

7. $m = 8$. Here $s = 57$, $t = 73$. Since 73 is a prime and since 2 has exponent 9 mod 73, we see immediately that, without loss of generality, we may take

$$(7.1) \quad D_8 \equiv \{1, 2, 4, 8, 16, 32, 64, 55, 37\} \pmod{73},$$

a difference set left fixed by all multipliers, namely by the powers of (2). For the discussion of D_{64} it is convenient to exhibit explicitly the cycles mod 73 under multiplication by 2. These are the cycle consisting of 0 alone, together with the following:

$$\begin{aligned} (\text{a}) : & 1 \ 2 \ 4 \ 8 \ 16 \ 32 \ 64 \ 55 \ 37 \\ (\text{a}') : & 72 \ 71 \ 69 \ 65 \ 57 \ 41 \ 9 \ 18 \ 36 \\ (\text{b}) : & 3 \ 6 \ 12 \ 24 \ 48 \ 23 \ 46 \ 19 \ 38 \\ (\text{b}') : & 70 \ 67 \ 61 \ 49 \ 25 \ 50 \ 27 \ 54 \ 35 \\ (\text{c}) : & 5 \ 10 \ 20 \ 40 \ 7 \ 14 \ 28 \ 56 \ 39 \\ (\text{c}') : & 68 \ 63 \ 53 \ 33 \ 66 \ 59 \ 45 \ 17 \ 34 \\ (\text{d}) : & 11 \ 22 \ 44 \ 15 \ 30 \ 60 \ 47 \ 21 \ 42 \\ (\text{d}') : & 62 \ 51 \ 29 \ 58 \ 43 \ 13 \ 26 \ 52 \ 31 \end{aligned}$$

In this notation we may exhibit D_8 , $K = K(D_8)$ and $C(0)$ as follows:

$$\begin{aligned} D_8 : & (\text{a}), \quad K : 0, (\text{d}), (\text{d}'), (\text{b}'). \\ C(0) : & 0, (\text{b}), (\text{b}'), (\text{c}), (\text{c}'), (\text{d}), (\text{d}'). \end{aligned}$$

³ Since the above was written the author has made a preliminary investigation of the case $m = 16$ along the lines of §§10, 11 below. The work could undoubtedly be carried out by hand but would be quite tedious.

Since $f(2x) \equiv 2f(x) \pmod{73}$ and since $4 \cdot 19 \equiv 19 \pmod{57}$, we see that

$$(7.2) \quad f(19) \equiv f(38) \equiv 0 \pmod{73}.$$

In view of Lemma 3.1, we may assume without loss of generality that $f(18)$ is congruent modulo 73 to one of 11, 62 or 70. However, if $f(18) \equiv 62$, then, on the one hand,

$$16 - f(18) \equiv 27 \pmod{73}$$

whereas, on the other hand,

$$\begin{aligned} f(9 + 18) - f(9) &\equiv f(30) - f(48) \equiv f(2^3 \cdot 18) - f(2^8 \cdot 18) \\ &\equiv (8 - 37)f(18) \equiv 27 \pmod{73}, \end{aligned}$$

in contradiction to condition (C) for $x = 18$. Again, if $f(18) \equiv 70$, then

$$f(18) - 4 \equiv 66 \pmod{73}$$

whereas

$$\begin{aligned} f(15 + 18) - f(15) &\equiv f(24) - f(15) \equiv f(2^7 \cdot 18) - f(2^2 \cdot 18) \\ &\equiv (55 - 4)f(18) \equiv 66 \pmod{73}, \end{aligned}$$

in contradiction to (C) for $x = 18$. Therefore

$$(7.3) \quad f(18) \equiv 11 \pmod{73}.$$

We note next that 1, 18, 20 are representatives of the three cycles of length 18 under multiplication mod 57 by 2. Moreover, if g be defined by

$$g(x) \equiv f(20x) \pmod{73},$$

then g is equivalent to f by Lemma 3.1 and, in addition,

$$g(18) \equiv f(18), g(1) \equiv f(20), g(20) \equiv f(1) \pmod{73}.$$

Consequently there is no loss of generality in assuming that $f(1), f(20)$ lie in the cycles (b'), (d') respectively.

By condition (C), the set $C(0) = C(f(19))$ consists, modulo 73, of the integers

$$f(y + 19) - f(y) \quad \text{where} \quad y \not\equiv 0, 38 \pmod{57}.$$

Noting that $4 \cdot 19 \equiv 19 \pmod{57}$, we first take $y = \pm 1, \pm 18, \pm 20$ and then multiply successively by powers of 4. We deduce that

$$f(1) - 11, 11 - f(1), f(20) - 11, 11 - f(20), f(1) - f(20), f(20) - f(1)$$

represent the cycles (b), (b'), (c), (c'), (d), (d') in some order. A straightforward check, where $f(1)$ ranges over (b') and $f(20)$ over (d'), reduces the possibilities to the following:

$$\begin{aligned} f(1) : & 70 \ 61 \ 50 \ 50 \ 54 \ 54 \\ f(20) : & 58 \ 31 \ 62 \ 26 \ 51 \ 31 \end{aligned}$$

We reduce these possibilities to a single choice by using (C) for $x = 18$. Specifically, we exhibit as follows an element of $\pm(D_{16} - f(18))$ which is congruent mod 73 to

$$\begin{aligned} p &\equiv f(43 + 18) - f(43) \equiv f(4) - f(14) \\ &\equiv 2^2 f(1) - 2^7 f(1) \pmod{73}. \end{aligned}$$

If $f(1) \equiv 70 \pmod{73}$, then $p \equiv 61 - 54 \equiv 7 \equiv f(18) - 4 \pmod{73}$.

If $f(1) \equiv 50 \pmod{73}$, then $p \equiv 54 - 49 \equiv 5 \equiv 16 - f(18) \pmod{73}$.

If $f(1) \equiv 54 \pmod{73}$, then $p \equiv 70 - 50 \equiv 20 \equiv f(18) - 64 \pmod{73}$.

In each case we have a contradiction, leaving a unique possibility:

$$(7.4) \quad f(1) \equiv 61, \quad f(20) \equiv 31 \pmod{73}.$$

The difference set is now uniquely determined. Since there exists a Desarguesian (hence cyclic) plane of order $2^6 = 64$, we should actually have a difference set. To check this directly, we need only verify condition (C) for $x = 19, 1, 18$ and 20 . In terms of cycles mod 4161 under multiplication by 2, the difference set appears as follows:

$$(7.5) \quad D_{64} \text{ modulo } 4161 :$$

$$\begin{array}{ccccccccccccc} 57 & 114 & 228 & 456 & 912 & 1824 & 3648 & 3135 & 2109 & & 1387 & 2774; \\ 307 & 614 & 1228 & 2456 & 751 & 1502 & 3004 & 1847 & 3694 & 3227 & 2293 & 425 & 850 \\ & & 1700 & 3400 & 2639 & 1117 & 2234; \\ 371 & 742 & 1484 & 2968 & 1775 & 3550 & 2939 & 1717 & 3434 & 2707 & 1253 & 2506 & 851 \\ & & 1702 & 3404 & 2647 & 1133 & 2266; \\ 1413 & 2826 & 1491 & 2982 & 1803 & 3606 & 3051 & 1941 & 3882 & 3603 & 3045 & 1929 & 3858 \\ & & 3555 & 2949 & 1737 & 3474 & 2787. \end{array}$$

8. $m = 3$. Here $s = 7, t = 13$. The unique projective plane of order 3 may be represented by the difference set

$$(8.1) \quad D_3 \equiv \{0; 1, 3, 9\} \pmod{13}$$

which is left fixed by all its multipliers, namely, by the powers of (3). If $K = K(D_3)$, necessarily $3K \equiv K \pmod{13}$. This helps us to verify that

$$K \equiv \{4, 12, 10\} \pmod{13}.$$

We may assume that $f(2) \equiv 4 \pmod{13}$, and then f is completely determined by the requirement that $f(3x) \equiv 3f(x) \pmod{13}$. Hence we find that

$$(8.2) \quad D_9 \equiv \{0; 7, 21, 63; 2, 6, 18, 54, 71, 31\} \pmod{91}.$$

And D_9 must be left fixed by all its multipliers, namely, by the powers of (3).

9. $m = 5$. Here $s = 21, t = 31$. It is known that there is only one projective plane of order 5. This may be represented by the difference set

$$(9.1) \quad D_5 \equiv \{1, 5, 25; 11, 24, 27\} \pmod{31},$$

which is left fixed by all multipliers, namely by the powers of (5). We observe that the cycles mod 31 under multiplication by 5 consist of 0 and the following:

$$\begin{array}{ll} (a) : 1 \ 5 \ 25 & (a') : 6 \ 30 \ 26 \\ (b) : 2 \ 10 \ 19 & (b') : 12 \ 29 \ 21 \\ (c) : 3 \ 15 \ 13 & (c') : 16 \ 18 \ 22 \\ (d) : 4 \ 20 \ 7 & (d') : 11 \ 24 \ 27 \\ (e) : 8 \ 9 \ 14 & (e') : 17 \ 23 \ 22 \end{array}$$

In these terms, $K = K(D_5)$ and $C(0)$ are given as follows:

$$K : 0, (d), (b'), (e'). \quad C(0) : 0, (b), (b'), (c), (c'), (e), (e').$$

Since $f(5x) \equiv 5f(x) \pmod{31}$ and since $5 \cdot 7 \equiv -7 \pmod{21}$, we see that $5f(7) \equiv f(7) \pmod{31}$, whence

$$(9.2) \quad f(7) \equiv 0 \pmod{31}.$$

By applying condition (C) with $x = 7$, we see in view of (9.2) that the elements

$$f(y + 7) - f(y) \pmod{31}, \quad y \not\equiv 0, 14 \pmod{21},$$

must fill out $C(0)$. By first taking $y = 7, 15, 20, 6, 8, 13, 1$ and then multiplying by powers of 5, we deduce that *the elements*

$$(9.3) \quad \begin{aligned} & f(1) - f(6), f(6) - f(1), f(8) - f(6), \\ & f(6) - f(8), f(1) - f(8), f(8) - f(1) \end{aligned}$$

must represent the six cycles (b), (b'), (c), (c'), (e), (e') in some order. This focuses our attention on $f(1), f(6), f(8)$.

First we observe that 7, 1, 6, 8 belong to distinct cycles mod 21 under multiplication by 5. By this and (9.2), $f(1), f(6), f(8)$ *must represent the cycles (d), (b'), (e') in some order.*

Next we show as follows that $f(1), f(8)$ are essentially interchangeable: Since 8 is prime to 21, Lemma 3.1 allows us to replace f by the function g defined by

$$g(x) \equiv f(8x) \pmod{31};$$

moreover, since $8 \cdot 6 \equiv 6$ and $8 \cdot 8 \equiv 1 \pmod{21}$,

$$g(1) \equiv f(8), g(6) \equiv f(6), g(8) \equiv f(1) \pmod{31}.$$

Thus $f(1), f(8)$ *may be interchanged without altering the value of $f(6)$.*

If $f(6)$ is in (e'), then, by Lemma 3.1, we may assume without loss of generality that $f(6) \equiv 17 \pmod{31}$. But then

$$f(6 + 6) - f(6) \equiv 5f(6) - f(6) \equiv 6 \equiv f(6) - 11 \pmod{31},$$

contradicting condition (C) for $x = 6$. Therefore one of $f(1), f(8)$ is in (e'). Since $f(1), f(8)$ are interchangeable we may assume that $f(8)$ is in (e'); and furthermore, by Lemma 3.1, we may assume that

$$(9.4) \quad f(8) \equiv 22 \pmod{31}.$$

The condition (9.3) does not differentiate between $f(1)$, $f(6)$. Hence we assume temporarily that $f(1)$ is in (d), $f(6)$ is in (b'), and verify rapidly that (9.3) requires $f(1) \equiv 4$, $f(6) \equiv 12$. That is,

$$f(1), f(6) \equiv 4, 12 \pmod{31}$$

in one of the two possible orders. However, if $f(1) \equiv 4$, $f(6) \equiv 12$, then

$$f(6) - 27 \equiv 12 - 27 \equiv 16 \pmod{31}$$

and also

$$f(4 + 6) - f(4) \equiv f(11) - f(4) \equiv 25f(8) - 25f(1) \equiv 16 \pmod{31},$$

which contradicts condition (C) with $x = 6$. Therefore we must have

$$(9.5) \quad f(1) \equiv 12, \quad f(6) \equiv 4 \pmod{31}.$$

The difference set D_{25} is uniquely determined by (9.1), (9.2), (9.4), (9.5). That it is in fact a difference set may be checked by using condition (C) for $x = 7, 6, 1, 8$. In terms of cycles mod 651 under multiplication by 5, we get:

$$(9.6) \quad D_{25} \text{ mod } 651 :$$

$$\begin{aligned} & 21\ 105\ 525; 231\ 504\ 567; 217\ 434; \\ & 48\ 240\ 549\ 141\ 54\ 270; 59\ 295\ 173\ 214\ 419\ 142; \\ & 113\ 565\ 221\ 454\ 317\ 283. \end{aligned}$$

10. $m = 7$. Here $s = 43$, $t = 57$. It is known (Hall [9]) that the only projective plane of order 7 is the Desarguesian plane. This may be represented by the difference set

$$(10.1) \quad D_7 \equiv \{19; 38; 5, 35, 17; 30, 39, 45\} \pmod{57},$$

which is left fixed by all multipliers, namely by the powers of (7).

We begin by noting that $57 = 3 \cdot 19$. Since $7 \equiv 1 \pmod{3}$, we are able to simplify some of our considerations materially by working modulo 3 instead of modulo 57. Since 2 is a primitive root mod 19 such that $2^6 \equiv 7 \pmod{19}$, and since $40 \equiv 1 \pmod{3}$, $40 \equiv 2 \pmod{19}$, we find it convenient to list the following cycles mod 57 under multiplication by 40 :

(I)	(II)	(III)
1 40 4 46 16 13	2 23 8 35 32 26	3 6 12 24 48 39
7 52 28 37 55 34	14 47 56 17 53 11	21 42 27 54 51 45 $(\pmod{57})$
49 22 25 31 43 10	41 44 50 5 29 20	33 9 18 36 15 30

The eighteen entries in (I), (II) or (III) form a single cycle under multiplication by 40. On the other hand, the six columns in (I), (II) or (III) form six cycles mod 57 of length three under multiplication by 7; and these are congruent mod 3 to 1, 2 or 0 respectively. In addition to these eighteen cycles of length three (under multiplication by 7) there are three cycles of length one, namely 0, 19, 38. If we introduce the "basic" cycle B :

$$B \equiv \{1, 7, 49\} \pmod{57},$$

the cycles of length three may be designated by xB where x ranges over the first rows of (I), (II), (III). In this notation, D_7 and $K = K(D_7)$ appear as follows:

$$D_7: 19; 38; 35B; 39B. \quad K: B, 16B; 2B, 23B; 3B, 24B, 48B.$$

If we turn to condition (C) for $x = 1$, we are led to consider the function F defined by

$$(10.2) \quad F(y) \equiv f(1 + y) - f(y), \quad y \not\equiv 0, 42 \pmod{43}.$$

We observe first, in view of the properties of f , that

$$(10.3) \quad F(21) \equiv 0 \pmod{57}$$

and that

$$(10.4) \quad F(42 - y) \equiv -F(y) \pmod{57}, \quad 1 \leq y \leq 20.$$

The first easily obtained fact about F is the number of its zeros mod 3 (assuming, of course, that (C) holds):

LEMMA 10.1. *In terms of the value of $f(1)$ modulo 3, the following table gives the correct number of zeros mod 3 of F in the range $1 \leq y \leq 20$:*

$$\begin{array}{c} f(1) \pmod{3}: 0 \ 1 \ 2 \\ \text{No. of Zeros: } 6 \ 8 \ 5. \end{array}$$

Proof. Let V denote the set of integers $F(y)$, $1 \leq y \leq 20$. Then, in view of (10.3), (10.4), condition (C) for $x = 1$ asserts that the sets

$$D_7 - f(1), \quad f(1) - D_7, \quad 0, \quad V, \quad -V,$$

taken together, consist of 57 integers constituting a complete set of residues mod 57. In particular, precisely 19 of these integers are congruent to zero mod 3. Let a, b denote the number of integers congruent to zero mod 3 in $D_7 - f(1)$ and V , respectively. Then $2a + 1 + 2b = 19$, whence

$$b = 9 - a.$$

Since we know D_7 , we simply calculate a in each case and verify that b has the values stated in the lemma. This completes the proof of Lemma 10.1.

Before we can evaluate F explicitly we must consider the integer $s = 43$. We observe that 43 is a prime and that 34 is a primitive root mod 43 whose powers $34^n \pmod{43}$ may be listed as follows for $0 \leq n \leq 20$:

$$\begin{array}{ccccccccc} 1 & 34 & 38 & 2 & 25 & 33 & 4 \\ 7 & 23 & 8 & 14 & 3 & 16 & 28 & \\ 6 & 32 & 13 & 12 & 21 & 26 & 24 & \end{array} \quad (\text{mod } 43).$$

By design, each column of the table is half of a cycle mod 43 (of length six) under multiplication by 7. The table suggests the following definitions:

$$(10.5) \quad f_i \equiv f(34^i) \pmod{57}, \quad 0 \leq i \leq 6.$$

Thus, more explicitly,

$$(10.6) \quad \begin{aligned} f_0 &\equiv f(1), f_1 \equiv f(34), f_2 \equiv f(38), f_3 \equiv f(2), f_4 \equiv f(25), \\ &f_5 \equiv f(33), f_6 \equiv f(4) \quad \text{mod } 57. \end{aligned}$$

Since $f(7x) \equiv 7f(x)$ and $f(-x) \equiv f(x) \pmod{57}$, the following congruences are readily verified by use of the table:

$$(10.7) \quad (7^2 - 7)f_0 \equiv -F(6), \quad (7^2 - 7)f_5 \equiv F(16) \quad \text{mod } 57.$$

(10.8) Congruences modulo 57:

$$\begin{aligned} (0) \quad 7f_0 - 7f_2 &\equiv -F(7); \quad 7^2f_0 - f_2 \equiv F(5); \quad f_0 - f_3 \equiv -F(1); \\ (1) \quad f_1 - 7f_2 &\equiv F(8); \quad 7^2f_1 - 7^2f_3 \equiv -F(11); \quad 7f_1 - 7^2f_4 \equiv -F(20); \\ f_1 - f_5 &\equiv -F(9); \quad 7^2f_1 - f_5 \equiv F(10); \quad 7f_1 - 7^2f_6 \equiv F(19); \\ (2) \quad 7^2f_2 - 7f_3 &\equiv -F(13); \quad 7^2f_2 - 7^2f_3 \equiv F(12); \quad f_2 - f_6 \equiv F(4); \\ (3) \quad f_3 - 7f_4 &\equiv -F(2); \quad 7f_3 - 7f_6 \equiv -F(14); \\ (4) \quad f_4 - 7^2f_5 &\equiv F(17); \quad f_4 - 7^2f_6 \equiv -F(18); \quad 7f_4 - f_6 \equiv -F(3); \\ (5) \quad 7f_5 - 7f_6 &\equiv F(15). \end{aligned}$$

The next lemma essentially assures us that, in considering condition (C) for the essential values

$$x \equiv 34^n \pmod{43}, \quad 0 \leq n \leq 6,$$

it will not be necessary to add any further congruences to the congruences (10.7), (10.8) which were appropriate for $n = 0$:

LEMMA 10.2. *If there exists a difference set defined by D_7 and f , where (modulo 57)*

$$(10.9) \quad f_i \equiv a_i, \quad 0 \leq i \leq 6,$$

then there also exists an equivalent difference set corresponding to

$$(10.10) \quad f_i \equiv a_{i+1}, \quad 0 \leq i \leq 5,$$

$$(10.11) \quad f_6 \equiv 7a_0.$$

COROLLARY. *The values, reduced mod 3, of f_0, f_1, \dots, f_6 may be permuted cyclically.*

Proof. Suppose that f satisfies (10.9). By Lemma 3.1, we may replace f by a function g defined as follows:

$$g(x) \equiv f(34x) \pmod{57}$$

for all $x \not\equiv 0 \pmod{43}$. Defining g_i as in (10.5) and recalling that $34^7 \equiv 7 \pmod{43}$, we see that (10.10), (10.11) hold with f replaced by g . This proves Lemma 10.2. The corollary reflects the fact that $7 \equiv 1 \pmod{3}$.

Now we apply the corollary to get the following:

LEMMA 10.3. *We may assume without loss of generality that*

$$(10.12) \quad f_0 \equiv f_1 \equiv 2, \quad f_2 \equiv f_4 \equiv f_6 \equiv 0, \quad f_3 \equiv f_5 \equiv 1 \pmod{3}.$$

Proof. If f is to define a difference set, the seven integers

$$f_0, f_1, \dots, f_6$$

must (in particular) represent the seven distinct cycles of K . Of these cycles, the number congruent to 0, 1, 2 mod 3 is 3, 2, 2 respectively. Hence the values of f_0, f_1, \dots, f_6 , when reduced mod 3, must form an ordered sequence

$$(10.13) \quad a_0 a_1 a_2 a_3 a_4 a_5 a_6$$

in which three of the a 's are 0, two are 1 and two are 2. The proof of Lemma 10.3 consists in a demonstration that (apart from the cyclic permutations allowed by the corollary to Lemma 10.2) the only permissible sequence (10.13) is

$$(10.14) \quad 2201010.$$

We note by (10.7) that 6 and 16 are zeros mod 3 of F independently of the choice of f_0 . By this and Lemma 10.1, *the correct number of zeros mod 3 to be obtained from (10.8) is as follows:*

$$f_0 \pmod{3}: 0 \ 1 \ 2$$

$$\text{Correct No. of Zeros: } 4 \ 6 \ 3.$$

In the sequel we consider only zeros (mod 3) obtained from (10.8). We first prove:

$$(10.15) \quad f_0 \equiv f_1 \equiv f_2 \equiv 0 \pmod{3} \quad \text{is impossible.}$$

Suppose the contrary. Then, by (0) of (10.8), 7 and 5 are zeros; by (1), 8 and 11 are zeros; and we have reached the allowable total of 4 zeros. Thus, by (3), $f_3 \not\equiv f_4, f_6 \pmod{3}$. This forces $f_4 \equiv f_6 \pmod{3}$, and then (4) contributes the zeros 18, 3, a contradiction. Next we prove:

$$(10.16) \quad f_0 \equiv f_2 \equiv f_3 \equiv 0 \pmod{3} \quad \text{is impossible.}$$

Again suppose the contrary. Then from (0) we get the zeros 7, 5, 1 and from (1) we get the zeros 13, 12; which makes too many.

The next case is more difficult:

$$(10.17) \quad f_0 \equiv f_3 \equiv f_4 \equiv 0 \pmod{3} \quad \text{is impossible.}$$

If the contrary is true, we make a cyclic transformation to the case

$$(a) \quad f_1 \equiv f_4 \equiv f_5 \equiv 0 \pmod{3}.$$

In case (a), (1) contributes the zeros 20, 9, 10, and (4), the zero 17. Since

$f_0 \not\equiv 0 \pmod{3}$, we conclude that $f_0 \equiv 1 \pmod{3}$. Then a further transformation yields the case

$$(b) \quad f_6 \equiv f_2 \equiv f_3 \equiv 0, \quad f_5 \equiv 1 \pmod{3}.$$

In case (b), (2) contributes the zeros 13, 12, 4, and (3), the zero 14. Again we conclude that $f_0 \equiv 1$, whence

$$f_0 \equiv f_5 \equiv 1, \quad f_1 \equiv f_4 \equiv 2 \pmod{3}.$$

We should have two more zeros, but all we get is the zero 20, from (1). This is a contradiction, proving (10.17).

A similar proof will do for the next case, but we shall give an alternative form :

$$(10.18) \quad f_0 \equiv f_4 \equiv f_5 \equiv 0 \pmod{3} \text{ is impossible.}$$

Suppose the contrary. Then we must have one of

- (i) $f_1 \equiv f_2, \quad f_3 \equiv f_6 \pmod{3},$
- (ii) $f_1 \equiv f_3, \quad f_2 \equiv f_6 \pmod{3},$
- (iii) $f_1 \equiv f_6, \quad f_2 \equiv f_3 \pmod{3}.$

In case (i) the zeros are 8, 14, 17; which is too few. In case (ii) the zeros are 11, 4, 17; again too few. In case (iii) we first transform to

$$f_2 \equiv f_6 \equiv f_0 \equiv 0, \quad f_3 \equiv f_1, \quad f_4 \equiv f_5 \pmod{3}.$$

Then the zeros are 7, 5, 11, 4, 17; which is too many. This proves (10.18).

Combining (10.15)–(10.18), we deduce that (cyclically speaking) no two 0's can be adjacent in a sequence (10.13). Equivalently, we may assume that

$$(10.19) \quad f_2 \equiv f_4 \equiv f_6 \equiv 0 \pmod{3}.$$

Along with (10.19) we must have one of

- (p) $f_0 \equiv f_3, \quad f_1 \equiv f_5 \pmod{3},$
- (q) $f_0 \equiv f_5, \quad f_1 \equiv f_3 \pmod{3},$
- (r) $f_0 \equiv f_1, \quad f_3 \equiv f_5 \pmod{3}.$

In case (p) the six zeros are 1, 9, 10, 4, 18, 3; whence

$$f_0 \equiv f_3 \equiv 1, \quad f_1 \equiv f_5 \equiv 2 \pmod{3}.$$

We transform to the form

$$f_3 \equiv f_5 \equiv f_0 \equiv 0, \quad f_1 \equiv f_4 \equiv 1, \quad f_2 \equiv f_6 \equiv 2 \pmod{3}.$$

Then the zeros are 1, 20, 4; which is too few. In case (q) the zeros are 11, 4, 18, 3; contradicting the fact that $f_0 \equiv 1$ or $2 \pmod{3}$. In case (r) the zeros are 4, 18, 3. This forces

$$f_0 \equiv f_1 \equiv 2, \quad f_3 \equiv f_5 \equiv 1 \pmod{3}$$

and completes the proof of Lemma 10.3.

It will be convenient to recall at this stage that condition (C) for $x = 1$ amounts to the statement that *the 57 integers contained in*

$$(10.20) \quad (D_7 - f_0) \cup (f_0 - D_7) \cup \{\pm F(y) \mid 1 \leq y \leq 20\} \cup 0$$

form a complete set of residues mod 57. We shall exploit the fact that a knowledge of the values of the $f_i \bmod 3$ allows us to split (10.20) into three subclasses of 19 integers congruent mod 3 to 0, 1, 2 respectively.

LEMMA 10.4. *In addition to the congruences (10.12) of Lemma 10.3, we may assume without loss of generality that*

$$(10.21) \quad f_0 \equiv 2 \pmod{57}.$$

Proof. We assume (10.12). Then we see from the form of K that f_0 must be in $2B$ or $23B$. By this and Lemma 3.1, we may assume either (10.21) or

$$(10.22) \quad f_0 \equiv 23 \pmod{57}.$$

Here we assume (10.22) and derive a contradiction to condition (C).

By (10.22), the integers mod 57 congruent to zero mod 3 in $\pm(D_7 - f_0)$ are 6, 42, 12, 18, 51, 15, 39, 45. (These have been arranged for comparison with the columns of table (III).) By (10.22), (10.7) and table (II), the integers $\pm F(6) \bmod 57$ are 3, 54. Deleting these ten numbers from table (III) we get the incomplete table

(III')	---	---	---	24	48	---
	21	---	27	---	---	---
	33	9	---	36	---	30

From (10.7) we have an additional integer, $F(16) \equiv (7^2 - 7)f_5$. Taking this along with the integers in (10.8) which are divisible by 3 (these are determined using (10.12)) we see from (10.20) that *the integers*

$$(10.23) \quad (7^2 - 7)f_5, \quad f_2 - f_6, \quad f_4 - 7^2f_6, \quad 7f_4 - f_6$$

and their negatives must fill out (III') modulo 57. We know in addition, from the form of K and from (10.12), that f_1 is in $2B$, that f_2, f_4, f_6 represent $3B, 24B, 48B$ in some order and that f_3, f_5 represent $B, 16B$ in some order.

Since

$$7(f_4 - 7^2f_6) \equiv 7f_4 - f_6 \pmod{57},$$

we see from (III') that we must have

$$(10.24) \quad f_4 - 7^2f_6 \equiv 21 \text{ or } 36 \pmod{57}.$$

Since $-21 \equiv 36 \pmod{57}$, we must also have

$$(7^2 - 7)f_5 \equiv 9, 48, 27 \text{ or } 30 \pmod{57}.$$

Then, since f_5 is in B or $16B$, we verify easily that one of the following must hold :

- (a) f_5 in B ; $f_5 \equiv 7$; $f_2 - f_6 \equiv 27$ or $30 \pmod{57}$;
- (b) f_5 in $16B$; $f_5 \equiv 55$; $f_2 - f_6 \equiv 9$ or $48 \pmod{57}$.

We rule out case (b) by considering some of the elements congruent to $1 \pmod{3}$ in (10.20) as follows : We have

$$-F(9) \equiv f_1 - f_5 \equiv f_1 + 2, \quad F(10) \equiv 7^2 f_1 - f_5 \equiv 7^2 f_1 + 2 \pmod{57}.$$

Moreover, since $f_0 \equiv 23$ is in $23B$, f_1 must be in $2B$ and, consequently, one of $f_1 + 2$, $7^2 f_1 + 2$ must be congruent mod 57 to 4 or 16. Since 19, 39 are in D_7 and since

$$f_0 - 19 \equiv 4, \quad 39 - f_0 \equiv 16 \pmod{57},$$

we have a contradiction to condition (C). This leaves (a), or

$$(10.25) \quad f_5 \equiv 7; \quad f_2 - f_6 \equiv 27 \text{ or } 30 \pmod{57}.$$

Next we let f_6 range over $3B$, $24B$, $48B$, solve in each case for f_2 , f_4 from (10.25), (10.24), and insist that f_2 , f_4 , f_6 represent $3B$, $24B$, $48B$ in some order. This leaves the following possibilities :

- (p) $f_2 \equiv 21$; $f_4 \equiv 36$; $f_6 \equiv 48 \pmod{57}$;
- (q) $f_2 \equiv 48$ or 51 ; $f_4 \equiv 24$; $f_6 \equiv 21 \pmod{57}$.

In case (q) we have

$$-F(5) \equiv f_2 - 7^2 f_0 \equiv f_2 - 44 \equiv 4 \text{ or } 7 \pmod{57}.$$

Since 19, 30 are in D_7 and since

$$f_0 - 19 \equiv 4, \quad 30 - f_0 \equiv 7 \pmod{57},$$

both choices of f_2 in (q) are inadmissible.

In case (p) we have

$$F(17) \equiv f_4 - 7^2 f_5 \equiv 36 - 1 \equiv 35 \equiv f_0 - 45 \pmod{57}.$$

Since 45 is in D_7 , we have reached the final contradiction which proves Lemma 10.4.

THEOREM 10.1. *Every difference set D_{49} is equivalent to one defined by D_7 and f , where (modulo 57)*

$$(10.26) \quad \begin{aligned} f(1) &\equiv 2, & f(34) &\equiv 44, & f(38) &\equiv 3, & f(2) &\equiv 7, \\ f(25) &\equiv 24, & f(33) &\equiv 43, & f(4) &\equiv 48. \end{aligned}$$

Proof. By Lemma 10.4, we may assume (10.12), (10.21). The proof

begins in the same way as that of Lemma 10.4. Since $f_0 \equiv 2 \pmod{57}$, we find that *the integers (10.23) and their negatives must fill out the table*

$$\begin{array}{cccccc} -- & 6 & 12 & -- & 48 & 39 \\ -- & -- & -- & -- & 51 & 45 \\ -- & 9 & 18 & -- & -- & -- \end{array}$$

modulo 57. This yields the cases

- (I) $f_4 - 7^2 f_6 \equiv 18 \text{ or } 39 \pmod{57}$,
 (II) $f_4 - 7^2 f_6 \equiv 9 \text{ or } 48 \pmod{57}$.

We begin by disposing of case (I). Case (I) implies

$$(7^2 - 7)f_5 \equiv 6, 51, 9 \text{ or } 48 \pmod{57}.$$

From this and the fact that f_5 is in B or $16B$ we obtain the following possibilities :

$$(a) f_5 \equiv 7; \quad (b) f_5 \equiv 49 \pmod{57}.$$

We rule out case (a) as follows : We have

$$-F(9) \equiv f_1 - f_5 \equiv f_1 - 7, \quad F(10) \equiv 7^2 f_1 - 7 \pmod{57}.$$

Since f_1 is in $23B$, one of these values must be 37 or 40. However, D_7 contains 39, 19, and

$$39 - f_0 \equiv 37, \quad f_0 - 19 \equiv 40 \pmod{57},$$

a contradiction. This leaves case (b), whence

$$(10.27) \quad f_5 \equiv 49; \quad f_2 - f_6 \equiv 9 \text{ or } 48 \pmod{57}.$$

Letting f_6 range over $3B$, $24B$, $48B$, we find that the only admissible values for f_2, f_4, f_6 are given by

$$(10.28) \quad f_2 \equiv 24, \quad f_4 \equiv 33, \quad f_6 \equiv 15 \pmod{57}.$$

But then

$$F(7) \equiv 7f_2 - 7f_0 \equiv 54 - 14 \equiv 40 \equiv f_0 - 19 \pmod{57},$$

a contradiction which rules out case (I).

Now we turn to case (II). Here

$$(7^2 - 7)f_5 \equiv 12, 45, 18 \text{ or } 39 \pmod{57}$$

and hence

$$f_5 \equiv 16 \text{ or } 43 \pmod{57}.$$

Therefore we have the two cases :

- (P) $f_5 \equiv 16; \quad f_2 - f_6 \equiv 18 \text{ or } 39; \quad f_4 - 7^2 f_6 \equiv 9 \text{ or } 48 \pmod{57}$,
 (Q) $f_5 \equiv 43; \quad f_2 - f_6 \equiv 12 \text{ or } 45; \quad f_4 - 7^2 f_6 \equiv 9 \text{ or } 48 \pmod{57}$.

In case (P), a search for f_2, f_4, f_6 yields only

$$(P.1) \quad f_5 \equiv 16, \quad f_2 \equiv 15, \quad f_4 \equiv 33, \quad f_6 \equiv 54 \pmod{57},$$

$$(P.2) \quad f_5 \equiv 16, \quad f_2 \equiv 54, \quad f_4 \equiv 3, \quad f_6 \equiv 15 \pmod{57}.$$

Since (P.2) implies

$$-F(17) \equiv 7^2f_5 - f_4 \equiv 43 - 3 \equiv 40 \equiv f_0 - 19,$$

it is impossible. On the other hand, (P.1) implies

$$-F(5) \equiv f_2 - 7^2f_0 \equiv 15 - 41 \equiv 31 \pmod{57};$$

and we observe that $D_7 - f_0$ contains

$$30 - 2 \equiv 28 \pmod{57}.$$

For each choice of f_1 in $23B$, one of

$$-F(9) \equiv f_1 - 16, \quad F(10) \equiv 7^2f_1 - 16$$

is congruent mod 57 to 28 or 31. This disposes of case (P).

We are left with case (Q), which develops into the following:

$$(Q.1) \quad f_5 \equiv 43, \quad f_2 \equiv 48, \quad f_4 \equiv 24, \quad f_6 \equiv 3 \pmod{57},$$

$$(Q.2) \quad f_5 \equiv 43, \quad f_2 \equiv 15, \quad f_4 \equiv 24, \quad f_6 \equiv 3 \pmod{57},$$

$$(Q.3) \quad f_5 \equiv 43, \quad f_2 \equiv 3, \quad f_4 \equiv 24, \quad f_6 \equiv 48 \pmod{57}.$$

We dispose of (Q.1) on the grounds that

$$F(7) \equiv 37 \equiv 39 - f_0 \pmod{57}.$$

And we dispose of (Q.2) on the grounds that

$$F(7) \equiv 31 \equiv -F(5) \pmod{57}.$$

This leaves (Q.3), or

$$(10.29) \quad f_0 \equiv 2, \quad f_2 \equiv 3, \quad f_4 \equiv 24, \quad f_5 \equiv 43, \quad f_6 \equiv 48 \pmod{57}.$$

At this stage we pick out of (10.20) the 19 integers congruent to 1 mod 3 and substitute from (10.29). We find that the integers are

$$7, 19, 22, 28, 31, 37, 40, 43$$

together with

$$\begin{aligned} f_1 - 43, \quad 21 - f_1, \quad 15 - 7f_1, \quad 36 - 7f_1, \quad 7^2f_1 - 43, \\ f_3 - 54, \quad 2 - f_3, \quad 7f_3 - 33, \quad 7f_3 - 51, \quad 7^2f_3 - 33, \\ 7^2f_1 - 7^2f_3. \end{aligned}$$

We recall that f_1, f_3 must be in $23B, B$, respectively. Since

$$f_1 \equiv 23 \pmod{57} \quad \text{implies} \quad f_1 - 43 \equiv 37 \pmod{57};$$

$$f_1 \equiv 47 \pmod{57} \quad \text{implies} \quad 15 - 7f_1 \equiv 28 \pmod{57};$$

$$f_3 \equiv 1 \pmod{57} \quad \text{implies} \quad 7f_3 - 33 \equiv 31 \pmod{57};$$

$$f_3 \equiv 49 \pmod{57} \quad \text{implies} \quad 7^2f_3 - 33 \equiv 31 \pmod{57};$$

we are forced to conclude that

$$(10.30) \quad f_1 \equiv 44, \quad f_3 \equiv 7 \pmod{57}.$$

In view of (10.6), the congruences (10.29), (10.30) imply the congruences (10.26) of Theorem 10.1. That we actually have a difference set may be verified by examining (10.20) for each of the seven cases obtained by applying cyclic transformations as explained in Lemma 10.2. This completes the proof of Theorem 10.1.

The corresponding difference set is given as follows:

$$(10.31) \quad D_{49} \text{ modulo } 2451 :$$

$$\begin{array}{ccccccccc} 817; & 1634; & 215 & 1505 & 731; & 1290 & 1677 & 1935; \\ 6 & 42 & 294 & 2058 & 2151 & 351; & 29 & 203 & 1421 & 143 & 1001 & 2105; \\ 187 & 1309 & 1810 & 415 & 454 & 727; & 414 & 447 & 678 & 2295 & 1359 & 2160; \\ 513 & 1140 & 627 & 1938 & 1311 & 1824; & 597 & 1728 & 2292 & 1338 & 2013 & 1836; \\ & & 883 & 1279 & 1600 & 1396 & 2419 & 2227. \end{array}$$

11. $m = 9$. Here $s = 73$, $t = 91$. In particular (as for the case $m = 7$, considered in §10) s is a prime. For this reason the discussion is similar in some respects to that of §10. As a consequence we may and do omit some parts of the proof.

We may assume that D_9 is given by (8.2) and hence is fixed by all multipliers, namely by the powers of (3). The set $K = K(D_9)$ consists of the following six cycles of length six under multiplication mod 91 by 3:

$$\begin{aligned} (a) : 40 & 29 & 87 & 79 & 55 & 74; \quad (a') : 53 & 68 & 22 & 66 & 16 & 48; \quad (b) : 15 & 45 & 44 & 41 & 32 & 5; \\ (c) : 43 & 38 & 23 & 69 & 25 & 75; \quad (d) : & 8 & 24 & 72 & 34 & 11 & 33; \quad (d') : 60 & 89 & 85 & 73 & 37 & 20. \end{aligned}$$

Reducing these cycles modulo 13, we find that

$$(11.1) \quad (a) \equiv (a') \equiv P, \quad (b) \equiv 2P, \quad (c) \equiv 4P, \quad (d) \equiv (d') \equiv 8P \pmod{13}$$

where

$$(11.2) \quad P \equiv \{1, 3, 9\} \pmod{13}.$$

Somewhat as in §10, we shall find it convenient to work in part modulo 13 instead of modulo 91.

As in §10, we define

$$(11.3) \quad F(y) \equiv f(1 + y) - f(y) \pmod{91}$$

for all $y \not\equiv 0, 72 \pmod{73}$. By the properties of f ,

$$(11.4) \quad F(36) \equiv 0 \pmod{91},$$

$$(11.5) \quad F(72 - y) \equiv -F(y) \pmod{91}, \quad 1 \leq y \leq 35.$$

In contrast to Lemma 10.1, the following lemma gives the number of zeros of $F \bmod 13$ according to which of the sets (11.1) contains $f(1)$ reduced mod 13:

LEMMA 11.1. *In terms of the cycle mod 13 containing $f(1)$, the following table gives the correct number of zeros of $F \bmod 13$ in the range $1 \leq y \leq 35$:*

$f(1) \text{ in:}$	P	$2P$	$4P$	$8P$
No. of Zeros:	3	1	3	2.

Proof. Omitted.

Next we observe that 5 is a primitive root modulo the prime $s = 73$ and that the powers $5^n \bmod 73$, $0 \leq n \leq 35$, are given by the following table:

1	5	25	52	41	59	
3	15	2	10	50	31	
9	45	6	30	4	20	
27	62	18	17	12	60	(mod 73).
8	40	54	51	36	34	
24	47	16	7	35	29	

Each column of the table is half of a cycle under multiplication mod 73 by 3. Accordingly, we define

$$(11.6) \quad f(1) \equiv f_0, \quad f(5) \equiv f_1, \quad f(25) \equiv f_2, \quad f(52) \equiv f_3, \quad f(41) \equiv f_4, \quad f(59) \equiv f_5 \quad \text{mod } 91$$

and verify the following:

(11.7)

$$\begin{aligned} (3^4 - 3^2)f_0 &\equiv -F(8), & (3^4 - 3^3)f_2 &\equiv F(18), & (3^4 - 1)f_3 &\equiv F(21), \\ (3^5 - 3^4)f_4 &\equiv -F(35), & (3^3 - 1)f_5 &\equiv -F(13) && \text{mod } 91; \end{aligned}$$

(11.8) Modulo 91:

- | | | | |
|-----|-----------------------------------|-----------------------------------|-----------------------------------|
| (0) | $3^3f_0 - 3^2f_1 \equiv -F(27)$, | $3^3f_0 - 3^5f_1 \equiv F(26)$; | $f_0 - 3f_2 \equiv -F(1)$, |
| | $3f_0 - 3f_2 \equiv F(2)$, | $3^5f_0 - f_2 \equiv -F(24)$; | $3^4f_0 - 3^5f_3 \equiv F(7)$, |
| | $3^2f_0 - 3f_3 \equiv -F(9)$; | $3f_0 - 3^2f_4 \equiv -F(3)$, | $3^5f_0 - 3f_4 \equiv F(23)$; |
| (1) | $f_1 - 3^2f_2 \equiv -F(5)$, | $3f_1 - 3^5f_2 \equiv -F(15)$, | $3^5f_1 - f_2 \equiv F(25)$; |
| | $3^3f_1 - 3f_3 \equiv F(10)$; | $f_1 - 3^2f_4 \equiv F(4)$, | $3^3f_1 - 3^3f_4 \equiv -F(11)$, |
| | $3^4f_1 - f_4 \equiv F(32)$; | $3f_1 - f_5 \equiv F(14)$, | $3^2f_1 - 3^5f_5 \equiv -F(28)$, |
| | $3^4f_1 - 3^4f_5 \equiv -F(33)$; | | |
| (2) | $3^2f_2 - 3^5f_3 \equiv -F(6)$, | $3^5f_2 - 3^3f_3 \equiv -F(16)$, | $3^3f_2 - 3^3f_3 \equiv F(17)$; |
| | $3^4f_2 - 3^2f_5 \equiv -F(19)$; | | |
| (3) | $3^4f_3 - 3f_4 \equiv -F(22)$; | $f_3 - 3^2f_5 \equiv F(20)$, | $3^2f_3 - 3^5f_5 \equiv F(29)$, |
| | $3^2f_3 - 3f_5 \equiv -F(30)$; | | |
| (4) | $3^3f_4 - 3^3f_5 \equiv -F(12)$, | $f_4 - 3f_5 \equiv F(31)$, | $3^5f_4 - 3^4f_5 \equiv F(34)$. |

The most useful form of the analogue of Lemma 10.2 may be given as follows :

LEMMA 11.2. *If some row of the following table gives (modulo 91) an admissible set of values of the f_i , then so does every row:*

	f_0	f_1	f_2	f_3	f_4	f_5
(0)	a_0	a_1	a_2	a_3	a_4	a_5
(1)	a_1	a_2	a_3	a_4	a_5	$3a_0$
(2)	a_2	a_3	a_4	a_5	$3a_0$	$3a_1$
(3)	a_3	a_4	a_5	$3a_0$	$3a_1$	$3a_2$
(4)	a_4	a_5	$3a_0$	$3a_1$	$3a_2$	$3a_3$
(5)	a_5	$3a_0$	$3a_1$	$3a_2$	$3a_3$	$3a_4$

Proof. Omitted.

LEMMA 11.3. *We may assume without loss of generality that*

$$(11.9) \quad f_0 \in 2P, \quad f_1 \in 8P, \quad f_2 \in 4P, \quad f_3 \in P, \quad f_4 \in 8P, \quad f_5 \in P \pmod{13},$$

$$(11.10) \quad f_1 \equiv 3f_4, \quad f_3 \equiv 3f_5 \pmod{13}.$$

Proof. We regard each row of the table of Lemma 11.2 as an admissible set of values of the f_i (reduced, however, modulo 13 instead of modulo 91). In the course of the present proof the phrase “row (j) ” should be understood to refer to Lemma 11.2 and not to the formulas (11.8). We may and do assume without loss of generality that

$$(11.11) \quad a_0 \in 2P \pmod{13}.$$

We consider the various possible arrangements of the a 's mod 13 among the sets $P, 2P, 4P, 8P$ (keeping in mind the form of K and, more specifically, formula (11.1)) and, using rows (0)–(5) of the table as convenient, check the number of zeros mod 13 among the $F(y)$ given in (11.8). Since (11.7) yields precisely one zero mod 13, namely the zero 13, *the correct number of zeros mod 13 of F obtainable from (11.8) is given as follows:*

f_0 in :	P	$2P$	$4P$	$8P$
Correct No. of Zeros :	2	0	2	1

We first prove :

$$(11.12) \quad a_5 \in 4P, \quad a_2 \in a_1P, \quad a_4 \in a_3P \pmod{13} \quad \text{is impossible.}$$

To see this we assume the contrary. First we use row (0). Since $f_0 = a_0 \in 2P \pmod{13}$, there must be no zeros. Hence in particular,

$$a_1 - 3^2 a_2 \not\equiv 0, \quad 3a_1 - 3^5 a_2 \not\equiv 0 \pmod{13}.$$

Since 3 has exponent 3 mod 13 and since $a_2 \in a_1P \pmod{13}$, we conclude that

$$a_1 \equiv a_2 \pmod{13}.$$

Similarly

$$a_3 \equiv 3a_4 \text{ or } 3^2 a_4 \pmod{13}.$$

Next we apply row (5) and expect two zeros. However, since in this case

$$\begin{aligned} a_1 - a_2 &\equiv -F(6), \quad a_1 - 3a_2 \equiv -F(16), \quad 3a_1 - 3a_2 \equiv F(17), \\ 3a_3 - 3a_4 &\equiv -F(12), \quad 3a_3 - 3^2a_4 \equiv F(31), \quad a_3 - 3^2a_4 \equiv F(34) \end{aligned}$$

modulo 13, we actually get three zeros, namely 6, 17 and one of 31, 34. This contradiction proves (11.12). Next:

$$(11.13) \quad a_5 \in 4P, \quad a_4 \in a_1P, \quad a_3 \in a_2P \pmod{13} \quad \text{is impossible.}$$

We assume the contrary to (11.13). By row (0),

$$a_1 \equiv 3a_4, \quad a_2 \equiv 3^2a_3 \pmod{13}.$$

Next we use row (5) and expect two zeros but find only one, namely 19. This proves (11.13).

$$(11.14) \quad a_5 \in 4P \pmod{13} \quad \text{is impossible.}$$

If (11.14) is false, then, by (11.12), (11.13), we must have

$$a_3 \in a_1P, \quad a_4 \in a_3P \pmod{13}.$$

From this and row (0) we deduce that

$$a_1 \equiv a_3 \text{ or } 3^2a_3 \pmod{13}.$$

Then, using row (5) and insisting on two zeros, we see that these must be 20, 30; and this requires

$$a_2 \equiv 3^2a_4 \pmod{13}.$$

Next we use row (1). Since $f_0 = a_1$, we should have two zeros or one zero according as a_1 is in P or $8P$. The zeros must come from

$$\begin{aligned} a_1 - 3a_3 &\equiv -F(1), \quad 3a_1 - 3a_3 \equiv F(2), \quad 3^2a_1 - a_3 \equiv -F(24), \\ a_2 - 3a_4 &\equiv F(10) \pmod{31}. \end{aligned}$$

There is at most one zero, namely 2. Hence we must insist that

$$a_1 \equiv a_3 \in 8P, \quad a_2 \equiv 3^2a_4 \in P \pmod{13}.$$

Finally we use row (3). Since $f_0 = a_3 \in 8P$, there should be exactly one zero. But in fact we get three zeros, namely 23, 28, 33. This contradiction proves (11.14).

Next we prove

$$(11.15) \quad a_5 \in a_4P \pmod{13} \quad \text{is impossible.}$$

Indeed, if the contrary is true and if we use row (0), F should have no zero from (11.8) but in fact will have at least one zero, namely one of 12, 31, 34. Next:

$$(11.16) \quad a_4 \in 4P, \quad a_5 \in a_1P, \quad a_3 \in a_2P \pmod{13} \quad \text{is impossible.}$$

For if we deny (11.16) and use row (3), we should get at least one zero but, in fact, find none at all. Next:

$$(11.17) \quad a_4 \in 4P, \quad a_5 \in a_3P, \quad a_2 \in a_1P \pmod{13} \quad \text{is impossible.}$$

If we deny (11.17) and use row (0), we conclude that

$$a_1 \equiv a_2; \quad a_3 \equiv 3a_5 \pmod{13}.$$

But then we use row (1) and expect at least one zero but find none. We are now ready for

$$(11.18) \quad a_4 \in 4P \pmod{13} \quad \text{is impossible.}$$

In view of (11.16), (11.17), if (11.18) is false we must have

$$a_5 \in a_2P, \quad a_3 \in a_1P \pmod{13}.$$

By use of row (0), we conclude that

$$a_1 \equiv a_3 \text{ or } 3^2a_3; \quad a_2 \equiv a_5 \text{ or } 3^2a_5 \pmod{13}.$$

Next we turn to row (4). Since $f_0 = a_4 \in 4P$, we must have exactly two zeros from (10.8). The relevant congruences are

$$\begin{aligned} a_1 - a_3 &\equiv F(29), & a_1 - 3^2a_3 &\equiv 3^2F(20) \equiv -F(30), \\ a_2 - 3^2a_5 &\equiv 3^2F(11), & a_5 - a_2 &\equiv 3^2F(32) \equiv F(4), \end{aligned}$$

and the requirement of exactly two zeros forces

$$a_1 \equiv a_3; \quad a_2 \equiv 3^2a_5 \pmod{13}.$$

Finally we use row (1). Here there should be at most two zeros but there are certainly three, namely 2, 4, 32. Now we have a contradiction which proves (11.18). Next:

$$(11.19) \quad a_1 \in 4P, \quad a_5 \in a_2P, \quad a_4 \in a_3P \pmod{13} \quad \text{is impossible.}$$

Suppose that (11.19) is false and use row (3). Since $f_0 = a_3$, we must have at least one and at most two zeros. However, the zeros come from

$$3a_5 - a_2 \equiv -F(19), \quad a_3 - 3^2a_4 \equiv -F(27) \equiv F(26) \pmod{13}.$$

Hence we must have one of the following :

- (i) $a_2 \equiv 3a_5 \in P; \quad a_3 \equiv a_4 \text{ or } 3a_4 \pmod{13},$
- (ii) $a_3 \equiv 3^2a_4 \in P; \quad a_2 \equiv a_5 \text{ or } 3^2a_5 \pmod{13}.$

Now we use row (0). In this case there can be no zeros, whence (i) is ruled out and (ii) must hold. Thus, also, $a_2 \in 8P$. Therefore, if we use row (2), we should get just one zero. Since, for row (2),

$$a_3 - 3^2a_4 \equiv -F(5), \quad a_2 - 3^2a_5 \equiv -F(9) \pmod{13},$$

we see that 5 is a zero and hence that 9 cannot be; and this forces

$$a_2 \equiv a_5 \pmod{13}.$$

Now we use row (1). There should be two zeros but the only zero is 11; a contradiction which proves (11.19).

$$(11.20) \quad a_1 \in 4P \pmod{13} \quad \text{is impossible.}$$

If (11.20) is false, we see by (11.15), (11.19) that necessarily

$$a_5 \in a_3 P, \quad a_4 \in a_2 P \pmod{13}.$$

We use row (1). Since $a_1 \in 4P$, we should get two zeros from (11.8). However, the only possible source of a zero is the congruence

$$a_2 - 3a_4 \equiv F(10) \pmod{13}.$$

Therefore we have a contradiction which proves (11.20). Next:

$$(11.21) \quad a_2 \in 4P, \quad a_5 \in a_1 P, \quad a_4 \in a_3 P \pmod{13} \quad \text{is impossible.}$$

We assume the contrary to (11.21) and use row (4), observing that (11.8) should yield two zeros or one according as a_4 is in P or $8P$. The only source of a zero is the congruence

$$a_5 - 3^2 a_1 \equiv F(10) \pmod{13}.$$

We conclude that

$$a_5 \equiv 3^2 a_1 \in P; \quad a_3, a_4 \in 8P.$$

Now we use row (3). Since a_4 is in $8P$, there should be just one zero. The zeros arise from the congruences

$$a_3 - 3^2 a_4 \equiv -F(27) \equiv F(26),$$

which yield two or none. This contradiction proves (11.21). Next:

$$(11.22) \quad a_3 \in 4P, \quad a_5 \in a_1 P, \quad a_4 \in a_2 P \pmod{13} \quad \text{is impossible.}$$

For if we deny (11.22) and use row (0), we deduce that

$$a_1 \equiv 3a_5 \pmod{13}.$$

Then, on using row (3), we should have two zeros in (11.8), and the source of these is the congruences

$$a_4 - a_2 \equiv 3^2 F(14), \quad 3^2 a_4 - a_2 \equiv -F(28) \equiv -F(33) \pmod{13},$$

whence

$$a_2 \equiv 3^2 a_4 \pmod{13}.$$

If we use row (1) we should get two zeros or one from (11.8) according as a_1 is in P or $8P$. The zeros must arise from the congruences

$$a_1 - 3a_5 \equiv -3^2 F(3), \quad a_1 - 3^2 a_5 \equiv 3F(23), \quad a_2 - 3a_4 \equiv F(10) \pmod{13},$$

whence we see that a_1, a_5 are in $8P$ and a_2, a_4 are in P . Finally, by using row (4) we should get two zeros in (11.8). But the relevant congruences are

$$3a_5 - a_1 \equiv 3F(10), \quad 3a_4 - a_2 \equiv -F(3), \quad a_4 - a_2 \equiv 3F(23) \pmod{13},$$

and these yield only one zero. This completes the proof of (11.22). Next:

$$(11.23) \quad a_3 \in 4P \pmod{13} \quad \text{is impossible.}$$

If (11.23) is false, then, by (11.22), (11.15), we must have

$$a_5 \in a_2 P, \quad a_4 \in a_1 P \pmod{13}.$$

By row (0),

$$a_1 \equiv 3a_4; \quad a_2 \equiv a_5 \text{ or } 3^2a_5 \pmod{13}.$$

Since a_3 is in $4P$, we should get two zeros from (11.8) by using row (3). In fact we get none; a contradiction which proves (11.23).

By (11.14), (11.18), (11.20), (11.23), we must have a_2 in $4P$. And, by (11.15), (11.21), we are reduced to the case

$$(11.24) \quad a_0 \in 2P, \quad a_2 \in 4P, \quad a_4 \in a_1 P, \quad a_5 \in a_3 P \pmod{13}.$$

Using (11.24) along with row (0), we deduce that

$$a_1 \equiv 3a_4; \quad a_3 \equiv 3a_5 \pmod{13}.$$

When we use row (1), the only source of zeros is the congruences

$$a_1 - 3a_4 \equiv 3^2F(7), \quad a_1 - 3^2a_4 \equiv -3F(9) \pmod{13}.$$

Hence 7 is the only zero and

$$a_1 \equiv 3a_4 \in 8P, \quad a_3 \equiv 3a_5 \in P \pmod{13}.$$

This completes the proof of Lemma 11.3. That no contradictions arise from (11.24) will be clear from what follows.

LEMMA 11.4. *Let row (0) of Lemma 11.2 yield an admissible set of values of the f_i satisfying the conditions*

$$(11.25) \quad a_0 \in 2P, \quad a_1 \in 8P, \quad a_2 \in 4P, \quad a_3 \in P, \quad a_4 \in 8P, \quad a_5 \in P \pmod{13},$$

$$(11.26) \quad a_1 \equiv 3a_4, \quad a_3 \equiv 3a_5 \pmod{13}.$$

Assume (as we may without loss of generality) that

$$(11.27) \quad a_0 \equiv 5 \pmod{91}.$$

For $i = 1, 4$ let d_i denote an element of D_9 (uniquely determined by a_i) such that

$$(11.28) \quad d_i - a_i \equiv 0 \pmod{13}.$$

Then each of the following rows (numbered to correspond with the rows of Lemma 11.2) must consist of three integers which, together with their negatives and 0, make up the seven integers mod 91 which are divisible by 13:

(11.29) *Modulo 91:*

$$\begin{aligned} (0) \quad & -F(13) \equiv (3^3 - 1)a_5; & 13; & 26; \\ (1) \quad & 3^2F(13) \equiv 39; \quad 3^2F(7) \equiv a_1 - 3a_4; & & 3^2(d_1 - a_1); \\ (2) \quad & 3^3F(13) \equiv (3^4 - 3)a_1; \quad 3^3F(19) \equiv a_1 - 3a_4; \quad 3^3F(10) \equiv a_3 - 3^4a_5; \\ (3) \quad & -F(13) \equiv (3^4 - 3)a_2; \quad -F(24) \equiv 3^5a_3 - a_5; \quad -F(1) \equiv a_3 - 3a_5; \\ (4) \quad & 3^2F(13) \equiv (3^3 - 1)a_3; \quad -3^2F(7) \equiv 3^2a_1 - a_4; & & 3^2(d_4 - a_4); \\ (5) \quad & 3^3F(13) \equiv (3^4 - 3)a_4; \quad -3^3F(19) \equiv 3^2a_1 - a_4; \quad 3^3F(3) \equiv a_3 - 3^4a_5. \end{aligned}$$

Proof. Since $a_0 \pmod{13}$ is in $2P$, we see from (11.1) that a_0 is in (b). Hence, by Lemma 3.1, we may assume (11.27). For each row of Lemma 11.2 we pick out $F(13)$ from (11.7), those F 's divisible by 13 from (11.8), and those integers divisible by 13 from $D_9 - f_0$, getting three integers in all. We may multiply the three integers by the same power of 3 and we may change the signs of the integers independently without destroying the property stated in Lemma 11.4. After such alterations we exhibit the result in the appropriate row of Lemma 11.4.

THEOREM 11.1. *Every difference set D_{81} is equivalent to one defined by D_9 and f where (modulo 91)*

$$(11.30) \quad \begin{aligned} f(1) &\equiv 5, & f(5) &\equiv 24, & f(25) &\equiv 75, \\ f(52) &\equiv 68, & f(41) &\equiv 60, & f(59) &\equiv 40. \end{aligned}$$

Proof. We begin by working in terms of the rows (0)–(5) of Lemma 11.4. By row (0), since a_5 must be in the cycle (a) or (a') of K , we find that a_5 is restricted to the values

$$(11.31) \quad a_5 \equiv 40, 79; 68, 16 \pmod{91},$$

the first two being in (a) and the others in (a'). We know that a_1, a_4 must represent (d), (d') in some order. We let a_1 range over (d), (d'), evaluate $3^2(d_1 - a_1)$ in each case and test (using row (1)) for admissible values of a_4 . After eliminating inadmissible values we get the following table :

$a_1:$	24	11	85	73
$a_4:$	60	73	24	33
$a_1 - 3a_4:$	26	-26	13	-26
$3^4a_1 - 3a_1:$	-39	39	-13	-39
$3^4a_4 - 3a_4:$	39	-39	-39	26
$3^2a_1 - a_4:$	-26	26	13	-13.

Turning to row (2), we see from the table that $a_1 \equiv 85$ is inadmissible and that a_3, a_5 must satisfy

$$(11.32) \quad a_3 - 3^4a_5 \equiv \pm 13 \pmod{91}.$$

Then turning to row (5), we see from the table and (11.32) that $a_1 \equiv 73$ is inadmissible. Next we recall that a_3, a_5 must represent the cycles (a), (a') in some order. We let a_5 range over (11.31) and test (11.32) for admissible values of a_3 . At this stage we get the two tables

$$(11.33) \quad \begin{aligned} a_1: 24 &11 & a_3: 68 &16 \\ a_4: 60 &73 & a_5: 40 &79 \end{aligned}$$

In row (3), either choice for a_3, a_5 simply requires that

$$(3^4 - 3)a_2 \equiv \pm 26 \pmod{91}.$$

Since a_2 must represent the cycle (c), we deduce that

$$(11.34) \quad a_2 \equiv 38 \text{ or } 75 \pmod{91}.$$

Moreover, if the a 's are given by (11.26), (11.33), (11.34), each of the rows (0)-(5) of Lemma 11.4 has the stated property.

We now remove the ambiguities by evaluating further F 's, using row (0) of Lemma 11.2 and assuming (11.26). We begin with (11.7) and note that $F(8) \equiv 4 \pmod{91}$.

If $a_3 \equiv 16$, then $F(21) \equiv 4 \pmod{91}$.

If $a_4 \equiv 73$, then $F(33) \equiv -4 \pmod{91}$.

Therefore

$$(11.35) \quad a_0 \equiv 5, \quad a_1 \equiv 24, \quad a_3 \equiv 68, \quad a_4 \equiv 60, \quad a_5 \equiv 40 \pmod{91}.$$

In view of (11.35), if $a_2 \equiv 38 \pmod{91}$ then, from (11.8),

$$F(7) \equiv -12 \equiv -F(19) \pmod{91}.$$

This final contradiction shows that

$$(11.36) \quad a_2 \equiv 75 \pmod{91}$$

and completes the proof of Theorem 11.1.

The difference set D_{81} is now easily exhibited.

BIBLIOGRAPHY

1. J. Singer, *A theorem in finite projective geometry and some application to number theory*, Trans. Amer. Math. Soc. vol. 43 (1938) pp. 377–385.
2. Marshall Hall, *Cyclic projective planes*, Duke Math. J. vol. 14 (1947) pp. 1079–1090.
3. H. B. Mann and T. A. Evans, *On simple difference sets*, Sankhyā vol. 11 (1951) pp. 357–364.
4. S. Chowla and H. J. Ryser, *Combinatorial problems*, Canad. J. Math. vol. 2 (1950) pp. 93–99.
5. Marshall Hall and H. J. Ryser, *Cyclic incidence matrices*, Canad. J. Math. vol. 3 (1951) pp. 495–502.
6. H. B. Mann, *Some theorems on difference sets*, Canad. J. Math. vol. 4 (1952) pp. 222–226.
7. T. G. Ostrom, *Concerning difference sets*, Canad. J. Math. vol. 5 (1953) pp. 421–424.
8. Marshall Hall, *A survey of difference sets*, Proc. Amer. Math. Soc. vol. 7 (1956) pp. 975–986.
9. ———, *Uniqueness of the projective plane with 57 points*, Proc. Amer. Math. Soc. vol. 4 (1953) pp. 912–916. Correction, vol. 5 (1954) pp. 994–997.

UNIVERSITY OF WISCONSIN,
MADISON, WISCONSIN

ON HOMOMORPHISMS OF PROJECTIVE PLANES

BY

D. R. HUGHES

1. Introduction. In [4] Klingenberg has considered homomorphisms of projective planes, mostly for the Desarguesian case, and in this paper we show that many of these investigations can be extended to the general case, obtaining analogous results. It is shown that no finite projective plane possesses any homomorphisms other than isomorphisms, and that certain types of collineations (specifically, the central collineations) of a projective plane always induce collineations in any homomorphic image, under certain more or less obvious restrictions.

It seems to the author that this approach offers a possible way to solve some of the outstanding problems in the study of finite projective planes, since it is known that every finite projective plane is a homomorphic image of a free plane [2; 5], and the free planes are known to possess considerably many collineations (the collineation groups of free planes appear to be undetermined at the present time). A theorem stating that every finite projective plane possesses non-trivial collineations would be of immense value in classifying the finite planes, since a good deal is known about finite planes with collineations; the proof of such a theorem must use finiteness in a strong way, since Hall has displayed an infinite projective plane with no non-identity collineations. Under any circumstances, the study of homomorphisms of projective planes appears to be a subject of interest in its own right, whether or not it sheds light on the finite problems.

2. Planar ternary rings. Suppose π is a collection of (undefined) objects called *points* and *lines*, together with an *incidence relation* (i.e., point on line, line contains point, etc.), all satisfying : (i) if P and Q are distinct points of π , then there is exactly one line $L = PQ$ of π which contains both P and Q ; (ii) if K and L are distinct lines of π , then there is exactly one point $P = KL$ of π which lies on both K and L ; (iii) π contains a set of four points no three of which are on a single line (i.e., no three of which are *collinear*). Then we say that π is a *projective plane* (see [2; 6] for more discussion). Then there is a (finite or infinite) cardinal n such that every point is on $n + 1$ lines, every line contains $n + 1$ points, and π contains altogether $n^2 + n + 1$ points and $n^2 + n + 1$ lines; n is called the *order* of π , and π is said to be finite if n is finite. We now proceed to a description of the process of coordinatizing a projective plane; more material on this can be found in [2; 3; 6], but we shall use the technique of [3] exclusively here.

Let π be a projective plane of order n , and let X, Y, O be three non-collinear points in π ; let $L_\infty = XY$, $L_1 = OY$, $L_2 = OX$. Let R be a set

of symbols, of cardinal n , and assume that 0 (zero) and 1 (one) are two distinct symbols of R . We assign “coordinates” $(0, b)$ to every point on L_1 , with the exception of the point Y , the only restriction being that O is assigned the coordinates $(0, 0)$; the point Y will be called (∞) , where “ ∞ ” is some symbol not in R . Let some point other than X or Y be chosen on L_∞ , and call this point (1) . If P is on L_2 , collinear with (1) and $(0, b)$, then the coordinates of P will be $(b, 0)$; the point X is not assigned coordinates by this rule, and we call it (0) . If P is in π , not on L_∞ , and PY contains $(a, 0)$, PX contains $(0, b)$, then the coordinates of P are (a, b) ; if P is on L_∞ , collinear with $(1, 0)$ and $(0, b)$, then P is called (b) . We have now assigned coordinates to every point of π . If L is a line of π , not containing Y , then if L contains (m) and $(0, k)$, L is assigned the coordinates $[m, k]$; $[\infty, (k, 0)]$ is the line through Y and $(0, k)$, while finally, L_∞ is called merely L_∞ . Now every line of π has also been assigned coordinates.

We define a *ternary function* F on R as follows: if (x, y) is on $[m, k]$, then $F(m, x, y) = k$. From the fact that π is a projective plane, it is easy to verify the following:

- (A) $F(a, 0, c) = F(0, b, c) = c$, for all $a, b, c \in R$.
- (B) $F(a, 1, 0) = F(1, a, 0) = a$, for all $a \in R$.
- (C) If $a, b, c, d \in R$, $a \neq c$, then there is a unique $x \in R$ such that $F(x, a, b) = F(x, c, d)$.
- (D) If $a, b, c, d \in R$, $a \neq c$, then there is a unique (ordered) pair $x, y \in R$ such that $F(a, x, y) = b$, $F(c, x, y) = d$.
- (E) If $a, b, c \in R$, then there is a unique $x \in R$ such that $F(a, b, x) = c$.

In general, if R is a non-empty set and F is a ternary function from R to R (i.e., $F(a, b, c) \in R$ for all $a, b, c \in R$), we say that (R, F) is a *ternary ring*. If (R, F) is a ternary ring containing at least the two distinct elements 0 and 1, and satisfying (A)–(E), then (R, F) is a *planar ternary ring*. It is well-known that a planar ternary ring defines a unique projective plane, in the obvious fashion (see [2; 6]).

Let (R, F) be a planar ternary ring. For all $a, b \in R$, we define $a + b = F(1, a, b)$ and $a \cdot b = ab = F(a, b, 0)$. Then R , under the operation $(+)$, is a loop with identity 0, while R^* , the set of non-zero elements of R , is a loop under the operation (\cdot) with identity 1; these are the additive and multiplicative loops of (R, F) , respectively. In general, if A is a set with the operation $(*)$ defined on it, we speak of the system $(A, *)$; thus the additive and multiplicative loops are $(R, +)$ and (R^*, \cdot) , respectively. Finally, a planar ternary ring for which $F(a, b, c) = ab + c$ for all $a, b, c \in R$, is said to be *linear*.

We will need the following theorem in the next section.

THEOREM 2.1. *If (R, F) is a planar ternary ring and S is a finite subset of R such that (S, F) is a ternary ring, then S consists of the zero alone or (S, F) is a planar ternary ring.*

Proof. See [3].

3. Homomorphisms. Let π and π_1 be projective planes and let ϕ be a mapping of the points of π onto the points of π_1 and of the lines of π onto the lines of π_1 ; if ϕ preserves incidence, then ϕ is a *homomorphism* of π onto π_1 . If ϕ is a one-to-one mapping, then it is an *isomorphism*, while if $\pi = \pi_1$ and ϕ is an isomorphism, then it is a *collineation*. Now suppose that ϕ is a homomorphism of the projective plane π onto the projective plane π_1 , and let (R_1, F_1) be a planar ternary ring for π_1 , where $O_1 = (0, 0)$, $Y_1 = (\infty)$, $X_1 = (0)$, $U_1 = (1, 1)$, all in π_1 . Letting O, Y, X, U be points in π which map onto O_1, Y_1, X_1, U_1 , respectively, it is clear that we can choose a planar ternary ring (R, F) for π such that $O = (0, 0)$, $Y = (\infty)$, $X = (0)$, $U = (1, 1)$. In general, in what follows, we shall use the symbols 0, 1, ∞ indiscriminately for both π and π_1 , if it is clear from the context which plane is meant.

From the nature of the two planar ternary rings, any point $(0, y)$ in π is mapped onto a point $(0, y_1)$ or onto (∞) , by ϕ ; similarly, $(x, 0)$ goes onto $(x_1, 0)$ or (0) , and (m) goes onto (m_1) or onto (∞) . Let J be the subset of R consisting of all elements y such that $(0, y)\phi = (\infty)$, and let Z be the subset consisting of all elements y such that $(0, y)\phi = (0, 0)$.

LEMMA 3.1.

- (i) $(x, 0)\phi = (0)$ if and only if $x \in J$.
- (ii) $(x, 0)\phi = (0, 0)$ if and only if $x \in Z$.
- (iii) $(m)\phi = (\infty)$ if and only if $m \in J$.
- (iv) $(m)\phi = (0)$ if and only if $m \in Z$.
- (v) $(x, y)\phi$ is on L_∞ if and only if at least one of x, y , is in J .
- (vi) $(x, y)\phi = (0, 0)$ if and only if both x and y are in Z .

Proof. These statements are all easy to prove; for instance :

(i) Suppose $(x, 0)\phi = (0)$. The points $(x, 0)$, $(0, x)$, (1) are collinear, so the points $(x, 0)\phi = (0)$, $(1)\phi = (1)$, $(0, x)\phi$ are collinear in π_1 . Hence $(0, x)\phi$ must be on L_∞ , so $(0, x)\phi = (\infty)$ and $x \in J$. The converse is similar.

(iii) Suppose $(m)\phi = (\infty)$. The points $(1, 0)$, $(0, m)$, (m) are collinear in π , so the points $(1, 0)$, $(0, m)\phi$, (∞) are collinear in π_1 ; thus $m \in J$. Again, the converse is straightforward.

LEMMA 3.2. *If $x \in R$, $x \neq 0$, then $x \in Z$ ($x \in J$) if and only if one of the following equivalent conditions holds, where b is any element not in Z or J :*

- (i) $xy = b$, for some $y \in J$ ($y \in Z$); (ii) $yx = b$, for some $y \in J$ ($y \in Z$).

Proof. Suppose $(x)\phi = (0)$. Then let $xy = b$, where b is not in J or Z ; this can be done if $x \neq 0$. Since (x) , $(y, 0)$, $(0, b)$ are collinear, so are (0) , $(y, 0)\phi$, $(0, b)\phi \neq (0, 0)$. Hence $(y, 0)\phi = (0)$, so $y \in J$ from Lemma 3.1 (i). The rest of the proof is quite similar.

Now we define R_0 to be the subset of R consisting of all elements not in J , and M to be the subset of R_0 consisting of all elements not in Z .

LEMMA 3.3. (R_0 , F) and (Z , F) are ternary rings.

Proof. If x , y , z are not in J , then let $k = F(x, y, z)$; this means that (x) , (y) , (z) and $(0, k)$ are collinear. But if $k \in J$, then $(0, k)\phi = (\infty)$, so $(y, z)\phi$ must be on L_∞ ; for certainly $(x)\phi \neq (\infty)$, and so the line through $(0, k)\phi$ and $(x)\phi$ is L_∞ . This contradicts Lemma 3.1 (v), so we must have $k \in R_0$; thus (R_0, F) is a ternary ring.

If $x, y, z \in Z$, then let $k = F(x, y, z)$, so that (x) , (y) , (z) , $(0, k)$ are collinear. Thus $(x)\phi = (0)$, $(y, z)\phi = (0, 0)$, $(0, k)\phi$ are collinear, and so $(0, k)\phi = (0, 0)$ and $k \in Z$. This proves the lemma.

Now suppose $(x, y)\phi = (x_1, y_1)$, where $x, y \in R_0$. It is easy to see that this implies $(x, 0)\phi = (x_1, 0)$, $(0, y)\phi = (0, y_1)$, $(0, x)\phi = (0, x_1)$, $(x)\phi = (x_1)$. Hence we can define ϕ as a mapping of R_0 onto R_1 so that $(x, y)\phi = (x\phi, y\phi)$, $(m)\phi = (m\phi)$, for all $m, x, y \in R_0$.

THEOREM 3.1. If $x, y, z \in R_0$, then $[F(x, y, z)]\phi = F_1(x\phi, y\phi, z\phi)$.

Proof. Let $x, y, z \in R_0$, and let $k = F(x, y, z)$; then (x) , (y) , (z) , $(0, k)$ are collinear. So $(x\phi)$, $(y\phi)$, $(z\phi)$, $(0, k\phi)$ are collinear, and $F_1(x\phi, y\phi, z\phi) = k\phi = [F(x, y, z)]\phi$.

We have now shown that a homomorphism of a projective plane implies a homomorphism of a more classical type: an algebraic homomorphism. A good deal more about the structure of R_0 , Z , and M can be said.

THEOREM 3.2. The addition in (R_0, F) is a loop and $(Z, +)$ is a normal subloop of $(R_0, +)$. If $x, y \in R_0$, then $x\phi = y\phi$ if and only if $x = z + y$, where $z \in Z$.

Proof. If we show that $(R_0, +)$ is a loop, then since $(Z, +)$ is clearly the kernel of the homomorphism of $(R_0, +)$ onto $(R_1, +)$, from Theorem 3.1, $(Z, +)$ will be a normal subloop of $(R_0, +)$ (for a discussion of the theory of homomorphisms of loops, see Bruck [1]). The rest of the theorem will follow immediately.

Suppose $x + y = k$, where $x, k \in R_0$. Then (1) , (x, y) , $(0, k)$ are collinear and it is straightforward to show that y is not in J . Similarly, if $x + y = k$, where $y, k \in R_0$, then $x \in R_0$. So $(R_0, +)$ is a subloop of the loop $(R, +)$, and the theorem is proved.

THEOREM 3.3. (M, \cdot) is a loop, and if $I = Z + 1$, then (I, \cdot) is a normal subloop of (M, \cdot) .

Proof. Since $1 \in M$, we need only show that the solutions x, y of $ax = ya = b$, for $a, b \in M$, are also in M . Consider $ax = b$; if $x \in J$, then from Lemma 3.2 we must have $a \in Z$, while if $x \in Z$, then $a \in J$. These are both contradictory, so $x \in M$. Similarly, $y \in M$. The rest of the theorem follows from the fact that I is the kernel of the homomorphism ϕ of (M, \cdot) onto (R_1^*, \cdot) .

It is now easy to see that for all $u, x, y \in R_0$ and all $z \in Z$, $[F(z, x, y)]\phi = [F(u, z, y)]\phi = y\phi$, and $[F(x, y, z)]\phi = (xy)\phi$. In particular, $xZ \subseteq Z$ and $Zx \subseteq Z$ for all $x \in R_0$; furthermore, if $x, y \in R_0$, $xy \in Z$, then at least one of x, y is in Z . So Z is analogous to a prime ideal for the system (R_0, F) , which is itself analogous to a ring. These properties can be compared to the results of Klingenberg [4], in which specialization of the planar ternary ring (R, F) results in (R_0, F) becoming a local ring without zero divisors, and Z a maximal (right or left) ideal. Without passing to this case, we can draw some very simple conclusions about the ternary rings under consideration. Firstly, algebraic properties of (R, F) are going to be reflected in similar algebraic properties in (R_1, F_1) ; thus if (R, F) is linear, left or right distributive, has associative addition or multiplication, etc., these same properties will be possessed by (R_1, F_1) . We can also use Theorem 2.1.

THEOREM 3.4. *The ternary ring (Z, F) is never planar, and hence Z is infinite or $Z = 0$.*

Proof. If (Z, F) is planar, then it possesses a multiplicative identity, which must be the element 1, since Z is contained in a planar ternary ring. But certainly 1 is not in Z , whence the theorem follows from Theorem 2.1.

COROLLARY 1. *If π is a projective plane, ϕ a homomorphism of π onto the projective plane π_1 , and if ϕ is not an isomorphism, then the set of points on a line L of π which map onto a single point on $L\phi$ in π_1 is infinite in number.*

COROLLARY 2. *A finite projective plane possesses no homomorphisms other than isomorphisms.¹*

A “converse” to much of the preceding material can be carried through; i.e., if S is a subset of the planar ternary ring (R, F) , under what circumstances will S play the role of R_0 , for some homomorphism ϕ ? The answer is easy to find, in the sense that necessary and sufficient conditions can be given; unfortunately, both the answer and its derivation are complicated, do not look very suggestive, and add nothing. Thus we omit them; in the somewhat special case considered by Klingenberg, this part of the work is of considerable interest in its own right, and it would be gratifying if investigations in this direction could be made more fruitful.

4. Configuration theorems and collineations. We will now examine briefly some conditions under which collineations are “preserved” under homomorphism, and point out some further problems in this direction. If π is a projective plane and ϕ a homomorphism of π onto a projective plane π_1 , then it is clear that a collineation θ of π will be preserved if θ moves the equivalence classes of ϕ around as entities. With obvious exceptions, we

¹ Professor Reinhold Baer has pointed out to the author that a purely “geometric” proof of this result is also possible.

shall see that the “central collineations” always have this property. A central collineation is one which fixes all the points on some line (the *axis*) and thus necessarily (see [6]) fixes all the lines through some point (the *center*, or *vertex*); if the central collineation is not the identity, then no points (lines) are fixed excepting the center and the points on the axis (the axis and the lines through the center). A more detailed discussion of this and similar topics can be found in [6].

LEMMA 4.1. *Let π be a projective plane and (R, F) a planar ternary ring for π . The mapping*

$$\begin{array}{ll} (x, y) \rightarrow (x, y + a) & [m, k] \rightarrow [m, k + a] \\ (m) \rightarrow (m) & [\infty, (k, 0)] \rightarrow [\infty, (k, 0)] \\ (\infty) \rightarrow (\infty) & L_\infty \rightarrow L_\infty \end{array}$$

for fixed $a \in R$, is a collineation of π if and only if $F(m, x, y + a) = F(m, x, y) + a$ for all $m, x, y \in R$.

Proof. The point (x, y) is on $[m, k]$ if and only if $F(m, x, y) = k$, while $(x, y + a)$ is on $[m, k + a]$ if and only if $F(m, x, y + a) = k + a = F(m, x, y) + a$. The rest of the proof is trivial.

LEMMA 4.2. *Let π be a projective plane and (R, F) a planar ternary ring for π . The mapping*

$$\begin{array}{ll} (x, y) \rightarrow (ax, y) & [m, k] \rightarrow [ma', k] \\ (m) \rightarrow (ma') & [\infty, (k, 0)] \rightarrow [\infty, (ak, 0)] \\ (\infty) \rightarrow (\infty) & L_\infty \rightarrow L_\infty \end{array}$$

for fixed $a \in R$, where $a'a = 1$, is a collineation of π if and only if $F(m, x, y) = F(ma', ax, y)$ for all $m, x, y \in R$.

Proof. Completely similar to the proof of Lemma 4.1.

LEMMA 4.3. *If θ is a central collineation of π with axis K and center Q , then: (i) if Q is on K , θ can be represented in the form of Lemma 4.1, using any planar ternary ring with $Q = Y$, $K = L$; (ii) if Q is not on K , θ can be represented in the form of Lemma 4.2, using any planar ternary ring with $Q = X$, $K = OY$.*

Proof. The proof is easy; or see [6].

THEOREM 4.1. *Suppose π and π_1 are projective planes and ϕ is a homomorphism of π onto π_1 . Let θ be a central collineation of π with axis K and center Q , where Q is on K . If there is a point P_1 of π_1 such that both $P_1\phi$ and $(P_1\theta)\phi$ are not on $K\phi$, then the mapping $\theta_1: P\phi \rightarrow (P\theta)\phi$, $L\phi \rightarrow (L\theta)\phi$, for points $P\phi$ of π_1 and lines $L\phi$ of π_1 , is a central collineation of π_1 with center $Q\phi$, axis $K\phi$.*

Proof. We can choose the planar ternary rings (R, F) and (R_1, F_1) of §3 so that $K = L_\infty$, $Q = Y$, $P_1 = O = (0, 0)$. Then $P_1\theta = (0, a)$ for some $a \in R$, where a is not in J . Lemmas 4.1 and 4.3 assure us that $(x, y)\theta = (x, y + a)$, $[m, k]\theta = [m, k + a]$, etc., all $m, x, y \in R$; furthermore, $F(m, x, y) + a = F(m, x, y + a)$. Hence, letting $a_1 = a\phi$, $F_1(m, x, y) + a_1 = F_1(m, x, y + a_1)$, for all $m, x, y \in R_1$. Hence, from Lemma 4.1, the theorem is proven.

THEOREM 4.2. Suppose π and π_1 are projective planes and ϕ is a homomorphism of π onto π_1 . Let θ be a central collineation of π with axis K and center Q , where Q is not on K , and also $Q\phi$ is not on $K\phi$. If there is a point P_1 of π such that $P_1\phi$ and $(P_1\theta)\phi$ are both not on $K\phi$, and neither equals $Q\phi$, then the mapping $\theta_1: P\phi \rightarrow (P\theta)\phi$, $L\phi \rightarrow (L\theta)\phi$, for points $P\phi$ of π_1 and lines $L\phi$ of π_1 , is a central collineation of π_1 with axis $K\phi$ and center $Q\phi$.

Proof. Completely analogous to the proof of Theorem 4.1.

THEOREM 4.3. Suppose π and π_1 are projective planes and ϕ is a homomorphism of π onto π_1 . Let θ be a central collineation of π with axis K and center Q , where Q is not on K but $Q\phi$ is on $K\phi$. If there is a point P_1 of π such that $P_1\phi$ and $(P_1\theta)\phi$ are both not on $K\phi$, then the mapping $\theta_1: P\phi \rightarrow (P\theta)\phi$, $L\phi \rightarrow (L\theta)\phi$, for points $P\phi$ of π_1 and lines $L\phi$ of π_1 , is a central collineation of π_1 with axis $K\phi$ and center $Q\phi$.

Proof. We choose planar ternary rings (R, F) and (R_1, F_1) such that $K = L_\infty$, $P_1 = (0, 0)$, $Q = (0, b)$, $P_1\theta = (0, a)$. Then $b \in J$ since $Q\phi = (\infty)$, while $a \in R_0$.

Let $y \in R_0$ and let $(0, y)\theta = (0, z)$. Furthermore, let $my = b$; since $[m, b]$ is a line fixed by θ and $(y, 0)$ is on $[m, b]$, $(y, 0)\theta$ is also on $[m, b]$. But $(0), (y, 0), (0, 0)$ are collinear, so $(0), (y, 0)\theta, (0, a)$ are collinear; hence $(y, 0)\theta = (u, a)$, where $F(m, u, a) = b$.

The points $(1), (y, 0), (0, y)$ are collinear, so $(1), (u, a), (0, z)$ are collinear. Hence $u + a = z$. Now we transfer attention to π_1 . Since $(y, 0)\phi = (y', 0)$ is not on L_∞ , but $(0, b)\phi = (\infty)$, the line $[m, b]\phi$ must be $[\infty, (y', 0)]$. So $(u, a)\phi = (u\phi, a\phi)$ is on $[\infty, (y', 0)]$, whence $u\phi = y' = y\phi$. Thus, from the equation $u + a = z$, we have $y\phi + a\phi = u\phi + a\phi = z\phi$. So $(0, y)\theta = (0, z)$ implies $z\phi = y\phi + a\phi$.

Now let (x, y) be a point of π , where $x, y \in R_0$, and let $F(n, x, y) = b$. Then $(x, y)\theta = (u, v)$ implies $F(n, u, v) = b$. Furthermore, if $F(m, x, y) = k$, then $F(m, u, v) = t$, where $t\phi = k\phi + a\phi$, since $(0, k)\theta = (0, t)$. In π_1 , $[n, b]\phi = [\infty, (x\phi, 0)]$, so $u\phi = x\phi$. Thus if $F_1(m\phi, x\phi, y\phi) = k\phi$, then $F_1(m\phi, x\phi, v\phi) = t\phi = k\phi + a\phi = F_1(m\phi, x\phi, y\phi) + a\phi$. Letting $m = 0$, this gives $v\phi = y\phi + a\phi$. So, for any $m \in R_0$, $F_1(m\phi, x\phi, y\phi + a\phi) = F_1(m\phi, x\phi, y\phi) + a\phi$. Now from Lemma 4.1, we are done.

The preceding three theorems can also be proved as configuration theorems, since the existence of central collineations is equivalent to the existence of

certain instances of Desargues' Theorem [6]. It is well known that if every "possible" central collineation with axis K , center Q , exists, then (R, F) can be chosen to be linear with associative addition or multiplication, according as Q is on K or not (see [6]; this can also be proved directly from Lemmas 4.1 and 4.2). In such a case the additive or multiplicative group, as the case may be, is isomorphic to the group of central collineations under discussion. Thus Theorem 3.1 implies that if π has a "complete" central collineation group with axis K and center Q , and either Q is on K or Q is not on K and also $Q\phi$ is not on $K\phi$, then $\pi\phi$ also has a "complete" central collineation group. But Theorem 4.3 asserts that if Q is not on K while $Q\phi$ is on $K\phi$, then $\pi\phi$ also has a "complete" central collineation group; note however, that since the center has been moved onto the axis, π will have a special planar ternary ring with associative multiplication, while $\pi\phi$ will have one with associative addition. Furthermore, it is not true in general that if G is a collineation group of π which gives rise to a collineation group H in $\pi\phi$, then H is a homomorphic image of G ; in the central collineation cases, H is a homomorphic image of a subgroup of G .

It would be of considerable interest to know if any other types of collineations are preserved under homomorphism. The free planes [2; 6] for instance, possess no non-identity central collineations at all. But every finite projective plane is a homomorphic image of a free plane of finite rank and a central problem in the study of finite planes is the question of the existence of non-identity collineations. The collineation groups of free planes are likely to be interesting objects in their own right, and might repay some close study.

BIBLIOGRAPHY

1. R. H. Bruck, *Contributions to the theory of loops*, Trans. Amer. Math. Soc. vol. 60 (1946) pp. 245–354.
2. Marshall Hall, Jr., *Projective planes*, Trans. Amer. Math. Soc. vol. 54 (1943) pp. 229–277.
3. D. R. Hughes, *Planar division neo-rings*, Trans. Amer. Math. Soc. vol. 80 (1955) pp. 502–527.
4. Wilhelm Klingenberg, *Projektive Geometrien mit Homomorphismus*, Math. Ann. vol. 132 (1956) pp. 180–200.
5. L. I. Koperkina, *Decompositions of projective planes*, Izv. Akad. Nauk. SSSR Ser. Math. vol. 9 (1945) pp. 495–526.
6. Günter Pickert, *Projektive Ebenen*, Berlin, 1955.

THE OHIO STATE UNIVERSITY,
COLUMBUS, OHIO

FINITE DIVISION ALGEBRAS AND FINITE PLANES¹

BY

A. A. ALBERT

1. Introduction. We shall discuss here some aspects of the theory of finite nonassociative division algebras and the finite projective planes which they determine. We shall begin our discussion with a more detailed exposition of the theory of isotopes for such algebras than any given elsewhere, and shall show how that theory leads to what appears to be a reasonable procedure for determining all division algebras with either 16 or 32 elements. Every division algebra \mathfrak{D} determines a projective plane $\mathfrak{M}(\mathfrak{D})$ which is Desarguesian (in the finite case) if and only if \mathfrak{D} is the unique finite field determined by its order. We shall show that *the planes determined by two division algebras \mathfrak{D} and \mathfrak{D}' are isomorphic if and only if \mathfrak{D} and \mathfrak{D}' are isotopic over their prime subfield*. This result may then be used to express the collineations of $\mathfrak{M}(\mathfrak{D})$ in terms of the known elementary collineations (translations and shears) and the *autotopes* of \mathfrak{D} . We shall conclude with the definition of a class of *central*² division algebras of dimension n over the finite field \mathfrak{F}_q of q elements for every $n > 2$ and $q > 2$.

2. Elementary properties and isotopy. Consider an algebra \mathfrak{D} , of finite dimension n , over an arbitrary field \mathfrak{F} . Then \mathfrak{D} consists of an n -dimensional vector space \mathfrak{D} over \mathfrak{F} and a product operation xy on $(\mathfrak{D}, \mathfrak{D})$ to \mathfrak{D} which is a bilinear function over \mathfrak{F} of x and y . Thus the right multiplications of \mathfrak{D} , defined by

$$(1) \quad x \rightarrow xy = xR_y$$

for every y of \mathfrak{D} , are linear transformations R_y of the vector space \mathfrak{D} . The set $R(\mathfrak{D})$, of all the R_y , is a vector space over \mathfrak{F} of linear transformations, and the mapping

$$(2) \quad y \rightarrow R_y$$

is a linear mapping over \mathfrak{F} of \mathfrak{D} onto $R(\mathfrak{D})$.

We shall call \mathfrak{D} an algebra *without divisors of zero* if all products xy , with $x \neq 0$ and $y \neq 0$ in \mathfrak{D} , are not zero. Thus \mathfrak{D} is an algebra without divisors

¹ This paper was sponsored, in part, by NSF Grant G-4792.

² This definition is the one given in a paper entitled *Finite noncommutative division algebras*, the Proceedings of the American Mathematical Society vol. 9 (1958) pp. 928-932. In that paper we showed that the algebras, called twisted fields, are not commutative whenever the defining parameter c is not -1 . In the present paper we shall show that the twisted field \mathfrak{D}_c defined by a finite field \mathfrak{K} over \mathfrak{F}_q is always central over \mathfrak{F}_q .

of zero if and only if $R(\mathfrak{D})$ has dimension n over \mathfrak{F} , and all nonzero elements of $R(\mathfrak{D})$ are nonsingular. The equation $xa = b$ then has the unique solution $x = bR_a^{-1}$.

In a similar fashion we may define the left multiplications L_x by $y \rightarrow xy = yL_x$, the space $L(\mathfrak{D})$ of all L_x , and the mapping $x \rightarrow L_x$ of \mathfrak{D} onto $L(\mathfrak{D})$. It is now easy to see that \mathfrak{D} has no divisors of zero if and only if the mathematical system \mathfrak{D}^* , consisting of the nonzero elements of \mathfrak{D} and the product operation xy , is a quasigroup.

An algebra \mathfrak{D} is called a *division algebra* if \mathfrak{D}^* is a loop. Thus an algebra \mathfrak{D} is a division algebra if and only if \mathfrak{D} is without divisors of zero, and has a unity element e . This element e has the property that R_e is the identity transformation I , and $L_e = I$, a property equivalent to $eR_y = y$ for every y of \mathfrak{D} .

Any two algebras \mathfrak{D} and \mathfrak{D}_0 over \mathfrak{F} , which have the same dimension over \mathfrak{F} , may be regarded as being the same vector space, that is, as having the same elements. Then \mathfrak{D} and \mathfrak{D}_0 are said to be *isotopic* over \mathfrak{F} if there exist three nonsingular linear transformations P, Q, S over \mathfrak{F} , of the vector space \mathfrak{D} , such that the product (x, y) of \mathfrak{D}_0 is expressible in terms of the product xy of \mathfrak{D} by

$$(3) \quad (x, y)S = (xP)(yQ).$$

Every algebra without divisors of zero is isotopic to a division algebra. Indeed, let g and h be any two nonzero elements of \mathfrak{D} so that R_h and L_g are nonsingular. Then we may determine nonsingular transformations P and Q by

$$(4) \quad P = R_h^{-1}, \quad Q = L_g^{-1}.$$

Define an isotope \mathfrak{D}_0 of \mathfrak{D} by

$$(5) \quad (x, y) = (xP)(yQ),$$

for this P and Q , and let

$$(6) \quad f = gh = hL_g = gR_h = hQ^{-1} = gP^{-1}.$$

Then $fP = g$, $fQ = h$, and we have

$$(7) \quad (f, y) = g(yL_g^{-1}) = y, \quad (x, f) = (xP)h = xR_h^{-1}R_h = x,$$

so that f is the unity quantity of \mathfrak{D}_0 .

Suppose then that \mathfrak{D} is a division algebra over \mathfrak{F} , and that e is its unity element. Let f be any nonzero element of \mathfrak{D} and write

$$(8) \quad (x, fy) = xy.$$

The product (x, z) determined in this way defines an isotope \mathfrak{D}_0 of \mathfrak{D} , since (8) is actually equivalent to

$$(9) \quad (x, y) = x(yL_f^{-1}).$$

This isotope has f as unity element since

$$(x, f) = x(fL_f^{-1}) = x[(eL_f)L_f^{-1}] = xe = x \quad \text{and} \\ (f, y) = f(yL_f^{-1}) = yL_f^{-1}L_f = y.$$

We state this well known property as follows.

LEMMA 1. *Let \mathfrak{D} be an algebra without divisors of zero and f be any nonzero element of \mathfrak{D} . Then there exists a division algebra \mathfrak{D}_0 isotopic to \mathfrak{D} and with f as its unity element.*

If $P = Q = S$ the formula $(x, y)P = (xP)(yP)$ implies that the mapping P of x onto xP and the product xy onto the product (x, y) is an isomorphism of \mathfrak{D} onto \mathfrak{D}_0 . It implies that if \mathfrak{D} is isotopic to \mathfrak{D}_0 then \mathfrak{D} is also isotopic to a *principal* isotope defined by

$$(10) \quad (x, y) = (xP)(yQ).$$

Moreover we have the following result.

LEMMA 2. *Let \mathfrak{D} and \mathfrak{D}_0 be isotopic division algebras over \mathfrak{F} , so that they may be regarded as being the same vector space over \mathfrak{F} . Then there exists an algebra \mathfrak{D}_{00} isomorphic over \mathfrak{F} to \mathfrak{D}_0 and having the same unity element e as \mathfrak{D} .*

For the result above implies that \mathfrak{D}_0 is isomorphic to an algebra defined by (10). Hence assume that \mathfrak{D}_0 is defined by (10). If f is the unity quantity of \mathfrak{D}_0 we have $(x, f) = x - xPR_{fQ}$ and $P^{-1} = R_{fQ}$. Also $(f, y) = y = yQL_{fP}$ and $Q^{-1} = L_{fP}$. Then $[e(fQ)]P = f(QP) = eP^{-1}P = e$, and $[(fP)e]Q = f(PQ) = eQ^{-1}Q = e$. We now define $S = (PQ)^{-1} = L_{fP}R_{fQ}$ and have $eS = f$. Define an algebra \mathfrak{D}_{00} isomorphic to \mathfrak{D}_0 by the product $x \cdot y = (xS, yS)S^{-1}$ and have $e \cdot y = (eS, yS)S^{-1} = (f, yS)S^{-1} = uSS^{-1} = y$, $x \cdot e = (xS, eS)S = (xS, f)S^{-1} = (xS)S^{-1} = x$. This completes our proof.

Let us now probe more deeply into the nature of a division algebra from the linear transformation viewpoint. We have seen that every division algebra \mathfrak{D} of dimension n over \mathfrak{F} consists of the following mathematical items:

- (a) An n -dimensional vector space \mathfrak{D} over \mathfrak{F} ;
- (b) An n -dimensional vector space $R(\mathfrak{D})$ of linear transformations R over \mathfrak{F} of \mathfrak{D} such that the nonzero elements of $R(\mathfrak{D})$ are nonsingular and the identity transformation I is in $R(\mathfrak{D})$;
- (c) A one-to-one mapping $x \rightarrow R_x$ of \mathfrak{D} onto $R(\mathfrak{D})$ so that there is a unique element e in \mathfrak{D} with $R_e = I$. Then the mapping $x \rightarrow R_x$ must have the property that $eR_x = x$.

We should now observe that, in a sense, only items (a) and (b) are essential. Indeed suppose that \mathfrak{D} is defined by (a), (b), (c). We select any nonzero element f of \mathfrak{D} and propose to define a new mapping

$$(11) \quad x \rightarrow R_x^{(0)},$$

of \mathfrak{D} onto $R(\mathfrak{D})$, with $R_f^{(0)} = I$ and $R_x^{(0)} \neq 0$ if $x \neq 0$. For every linear transformation R in $R(\mathfrak{D})$ we define a corresponding element x by

$$(12) \quad x = fR.$$

This maps $R(\mathfrak{D})$ into \mathfrak{D} . But $fR = fS$ for R and S in the vector space $R(\mathfrak{D})$, if and only if $f(R - S) = 0$, and if $R \neq S$ this is impossible by (6). Hence (12) defines a one-to-one mapping

$$(13) \quad R \rightarrow x = fR,$$

of $R(\mathfrak{D})$ onto \mathfrak{D} , and we can use (13) to write

$$(14) \quad R = R_x^{(0)}, \quad x = fR_x^{(0)}, \quad R_f^{(0)} = I.$$

We may now say that the third element (c), in our collection of the items which define a division algebra \mathfrak{D} , is merely in the nature of the selection of the unity element f . We observe further, that all selections of f yield isotopes of \mathfrak{D} . For, if \mathfrak{D} is defined by the mapping $x \rightarrow R_x$, and we use (12) to define $y = fR_x = fx = xL_f$, then we have

$$(15) \quad R_y^{(0)} = R_x = R_{yQ}, \quad Q = L_f^{-1}.$$

In any actual construction of division algebras we will select a basis

$$(16) \quad e = e_1, e_2, \dots, e_n$$

of a proposed division algebra \mathfrak{D} with unity element e , and represent the elements of $R(\mathfrak{D})$ by n -rowed square matrices. Then we can assume that the elements of the proposed algebra \mathfrak{D} are n -tuples

$$(17) \quad x = (\xi_1, \dots, \xi_n) = \xi_1 e_1 + \dots + \xi_n e_n \quad (\xi_i \text{ in } \mathfrak{F}),$$

and shall normalize all proposed constructions by insisting on the hypothesis that

$$(18) \quad e = (1, 0, \dots, 0)$$

is always our unity element. We now define \mathfrak{D} by a product formula

$$(19) \quad xy = (\xi_1, \dots, \xi_n)S_y,$$

where S_y is an n -rowed square matrix whose elements are linear forms in the coordinates η_j of $y = (\eta_1, \dots, \eta_n)$. The condition that \mathfrak{D} be an algebra without divisors of zero is that the determinant

$$(20) \quad \Delta(y) = |S_y|,$$

a form (homogeneous polynomial) of degree n in η_1, \dots, η_n , be *not a null form*, that is, $\Delta(y) = 0$ only if $y = (0, 0, \dots, 0)$.

The discussion we have given now implies that \mathfrak{D} may be regarded as consisting of the following items:

- (a) The space \mathfrak{B}_n of all n -tuples $a = (\alpha_1, \dots, \alpha_n)$ with coordinates α_i in \mathfrak{F} .

(b) A vector space $S(\mathfrak{D})$ over \mathfrak{F} of n -rowed square matrices S with elements in \mathfrak{F} , and a basis $S_1 = I, S_2, \dots, S_n$ such that, if $y = (\eta_1, \dots, \eta_n) \neq 0$, then $\eta_1S_1 + \dots + \eta_nS_n$ is nonsingular.

(c) The mapping of $y = eS$ onto $S = S_y$ for every S of $R(\mathfrak{D})$.

The definition above can be used if we replace $S(\mathfrak{D})$ by $T(\mathfrak{D}_0) = PS(\mathfrak{D})P^{-1}$ for any fixed nonsingular P . For clearly

$$(21) \quad T = PSP^{-1} = P(\eta_1S_1 + \dots + \eta_nS_n)P^{-1}$$

is nonsingular when $y = (\eta_1, \dots, \eta_n) \neq 0$. We now assume that \mathfrak{D} is defined by $xy = xS_y$ and we define a mapping Q by

$$(22) \quad y \rightarrow z = yQ = ePS_yP^{-1}.$$

Then Q is nonsingular since PS_yP^{-1} is nonsingular if $y \neq 0$ and so $z \neq 0$ if $y \neq 0$. Define a product (x, z) by

$$(23) \quad (x, z) = x(PS_yP^{-1}), \quad yQ = z.$$

This defines an isotope \mathfrak{D}_0 of \mathfrak{D} which has e as unity quantity since $eQ = e$, $PS_eP^{-1} = PIP^{-1} = I$, $(e, z) = ePS_yP^{-1} = Z$. We state this result as follows:

LEMMA 3. *Let \mathfrak{D} be a division algebra over \mathfrak{F} with unity quantity e so that \mathfrak{D} may be regarded as being the set of all n -tuples $a = (\alpha_1, \dots, \alpha_n)$ with α_i in \mathfrak{F} and $e = (1, 0, \dots, 0)$. Let $S(\mathfrak{D})$ be the space of the matrices S_x of the right multiplications of \mathfrak{D} relative to the basis leading to the n -tuples, and P be any nonsingular n -rowed square matrix. Then there exists an isotope \mathfrak{D}_0 of \mathfrak{D} with the same unity element e as \mathfrak{D} and with $S(\mathfrak{D}_0) = PS(\mathfrak{D})P^{-1}$.*

The following result will also be useful in the construction of nonassociative division algebras.

LEMMA 4. *Let the hypotheses of Lemma 3 be satisfied and let T be any nonzero element of $S(\mathfrak{D})$. Then there exists an isotope \mathfrak{D}_0 of \mathfrak{D} with e as unity element, and either $S(\mathfrak{D}_0) = T^{-1}S(\mathfrak{D})$ or $S(\mathfrak{D}_0) = [S(\mathfrak{D})]T^{-1}$, so that T^{-1} is in $S(\mathfrak{D}_0)$.*

For $T^{-1}S(\mathfrak{D})$ is an n -dimensional space of nonsingular linear transformations on \mathfrak{D} and the zero transformation, and it has the essential property that $T^{-1}T = I$ is in the set. The mapping Q defined by

$$(24) \quad y \rightarrow e_1T^{-1}S_y = z = yQ,$$

is one-to-one since $z = (e_1T^{-1})y = 0$ implies that $y = 0$. Now $T = R_b$ for a nonzero b in the division algebra \mathfrak{D} , and

$$(25) \quad bQ = e_1T^{-1}T = e_1.$$

Thus the algebra defined by

$$(26) \quad (d, z) = x(T^{-1}S_y) \quad (y = zQ^{-1})$$

is an isotope \mathfrak{D}_0 of \mathfrak{D} with $S(\mathfrak{D}_0) = T^{-1}S(\mathfrak{D})$ and $b = eQ^{-1}$, $(x, e_1) = xT^{-1}S_b = x$, while $(e_1, z) = e_1T^{-1}S_y = Z$ by (24).

3. Right powers. If b is any element of an algebra \mathfrak{D} we define the right powers of b by the inductive formula

$$b^1 = b, \quad b^k = b^{k-1}b \quad (k \geq 2).$$

When \mathfrak{D} has a unity quantity e we have

$$b^k = eR_b^k \quad (k \geq 1).$$

If \mathfrak{D} is a division algebra, and we use the n -tuple representation of Lemmas 1 and 2, we have

$$(27) \quad b^k = e_1S_b^k.$$

Thus the characteristic function

$$(28) \quad f(\lambda; b) = |\lambda I - S_b| = \lambda^n + \alpha_1\lambda^{n-1} + \cdots + \alpha_n,$$

of the matrix S_b , defines a polynomial such that

$$(29) \quad f(b; b) = b^n + \alpha_1b^{n-1} + \cdots + \alpha_ne_1 = 0.$$

We shall call $f(\lambda; b)$ the *right characteristic function* of b .

There must now exist a polynomial

$$(30) \quad \phi(\lambda; b)$$

called the *right minimum function* of b , which is the polynomial of least degree in $\mathfrak{F}[\lambda]$ such that $\phi(b; b) = e_1\phi(S_b; b) = 0$. As usual we know that $\phi(\lambda; b)$ divides every polynomial $\psi(\lambda)$ such that $\psi(b) = e_1\psi(S_b) = 0$. In particular $\phi(\lambda; b)$ divides the minimum function of S_b .

Suppose now that b is a nonscalar element of \mathfrak{D} , that is, $b \neq \lambda e$ for any λ in \mathfrak{F} . Assume that $g(\lambda)$ is any irreducible factor of the right characteristic function $f(\lambda; b)$ of b . Then $g(\lambda)$ is a factor of the minimum function $\phi(\lambda)$ of the corresponding right multiplication R_b and $\phi(\lambda) = g(\lambda)h(\lambda)$, $\phi(R_b) = 0$. Hence $\phi(\lambda) = g(\lambda)$ and $g(R_b) = 0$ or $g(R_b) \neq 0$, $g(R_b)h(R_b) = 0$, $h(R_b) \neq 0$. Thus $g(R_b)$ must always be singular. Hence there exists a nonzero element f in \mathfrak{D} such that

$$(31) \quad f[g(R_b)] = 0.$$

We now define a new algebra \mathfrak{D}_0 by using the product $(x, y) = x(yL_f^{-1})$ and have (10). Thus \mathfrak{D}_0 is an isotope of \mathfrak{D} with unity element f . Put $c = fb = bL_f$ and see that $(x, c) = x(cL_f^{-1}) = xb = xR_b$, that is, $R_c^{(0)} = R_b$, $g(R_c^{(0)}) = g(R_b)$, $f[g(R_c^{(0)})] = 0$. It follows that the right minimum function of c in \mathfrak{D}_0 divides the irreducible polynomial $g(\lambda)$ and must coincide with it. Thus we have the following result.

LEMMA 5. *Let b be a nonscalar element of a division algebra \mathfrak{D} and $g(\lambda)$ be an irreducible factor of the right characteristic function of b in \mathfrak{D} . Then there exists an isotope \mathfrak{D}_0 of \mathfrak{D} and an element c in \mathfrak{D}_0 whose right minimum function is $g(\lambda)$.*

Note that the polynomial $g(\lambda)$ can never be linear. Indeed if $g(\lambda) = \lambda - \alpha$ then $R_b - \alpha I$ is singular and there exists a nonzero element f in \mathfrak{D} such that $fb = \alpha_1 f = f(\alpha e)$, $g(b - \alpha e) = 0$, $b \neq \alpha e$ by hypothesis. This is impossible in a division algebra.

4. Algebras over finite fields. Let \mathfrak{D} be a division algebra of dimension $n \geq 4$ over a finite field \mathfrak{F} . Then \mathfrak{D} has a basis $e = e_1, e_2, \dots, e_n$, and the vector space $S(\mathfrak{D})$, of the matrices of the right multiplications $R(y) = R_y$, has a corresponding basis $S_1 = I, S_2, \dots, S_n$, where S_i is the matrix of $R(e_i)$. The general element $S(\mathfrak{D})$ is the matrix $S(y) = \eta_1 S_1 + \dots + \eta_n S_n$, and the right characteristic function of y is

$$(32) \quad |\lambda I - S(y)| = \lambda^n + \sigma_1(y)\lambda^{n-1} + \dots + \sigma_n(y),$$

where $\sigma_i(y)$ is a homogeneous polynomial of degree i in η_1, \dots, η_n .

LEMMA 6. *If $n \geq 4$ there exists a nonscalar element b in \mathfrak{D} whose right characteristic function has the form $\lambda^n + \beta_1\lambda^{n-1} + \dots + \beta_n$ with β_i in \mathfrak{F} and $\beta_2 = 0$. If $n > 4$ there exists an element b whose right characteristic function has this form with $\beta_1 = \beta_2 = 0$.*

For $\beta_i = \rho_i(b)$, and $\rho_2(y)$ is a quadratic form in η_1, \dots, η_n . Then we can put $\eta_1 = 0$ and $y^* = (0, \eta_2, \dots, \eta_n)$ to see that $\rho_2(y^*)$ is a quadratic form in η_2, \dots, η_n . If $n > 3$ we know that $\rho_2(y^*)$ is a quadratic form in three variables, and every such form over a finite field is a null form. Hence, we can always select nonzero elements $\alpha_2, \dots, \alpha_n$ in \mathfrak{F} , such that $b = (0, \alpha_2, \dots, \alpha_n) \neq 0$, b is nonscalar, $\rho_2(b) = 0$. If $n > 4$ the polynomial $\rho_1(y)$ is a linear form in η_1, \dots, η_n . We can solve the equation $\rho_1(y) = 0$ for one η_i in terms of the remaining η_i and have a corresponding set of vectors y^* whose coordinates are linear forms in $n - 1$ indeterminates. Then $\rho_2(y^*)$ is a quadratic form in these $n - 1$ indeterminates. Since $n > 4$ there are at least four indeterminates in $\rho_2(y^*)$, and we can take $\eta_1 = 0$ and have a corresponding element y^{**} with $\rho_2(y^{**}) = 0$, $\rho_2(y^{**})$ a quadratic form in at least three indeterminates. This is a null form, and we can determine a corresponding nonscalar element b with $\rho_1(b) = \rho_2(b) = 0$, as desired.

Let us now consider the case where \mathfrak{F} is the field of two elements and $n = 4, 5$. If $n = 4$ we select an element b as in Lemma 4 with right characteristic function

$$(33) \quad f(\lambda) = \lambda^4 + \beta_2\lambda^3 + \beta_3\lambda + \beta_4.$$

If $f(\lambda)$ is reducible it cannot have a linear factor and so must be the square of the unique irreducible quadratic polynomial $\lambda^2 + \lambda + 1$, that is,

$f(\lambda) = \lambda^4 + \lambda^2 + 1$, which is clearly false in (33). Hence $f(\lambda)$ must be irreducible. Then $1 + \beta_1 + \beta_3 + \beta_4 \neq 0$, $1 + \beta_1 + \beta_3 + \beta_4 = 1$, $\beta_4 = \beta_1 + \beta_3$. But S_b is nonsingular and so $\beta_4 = 1$, $\beta_1 + \beta_3 = 0$. Thus we have the possible values

$$(34) \quad \beta_1 = 1, \quad \beta_3 = 0, \quad f(\lambda) = \lambda^4 + \lambda^3 + 1,$$

and

$$(35) \quad \beta_1 = 0, \quad \beta_3 = 1, \quad f(\lambda) = \lambda^4 + \lambda + 1.$$

In the former case $[S(b)]^{-1}$ is a root of the irreducible polynomial $\lambda^4 + \lambda + 1$. The polynomial $\lambda^4 + \lambda + 1$ is then the right minimum function of an element b of an isotope \mathfrak{D}_0 of \mathfrak{D} and we have derived the following result.

THEOREM 1. *Let \mathfrak{D} be a division algebra of dimension four over the field \mathfrak{F}_2 of two elements. Then there exists an isotope \mathfrak{D}_0 of \mathfrak{D} containing an element b such that*

$$(36) \quad e = e_1, \quad b = e_2, \quad b^2 = e_3, \quad b^3 = b^2b = e_4$$

form a basis over \mathfrak{F}_2 of \mathfrak{D} , where

$$(37) \quad b^4 = b^3b = e + b.$$

Observe that the result above implies that

$$(38) \quad L_b = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}, \quad L_{bb} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \\ \delta_1 & \delta_2 & \delta_3 & \delta_4 \end{pmatrix},$$

$$L_{(bb)b} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \end{pmatrix},$$

where the 24 parameters $\alpha_i, \beta_i, \gamma_i, \delta_i, \rho_i, \sigma_i$ must be selected in \mathfrak{F}_2 so that the seven matrices L_b , L_{bb} , $L_{(bb)b}$, $L_b + L_{bb}$, $L_b + L_{(bb)b}$, $L_{bb} + L_{(bb)b}$, $L_b + L_{bb} + L_{(bb)b}$ have determinant unity and the seven sums of these matrices and I also have determinant unity. Thus the solutions of these 14 equations in 24 parameters yield a set of division algebras with 16 elements to which all algebras with 16 elements are isotopic.

We next consider the case $n = 5$. Then the right characteristic equation of an element b of \mathfrak{D} can be taken to have the form $\lambda^5 + \beta_3\lambda^2 + \beta_4\lambda + \beta_5$, where $1 + \beta_3 + \beta_4 + \beta_5 = 1$, $\beta_3 + \beta_4 + \beta_5 = 0$, $\beta_5 = 1$, $\beta_3 + \beta_4 = 1$. Then we have the solutions with $\beta_3 = 1$, $\beta_4 = 0$, or $\beta_3 = 0$ and $\beta_4 = 1$, so that the two possibilities are $f(\lambda) = \lambda^5 + \lambda^2 + 1$ and $f(\lambda) = \lambda^5 + \lambda + 1$.

The only irreducible cubics over \mathfrak{F}_2 are $\lambda^3 + \lambda + 1$ and $\lambda^3 + \lambda^2 + 1$, and the product $(\lambda^3 + \lambda + 1)(\lambda^2 + \lambda + 1) = \lambda^5 + \lambda^4 + \lambda + 1 \neq f(\lambda)$. However $\lambda^5 + \lambda + 1 = (\lambda^3 + \lambda^2 + 1)(\lambda^2 + \lambda + 1)$ and we have derived the following result.

THEOREM 2. *Let \mathfrak{D} be a division algebra of dimension 5 over the field \mathfrak{F}_2 of two elements. Then \mathfrak{D} either has a basis*

$$(39) \quad e_1 = e, \quad b = e_2, \quad b^2 = e_3, \quad b^2b = e_4, \quad (b^2b)b = e_5$$

where $[(b^2b)b]b = b^2 + e$, or there exists an isotope of \mathfrak{D} containing an element b such that $b^2 = b + e$.

Observe that, in the former case \mathfrak{D} is generated by b . In the latter case the characteristic function of R_b is the product of the invariant factors of $I - R_b$ and one of these is $\lambda^2 + \lambda + 1$. We may then use Lemma 4 to replace R_b by a similar matrix whose minimum function must be $f(\lambda) = \lambda^5 + \lambda + 1$. Then \mathfrak{D} has a basis $e = e_1, b, b^2, b^2b, (b^2b)b$ with $b^5 = [(b^2b)b]b = b + e$. The results on left multiplications may then be derived as in the case $n = 4$, and a procedure for determining \mathfrak{D} can then be obtained.

5. Left vector spaces. Let \mathfrak{D} be an n -dimensional division algebra over a field \mathfrak{F} , and \mathfrak{D}_r be the set of all r -tuples

$$(40) \quad A = (a_1, \dots, a_r),$$

with coordinates a_i in \mathfrak{D} . Then \mathfrak{D}_r may be regarded as being a *left vector space* over \mathfrak{D} , that is, every left linear combination $d_1A_1 + d_tA_t$, of elements A_i in \mathfrak{D}_r with coefficients d_i in \mathfrak{D} , is in \mathfrak{D}_r .

A subset \mathfrak{N} of \mathfrak{D}_r will be called a *left vector space over \mathfrak{D}* if \mathfrak{N} is a subspace over \mathfrak{F} of \mathfrak{D}_r , and

$$(41) \quad dA = (da_1, \dots, da_r)$$

is in \mathfrak{N} for every d of \mathfrak{D} and every A of \mathfrak{D}_r . Then \mathfrak{N} is a left *subspace* over \mathfrak{D} of \mathfrak{D}_r .

A subset \mathfrak{N} of \mathfrak{D}_r is said to be *spanned* over \mathfrak{D} by t of its elements U_1, \dots, U_t , if every element A of \mathfrak{N} is expressible in the form

$$(42) \quad A = d_1U_1 + \dots + d_tU_t$$

for coefficients d_i in \mathfrak{D} . A finitely spanned subset \mathfrak{N} of \mathfrak{D}_r will be said to be *closed* over \mathfrak{D} , relative to the spanning subset U_1, \dots, U_t which spans \mathfrak{N} , if every element A of (42) is in \mathfrak{N} . A finitely spanned subset \mathfrak{N} of \mathfrak{D}_r , closed over \mathfrak{D} , is a subspace over \mathfrak{F} of \mathfrak{D}_r . However \mathfrak{N} need not be a left subspace over \mathfrak{D} of \mathfrak{D}_r .

Indeed, let $r = 2$ and \mathfrak{N} be spanned over \mathfrak{D} by $U = (a, 1)$ for a in \mathfrak{D} . Then \mathfrak{N} will be closed over \mathfrak{D} if \mathfrak{N} consists of all $dU = (da, d)$ for d in \mathfrak{D} .

If c is in \mathfrak{D} the element $c(dU) = (c \cdot da, cd)$ is in \mathfrak{N} if and only if $(c \cdot da, cd) = (ba, b)$ for b in \mathfrak{D} . But then $b = cd$, $ba = (cd)a = c(da)$ which is true for all a, c, d of \mathfrak{D} if and only if \mathfrak{D} is associative.

We observe that, if E_i is the r -tuple with the unity element 1 of \mathfrak{D} as its i th coordinate and zeros elsewhere, then E_1, \dots, E_r span \mathfrak{D}_r . Moreover, any subset of \mathfrak{D}_r , spanned by some of the E_j and closed over \mathfrak{D} , is a left vector subspace over \mathfrak{D} of \mathfrak{D}_r .

6. Finite projective planes. A *finite projective plane* \mathfrak{M} of order μ is a set of $\nu = \mu^2 + \mu + 1$ elements A_i , called *points*, such that \mathfrak{M} contains exactly ν subsets L_j , called *lines*, where each line consists of exactly $\mu + 1$ points and the intersection of any two distinct lines of \mathfrak{M} is a single point of \mathfrak{M} . Two projective planes \mathfrak{M} and \mathfrak{M}' are said to be *isomorphic* if they have the same order and there exists a nonsingular mapping

$$\sigma : A_i \rightarrow A_i^{\circ}$$

of \mathfrak{M} onto \mathfrak{M}' which maps the ν lines L_i of \mathfrak{M} onto the ν lines L_i° of \mathfrak{M}' . A *collineation* of \mathfrak{M} is an isomorphism σ of \mathfrak{M} onto \mathfrak{M} . The set of all collineations is a group called the *collineation group* of \mathfrak{M} .

Every finite projective plane can be *coordinatized* by what is called a *planar ternary ring*. Each such ring is a set \mathfrak{D} , and an operation $\Phi(x, m, b)$ on $(\mathfrak{D}, \mathfrak{D}, \mathfrak{D})$ to \mathfrak{D} , such that

(i) There is a unique element 1 in \mathfrak{D} such that the system consisting of the set \mathfrak{D} and the binary operation $a + b = \Phi(1, m, b)$ is a *loop* whose identity element $0 \neq 1$.

(ii) The ternary operation has the property $\Phi(0, m, b) = \Phi(x, 0, b) = b$ for every m, x , and b of \mathfrak{D} .

(iii) The system \mathfrak{D}^* , consisting of the nonzero elements of \mathfrak{D} and the binary operation $\Phi(x, m, 0)$, is a *loop* with identity element 1.

(iv) The $\mu - 2$ systems $\mathfrak{D}(m)$ consisting of \mathfrak{D} and the binary operations $\Phi(x, m, b)$ for fixed $m \neq 0, 1$ are *quasigroups*.

(v) If $m_1 \neq m_2$ the equation $\Phi(x, m_1, a) = \Phi(x, m_2, b)$ has a unique solution x in \mathfrak{D} for every a and b of \mathfrak{D} .

(vi) If $x_1 \neq x_2$ and $y_1 \neq y_2$ the equations $y_1 = \Phi(x_1, m, b)$ and $y_2 = \Phi(x_2, m, b)$ have a unique solution pair m, b .

Every planar ternary ring \mathfrak{D} determines a finite projective plane $\mathfrak{M}(\mathfrak{D})$. The points of $\mathfrak{M}(\mathfrak{D})$ are triples (a, b, c) with a, b, c in \mathfrak{D} and may be listed as follows :

- I. The point $E_2 = (0, 1, 0)$ called the *point at infinity* (on the y -axis).
- II. The points $(1, m, 0) = E_1 + mE_2$ where m ranges over the μ elements of \mathfrak{D} . There are μ distinct points of this kind, and they and E_2 are called the *points on the line at infinity*.
- III. The μ^2 distinct *ordinary* points $(a, b, 1) = aE_1 + bE_2 + E_3$, where a and b range independently over all elements of \mathfrak{D} .

We have thus listed $\nu = \mu^2 + \mu + 1$ distinct points E_2 , $E_1 + mE_2$, $aE_1 + bE_2 + E_3$ of a set $\mathfrak{M}(\mathfrak{D})$ which is a subset of the set \mathfrak{D}_3 of all triples with elements in \mathfrak{D} . We now proceed to list the ν distinct lines of $\mathfrak{M}(\mathfrak{D})$. Each such line will be a subset consisting of $\mu + 1$ points of $\mathfrak{M}(\mathfrak{D})$, and every pair A , B of distinct points of $\mathfrak{M}(\mathfrak{D})$ will be elements on (that is, in the subset which comprises) exactly one line of $\mathfrak{M}(\mathfrak{D})$. We designate this line by $[A : B]$, and list the lines as follows:

IV. The line $[E_1 : E_2]$ is called the *line at infinity*. Its points are E_2 and all $(1, m, 0)$.

V. The μ lines $x = a$, where a ranges over all elements of \mathfrak{D} . Each such line consists of E_2 and the μ ordinary points $(a, y, 1)$ for a fixed element a of \mathfrak{D} and y ranging over all elements of \mathfrak{D} . Each such line is the line $[E_2 : (a, 0, 1)] = [E_2 : (a, y, 1)]$ for every y of \mathfrak{D} . In particular, the line $x = 0$ is the line $[E_2 : E_3]$ and is called the *y-axis*.

VI. There are μ^2 lines $y = \Phi(x, m, b)$ where m and b are any fixed elements of \mathfrak{D} . The point $(1, m, 0)$ is on the line $y = \Phi(x, m, b)$, and the ordinary points of this line are the points $(x, y, 1)$ with $y = \Phi(x, m, b)$.

Every division algebra \mathfrak{D} is a planar ternary ring, and so we have given a definition of a finite projective plane $\mathfrak{M}(\mathfrak{D})$ for every finite division algebra \mathfrak{D} .

7. The non-associative case. Consider a projective plane \mathfrak{M} , a point P of \mathfrak{M} and a line \mathfrak{L} of \mathfrak{M} . Suppose that \mathfrak{G} is a group of collineations of \mathfrak{M} satisfying the following two properties :

(a) If σ is in \mathfrak{G} then σ fixes every line through P and every point of \mathfrak{L} , that is, σ is a central collineation or perspectivity of \mathfrak{M} with center P and axis \mathfrak{L} .

(b) If A , B , and P are points of \mathfrak{M} such that $A \neq P$, $B \neq P$, neither A nor B is on \mathfrak{L} , but A , B , and P are on a line of \mathfrak{M} , then there exists a σ in \mathfrak{G} such that $A\sigma = B$. It can actually be shown that σ is unique.

Then we shall say that \mathfrak{M} is (P, \mathfrak{L}) -*transitive* or that \mathfrak{M} is (P, \mathfrak{L}) -*Desarguesian*.

A ternary ring \mathfrak{D} is called a *left Veblen-Wedderburn system* if $(x + y)z = xz + yz$ for all x, y, z in \mathfrak{D} . The lines defined in VI then have the equations $y = xm + b$. Right Veblen-Wedderburn systems are defined dually. It is well-known³ that the following results are all valid :

(c) A projective plane \mathfrak{M} is coordinatized by a left Veblen-Wedderburn system if and only if \mathfrak{M} is (P, \mathfrak{L}) -transitive with $\mathfrak{L} = [E_1 : E_2]$, and P any point on $[E_1 : E_2]$.

³ These results can be found in G. Pickert, *Projektive Ebenen*, Berlin, 1955.

(d) [The dual of (c)]. A projective plane \mathfrak{M} is coordinatized by a right Veblen-Wedderburn system if and only if \mathfrak{M} is $(E_2 : \mathfrak{L})$ -transitive for all lines \mathfrak{L} containing E_2 .

(e) [A corollary of (c) and (d)]. A plane \mathfrak{M} is coordinatized by a division ring if and only if \mathfrak{M} is (P, \mathfrak{L}) -transitive for $\mathfrak{L} = [E_1 : E_2]$ and all P on \mathfrak{L} , and for $P = E_2$ and all lines \mathfrak{L} containing E_2 .

(f) Let \mathfrak{M} be finite and (P, \mathfrak{L}) -transitive for two distinct lines \mathfrak{L} and all points P on \mathfrak{L} . Then \mathfrak{M} is Desarguesian, and $\mathfrak{M} = \mathfrak{M}(\mathfrak{D})$ for a finite field \mathfrak{D} .

(g) [The dual of (f)]. Let \mathfrak{M} be finite and (P, \mathfrak{L}) -transitive for two distinct points P and all lines \mathfrak{L} containing P . Then \mathfrak{M} is Desarguesian, and $\mathfrak{M} = \mathfrak{M}(\mathfrak{D})$ for a finite field \mathfrak{D} .

(h) [Corollary of (f) and (g)]. Let \mathfrak{M} be finite and not Desarguesian, and let \mathfrak{M} be coordinatized by a left (right) Veblen-Wedderburn system. Then no collineation of \mathfrak{M} moves the line $[E_1 : E_2]$ (the point E_2).

These results are all well known. The following result is an immediate consequence of (h).

THEOREM 3. *Let \mathfrak{D} be a finite division algebra and suppose that \mathfrak{D} is not associative. Then every collineation of the plane $\mathfrak{M}(\mathfrak{D})$ leaves the line $[E_1 : E_2]$ fixed and the point E_2 fixed.*

The following result follows from (f) and (g).

THEOREM 4. *Let \mathfrak{D} and \mathfrak{D}' be finite division algebras having the same number of elements, and suppose that \mathfrak{D} is not associative so that $\mathfrak{M}(\mathfrak{D})$ is not Desarguesian. Then every isomorphism of $\mathfrak{M}(\mathfrak{D})$ onto $\mathfrak{M}(\mathfrak{D}')$ maps the line at infinity of $\mathfrak{M}(\mathfrak{D})$ onto the line at infinity of $\mathfrak{M}(\mathfrak{D}')$, and the point at infinity E_2 of $\mathfrak{M}(\mathfrak{D})$ onto the point at infinity E'_2 of $\mathfrak{M}(\mathfrak{D}')$.*

For $\mathfrak{M}(\mathfrak{D}')$ must be non-Desarguesian. The image of $[E_1 : E_2]$ under an isomorphism δ is a line \mathfrak{L}^δ such that $\mathfrak{M}(\mathfrak{D}')$ is $(P', \mathfrak{L}^\delta)$ -transitive for every point P' on \mathfrak{L}^δ . By (f) we see that \mathfrak{L}^δ must be the line at infinity of $\mathfrak{M}(\mathfrak{D}')$. Also $\mathfrak{M}(\mathfrak{D}')$ must be $(E_2^\delta, \mathfrak{L}')$ -transitive for all lines \mathfrak{L}' of $\mathfrak{M}(\mathfrak{D}')$ containing E_2^δ , and (g) implies that $E_2^\delta = E_2$.

8. The elementary collineations of $\mathfrak{M}(\mathfrak{D})$. The collineation group $\mathfrak{G}(\mathfrak{D})$ of a plane $\mathfrak{M}(\mathfrak{D})$ defined by a nonassociative algebra \mathfrak{D} has some rather elementary subgroups. The first of these is the *translation group* $\mathfrak{T}(\mathfrak{D})$. This consists of all mappings $\tau = \tau(h, k)$, defined for every pair of elements h and k of \mathfrak{D} by

$$(43) \quad (0, 1, 0)\tau = (0, 1, 0), \quad (1, m, 0)\tau = (1, m, 0), \\ (a, b, 1)\tau = (a - h, b - k, 1).$$

Then $A\tau = A$ for every point A on $[E_1 : E_2]$. Also

$$(44) \quad [E_2 : (a, 0, 1)]\tau = [E_2 : (a - h, 0, 1)],$$

since $(a, b, 1)^{\tau} = (a - h, b - k, 1)$ is on the line $x = a - h$ for every b of \mathfrak{D} . Hence $\tau(h, k)$ merely permutes the lines $x = a$.

Finally, if $\tau = \tau(h, k)$, then

$$(45) \quad [(1, m, 0); (0, b, 1)]^{\tau} = [(1, m, 0); (0, b + hm - k, 1)].$$

For $(x, xm + b, 1)^{\tau} = (x - h, xm + b - k, 1) = (w, wm + c, 1)$ where $w = x - h$ and $wm + c = (x - h)m + c = xm - hm + c = xm + b - k$, providing that $c = b + hm - k$ as in (45). Thus $\tau = \tau(h, k)$ merely permutes the lines of $\mathfrak{M}(\mathfrak{D})$, and so defines a collineation of $\mathfrak{M}(\mathfrak{D})$. It should be clear that the set of all collineations of this type is a group $T(\mathfrak{D})$ with $\tau(0, 0)$ as identity element, and that $[\tau(h, k)]^{-1} = \tau(-h, -k)$.

The second group of elementary transformations is the group of what might be called the *shears* of $\mathfrak{M}(\mathfrak{D})$. Each shear $\tau = \tau(t)$ is defined for every t of \mathfrak{D} by

$$(46) \quad E_2^{\tau} = E_2, (1, m, 0)^{\tau} = (1, m - t, 0), (a, b, 1)^{\tau} = (a, b - at, 1).$$

Thus $\tau(t)$ leaves E_2 fixed and permutes the other points on $[E_1 : E_2]$ from which we see that

$$(47) \quad [E_1 : E_2]^{\tau(t)} = [E_1 : E_2].$$

The line $x = a$ consists of E_2 and all $(a, y, 1)$ and the images are E_2 and all $(a, y - at, 1)$. Thus every line $x = a$ is invariant under $\tau(t)$ and we write

$$(48) \quad [E_2 : (a, 0, 1)]^{\tau(t)} = [E_2 : (a, 0, 1)].$$

Finally we see that $(1, m, 0)^{\tau} = (1, m - t, 0)$ and $(x, xm + b, 1)^{\tau} = (x, xm + b - tx, 1)$ from which we see that

$$(49) \quad [(1, m, 0); (0, b, 1)]^{\tau} = [(1, m - t, 0); (0, b, 1)],$$

that is, $\tau(t)$ maps the line $y = xm + b$ onto the line $y = x(m - t) + b$.

We shall call the subgroup of $\mathfrak{G}(\mathfrak{D})$ generated by the translations and shears the *elementary* collineation group of $\mathfrak{M}(\mathfrak{D})$, and shall designate it by $\mathfrak{E}(\mathfrak{D})$. We shall refer to its elements as the *elementary collineations*. We observe that if $\rho = \tau(t)$, and $\tau = \tau(h, k)$, then

$$(50) \quad \begin{aligned} E_2^{\rho\tau} &= E_2, \quad (1, m, 0)^{\rho\tau} = (1, m - t, 0), \\ &\quad (a, b, 1)^{\rho\tau} = (a - h, b - at - k, 1), \end{aligned}$$

while

$$(51) \quad \begin{aligned} E_2^{\tau\rho} &= E_2, \quad (1, m, 0)^{\tau\rho} = (1, m - t, 0), \\ &\quad (a, b, 1)^{\tau\rho} = (a - h, b - k - at - ht, 1). \end{aligned}$$

It follows that

$$(52) \quad \tau(t)\tau(h, k) = \tau(h, k + ht)\tau(t).$$

But then every element of $\mathfrak{E}(\mathfrak{D})$ is expressible in the form $\tau = \tau(t)\tau(h, k)$, where t is uniquely determined by $(1, m, 0)^{\tau} = (1, m - t, 0)$. Also h and k

are uniquely determined by $(0, 0, 1)^r = (-h, -k, 1)$ and we have the following result.

THEOREM 5. *The elementary collineation group $\mathfrak{E}(\mathfrak{D})$ of a finite plane $\mathfrak{M}(\mathfrak{D})$ has the translation group as a normal subgroup. Every element of $\mathfrak{E}(\mathfrak{D})$ is uniquely expressible as the product $\tau = \tau(t)\tau(h, k)$ of a shear $\tau(t)$ and a translation $\tau(h, k)$.*

9. Isotopic algebras. Consider two isotopic division algebras with respective unity elements e and e' . Then there exist nonsingular linear transformations P, Q, S over \mathfrak{F} of \mathfrak{D} such that

$$(53) \quad xQ \cdot mP = (xm)S,$$

for every m and x of \mathfrak{D} where xm is the product of \mathfrak{D} and $x \cdot m$ the product of \mathfrak{D}' . Our standard representation of the points of $\mathfrak{M}(\mathfrak{D})$ is by the triples $E_2 = (0, e, 0), (e, m, 0)$, and (a, b, e) . The points of $\mathfrak{M}(\mathfrak{D}')$ are then represented by $E'_2 = (0, e', 0), (e', m, 0), (a, b, e')$. We now construct a mapping σ of $\mathfrak{M}(\mathfrak{D})$ onto $\mathfrak{M}(\mathfrak{D}')$ defined by

$$(54) \quad E_2^\sigma = E'_2, \quad (e, m, 0)^\sigma = (e', mP, 0), \quad (a, b, e)^\sigma = (aQ, bS, e').$$

Then we have

$$(55) \quad \begin{aligned} E_1^\sigma &= (e, 0, 0)^\sigma = (e', 0, 0) = E'_1, \quad E_2^\sigma = E'_2, \\ E_3E_3^\sigma &= (0, 0, e)^\sigma = (0, 0, e') = E'_3. \end{aligned}$$

Thus $[E_1 : E_2]^\sigma = [E'_1 : E'_2]$. The line $x = a$ consists of E_2 and all (a, y, e) and the image of this line under σ consists of E'_2 and all (aQ, yS, e') . Thus

$$(56) \quad [E_2 : (a, 0, e)]^\sigma = [E'_2 : (aQ, 0, e')]$$

so that σ maps the lines $x = a$ onto the lines $x = aQ$. Finally, the line $y = xm + b$ consists of $(e, m, 0)$ with image $(e', mP, 0)$ and all $(x, xm + b, e)$ where $(x, xm + b, e)^\sigma = [xQ, (xm + b)S, e']$. Since S is linear we have $(xm + b)S = (xm)S + bS$. By (53) we then have $(xm + b)S = xQ \cdot mP + bS$. But then

$$(57) \quad [(e, m, 0) : (0, b, e)]^\sigma = [(e', mP, 0) : (0, bS, e')].$$

This proves that $\mathfrak{M}(\mathfrak{D})$ and $\mathfrak{M}(\mathfrak{D}')$ are isomorphic planes. Moreover \mathfrak{D} and \mathfrak{D}' are isotopic over the prime field \mathfrak{F}_p in \mathfrak{F} .

Conversely, assume that \mathfrak{D} and \mathfrak{D}' are both division algebras, and that both have μ elements. Then $\mu = p^n$ where n is the dimension of both \mathfrak{D} and \mathfrak{D}' as being the same vector space over \mathfrak{F}_p . Suppose that σ is an isomorphism of $\mathfrak{M}(\mathfrak{D})$ onto $\mathfrak{M}(\mathfrak{D}')$ so that Theorem 4 implies that

$$(58) \quad E_2^\sigma = E'_2, \quad [E_1 : E_2]^\sigma = [E'_1 : E'_2].$$

Then

$$(59) \quad E_1^\sigma = (e', t, 0), \quad E_3^\sigma = (h, k, e'),$$

and we have already shown that there exists a collineation τ of $\mathfrak{M}(\mathfrak{D}')$ such that $E_1^{\sigma\tau} = E'_1$, $E_3^{\sigma\tau} = E'_3$. Thus we can assume that

$$(60) \quad E_1^\sigma = E_1, \quad E_2^\sigma = E_2, \quad E_3^\sigma = E_3.$$

There exists an isotope \mathfrak{D}'' of \mathfrak{D}' with unity element e and we have already shown that there exists an isomorphism ρ , of $\mathfrak{M}(\mathfrak{D}')$ onto $\mathfrak{M}(\mathfrak{D}'')$, with $(E'_2)^\rho = E_2$, $(E'_1)^\rho = E_1$, $(E'_3)^\rho = E_3$. Then $\sigma\rho$ is an isomorphism of $\mathfrak{M}(\mathfrak{D})$ onto $\mathfrak{M}(\mathfrak{D}'')$. We shall show that then \mathfrak{D} and \mathfrak{D}'' are isotopic, and thus that \mathfrak{D} and \mathfrak{D}' are isotopic. Hence let us assume that

$$(61) \quad e = e', \quad E_1^\sigma = E_1, \quad E_2^\sigma = E_2, \quad E_3^\sigma = E_3.$$

Since $[E_1 : E_2]^\sigma$ must be $[E_1 : E_2]$ there must exist a nonsingular transformation P of \mathfrak{D} , *not necessarily linear* over \mathfrak{F}_p , such that

$$(62) \quad (1, m, 0)^\sigma = (1, mP, 0), \quad 0P = 0.$$

The point $(0, b, 1)$ is on the line $[E_2 : E_3]$ and its image is also on $[E_2 : E_3]$. Thus there must exist a nonsingular transformation S of \mathfrak{D} such that

$$(63) \quad (0, b, 1)^\sigma = (0, bS, 1), \quad 0S = 0.$$

Finally $(a, 0, 1)$ and its image are on $[E_1 : E_3]$, and so there exists a nonsingular transformation Q of \mathfrak{D} such that

$$(64) \quad (a, 0, 1)^\sigma = (aQ, 0, 1), \quad 0Q = 0.$$

Observe that S and Q need not be linear. We now derive the following result.

LEMMA 7. *Let σ be an isomorphism of $\mathfrak{M}(\mathfrak{D})$ onto $\mathfrak{M}(\mathfrak{D}')$ such that $E_2^\sigma = E'_2$, $[E_1 : E_2]^\sigma = [E'_1 : E'_2]$ so that we can assume that $E_i^\sigma = E_i$ for $i = 1, 2, 3$ and that (62), (63) and (64) hold. Then*

$$(65) \quad (a, b, 1)^\sigma = (aQ, bS, 1).$$

For $(a, b, 1)$ is on the line $x = a$, that is, on $[E_2 : (a, 0, 1)]$. Hence $(a, b, 1)^\sigma = (x, y, 1)$ where $x = aQ$. Similarly $(a, b, 1)$ is on the line $[E_1 : (Q, b, 1)]$, and its image is on $[E_1 : (0, bS, 1)]$, so that $y = bS$ and our proof is complete.

We now consider the line $[(1, m, 0) : (0, b, 1)]$, that is, the line $y = xm + b$. Its image must be $[(1, mP, 0) : (0, bS, 1)]$ and this is the line $y = x \cdot mP + bS$ of $\mathfrak{M}(\mathfrak{D}')$. By Lemma 7 we have

$$(66) \quad (x, xm + b, 1) = (xQ, (xm + b)S, 1)$$

and this result is equivalent to the basic property

$$(67) \quad (xm + b)S = xQ \cdot mP + bS.$$

Take $b = 0$ and $x = e = 1$ in (67) to see that

$$(68) \quad f = 1Q, \quad f \cdot mP = mS, \quad S = PL_f^{(0)}$$

for every m of \mathfrak{D} . Then $bS = f \cdot (bP)$, and (67) for $x = 1$ becomes

$$f \cdot (m + b)P = f \cdot mP + f \cdot bP = f \cdot (mP + bP).$$

Since $f \neq 0$, and \mathfrak{D}' is a division algebra we have

$$(69) \quad (m + b)P = mP + bP.$$

But an additive transformation over \mathfrak{F}_p is linear over \mathfrak{F}_p . Hence P is a linear transformation over \mathfrak{F}_p . By (68) we know that S is also linear over \mathfrak{F}_p . We may also take $b = 0$ and $m = 1$ in (67) to obtain

$$(70) \quad xS = xQ \cdot g, \quad g = 1P, \quad QR_g^{(0)} = S,$$

so that $Q = S(R_g^{(0)})^{-1}$, and thus Q is also linear. But (67) implies that $(xm)S = xQ \cdot mP$ where P, Q and S are nonsingular linear transformations over \mathfrak{F}_p . Thus \mathfrak{D} and \mathfrak{D}' are isotopic over \mathfrak{F}_p .

We state this principal result as follows.

THEOREM 6. *Let \mathfrak{D} and \mathfrak{D}' be finite division algebras and \mathfrak{D} be not associative. Then the corresponding finite projective planes are isomorphic if and only if the algebras \mathfrak{D} and \mathfrak{D}' are isotopic over the prime field \mathfrak{F}_p .*

It should now be clear what the collineations of $\mathfrak{M}(\mathfrak{D})$ can be when \mathfrak{D} is not associative. For we have already seen that every collineation ρ of $\mathfrak{M}(\mathfrak{D})$ has the form $\rho = \sigma\tau$ where $E_i^\tau = E_i$ for $i = 1, 2, 3$ and σ is an elementary collineation, that is $\sigma = \sigma_1\sigma_2$ where σ_1 is a translation and σ_2 is a shear. Then $(1, m, 0)^\tau = (1, mP, 0)$, $(a, b, 1)^\tau = (aQ, bS, 1)$ for linear transformation P, Q, S such that

$$(71) \quad (xQ)(mP) = (xm)S.$$

But a relation of this kind is an *autotopism* of \mathfrak{D} . We can state our result as follows.

THEOREM 7. *Let \mathfrak{D} be a nonassociative division algebra over \mathfrak{F}_p . Then every collineation ρ of $\mathfrak{M}(\mathfrak{D})$ is the product $\rho = \sigma\tau$ of an elementary collineation σ , and a collineation τ such that $(0, 1, 0)^\tau = (0, 1, 0)$, $(1, m, 0)^\tau = (aQ, bS, 1)$ for linear transformations P, Q, S over \mathfrak{F}_p of \mathfrak{D} such that (71) holds and defines an autotopism of \mathfrak{D} . Conversely if \mathfrak{D} has an autotopism defined by (71) the mapping τ just defined is a collineation of $\mathfrak{M}(\mathfrak{D})$.*

Let us close these remarks by observing that (71) is equivalent to $QR_{mP} = R_mS$, as well as to $PL_{xQ} = L_xS$. Then $Q = SR_g^{-1}$, where $g = 1P$, and we have

$$(72) \quad R_{mP^{-1}} = SR_g^{-1}R_mS^{-1}.$$

Thus there can exist a nontrivial autotopism of \mathfrak{D} only if there exists an element $g \neq 0$ in \mathfrak{D} and a nonsingular linear transformation S such that $R_g^{-1}R(D) = S^{-1}R(D)S$. This implies that there exists a nonsingular linear transformation P such that $R_g^{-1}R_m = S^{-1}R_{mP^{-1}}S$. Then (71) holds for $Q = SR_g^{-1}$, $QR_{mP} = R_mS$, and our proof is complete.

We shall now define a class of finite nonassociative division algebras \mathfrak{D} .

10. **A class of finite division algebras.** Let \mathfrak{F} be a field of q elements, \mathfrak{K} be a field of degree n over \mathfrak{F} , where we assume that

$$(73) \quad q > 2, \quad n > 2.$$

Select c to be any element of \mathfrak{K} such that

$$(74) \quad c \neq a^{q-1}$$

for any a of \mathfrak{K} . Then we shall define a corresponding division algebra \mathfrak{D}_c .

We first observe that the mapping

$$(75) \quad x \rightarrow xS = x^q$$

is the generating automorphism of \mathfrak{K} over \mathfrak{F} . Define a product (x, y) by

$$(76) \quad (x, y) = x(yS) - (xS)(cy),$$

so that the mathematical system consisting of \mathfrak{K} and the product (x, y) is an algebra \mathfrak{K}_c over \mathfrak{F} . If $x \neq 0$ and $y \neq 0$ in \mathfrak{K} the equation $(x, y) = 0$ is equivalent to $(xy)(y^{q-1} - cx^{q-1}) = 0$ which contradicts our hypothesis (74) on c . Hence \mathfrak{K}_c has no divisors of zero. Then the linear transformations

$$(77) \quad \Lambda_x^{(c)} = SR_x - R_{c(xS)}, \quad \Omega_y^{(c)} = R_{yS} - SR_{cy},$$

are nonsingular for every $x \neq 0$ and every $y \neq 0$ of \mathfrak{K} . Thus we may define two nonsingular transformations A and B by

$$(78) \quad A^{-1} = \Omega_e^{(c)} = I - SR_e, \quad B^{-1} = \Lambda_e^{(c)} = S - R_c,$$

where $e = 1$ is the unity element of \mathfrak{K} .

For convenience we shall introduce the notation $R(a) = R_a$ for every a of \mathfrak{K} . Define quantities c_i in \mathfrak{K} by

$$(79) \quad c_i = c(cS)(cS^2) \cdots (cS^{i-1}) \quad (i = 1, \dots, n).$$

Then (79) implies that

$$(80) \quad c_1 = c, \quad c_2 = (c_1S)c, \dots, c_{i+1} = (c_iS)c = (cS^i)c_i \quad (i = 1, \dots, n-1),$$

and we have the special case

$$(81) \quad c_n = (c_{n-1}S)c = (cS^{n-1})c_{n-1} = \alpha e,$$

where α is the norm $\nu(c) = c(cS)(cS^2) \cdots (cS^{n-1}) = \alpha e = \alpha$ of c over \mathfrak{F} and is a nonzero element of \mathfrak{F} . The formula

$$(82) \quad (1 - \alpha)A = I + SR(c_1) + S^2R(c_2) + \cdots + S^{n-1}R(c_{n-1})$$

may then be verified by direct computation and $\alpha \neq 1$.

We shall now define a class of algebras called *twisted fields*. Each such algebra \mathfrak{D}_c is the isotope of \mathfrak{K}_c defined by

$$(83) \quad x \cdot y = (xA, yB).$$

The unity element of \mathfrak{D}_c is the element

$$(84) \quad f = eB^{-1} = e - c.$$

Then the multiplications of \mathfrak{D}_c are given by $x \cdot y = xR_y^{(c)} = yL_x^{(c)}$, where

$$(85) \quad R_y^{(c)} = (\Omega_e^{(c)})^{-1}\Omega_{yB}^{(c)}, \quad L_y^{(c)} = (\Lambda_e^{(c)})^{-1}\Lambda_{yA}^{(c)}.$$

Suppose now that g is an element of the center of \mathfrak{D}_c . Then $R_y^{(c)}R_g^{(c)} = R_{yg}^{(c)}$ for every y of \mathfrak{K} , and it follows from (77), (78), and (85) that

$$(86) \quad (R_{xs} - SR_{cx})A^{-1}(R_{hs} - SR_{ch}) = R_{zs} - SR_{cz},$$

where

$$(87) \quad yB = x, \quad gB = h, \quad (y \cdot g)B = z.$$

It is well known, from associative algebra theory, that every linear transformation T on \mathfrak{K} is uniquely⁴ expressible in the form $T = R(k_0) + SR(k_1) + S^2R(k_2) + \dots + S^{n-1}R(k_{n-1})$, for k_i in \mathfrak{K} . The fact that every S^i is an automorphism over \mathfrak{F} of \mathfrak{K} implies that $[R(k)]S^i = S^iR(kS^i)$ for every k of \mathfrak{K} . By the right member of (86) the term S^2 of the left member of (86) is zero, and we thus have

$$R_{xs}S^2R(c_2)R_{hs} - SR_{cx}SR(c_1)R_{hs} - R_{xs}SR(c_1)SR(ch) + SR(cx)SR(ch) = 0,$$

where we are using the form (82) of A . This yields the equation

$$(xS^3 - xS)c_2(hS) = (xS^3 - xS)(c_2h).$$

Since $n > 2$ we can find an element $x \neq xS^2$ in \mathfrak{K} , and our result implies that $hS = h$, that is, $h = \beta e$ for $\beta \neq 0$ in \mathfrak{F} . Then $g = hB^{-1} = (\beta e)(S - R_c) = \beta f$ is in the base field $f\mathfrak{F}$ of \mathfrak{D} . This completes a proof of the following final result.

THEOREM 8. *Every twisted field is a central division algebra over the base field \mathfrak{F} of the defining field \mathfrak{K} .*

THE UNIVERSITY OF CALIFORNIA,

LOS ANGELES, CALIFORNIA

THE UNIVERSITY OF CHICAGO,

CHICAGO, ILLINOIS

⁴ This is an immediate consequence of the result which states that the full matrix algebra is the cyclic algebra $(\mathfrak{K}, S, 1)$.

THE SIZE OF THE 10×10 ORTHOGONAL LATIN SQUARE PROBLEM

BY

L. J. PAIGE AND C. B. TOMPKINS

1. Introduction. The problem of the existence of a pair of orthogonal latin squares has been known and has remained partially solved since early days of combinatorial problems. This is a problem of forming a square array such as the array

$$\begin{array}{ccc} 11 & 22 & 33 \\ 23 & 31 & 12 \\ 32 & 13 & 21 \end{array}$$

in which each position is filled by an ordered pair of two marks in a way which uses each of the n^2 pairs possible and which uses each of the n marks exactly once as a first and once as a second mark of the pair in each row and in each column. The construction is trivially shown to be impossible for $n = 2$. There is a straightforward construction giving a set of m mutually orthogonal squares (that is, an array with the above properties but with ordered sets of m marks—a precise definition will be given in §2 below), where m is one less than the smallest factor in the canonical factorization of n into powers of different primes. However, for $n = 4k + 2$, this method does not give an orthogonal pair, and for these cases only more specialized knowledge is available. For $n = 6$, Tarry [1] proved that no orthogonal pair exists; this proof has been repeated and seems well established. For a large class of values of n , not including $n = 10$, Bruck and Ryser showed that certainly not as many as $n - 1$ mutually orthogonal pairs exist. Little is known about the case $n = 10$.¹

Relations between this problem and other combinatorial problems and applications of results will not occupy us here. Some of these relations are set forth in other papers in this volume; some applications have been explained by Mann [3]. A particularly interesting exposition of some of these problems has also been prepared by Hall [4].

Tarry's proof depended on determining a complete set of non-isomorphic latin squares on six marks. (The nature of isomorphisms admitted is stated more precisely in §2 below.) There were seventeen squares in this set, and each could be examined to show that it admitted no orthogonal

¹ Since this paper was prepared, R. C. Bose and S. S. Shrikhande, E. T. Parker, and others have shown that orthogonal pairs of latin squares exist for all $n > 6$. [See Proc. Nat. Acad. Sci. U.S.A., vol. 45 (1959) pp. 734–737 and 859–862]; however the principal part of the present paper is the computational problem of rejection of isomorphs of situations previously analyzed, and to this end it still may have some value.

complement. A complete set of non-isomorphic squares on seven marks has been constructed (initially with one omission) by Norton [5]. The number of squares in this set is 147. No complete set of squares on more than seven marks is known, but the rate of increase of the size of such sets is so great that there seems to be little merit in trying to list a complete set on ten marks.

It is also true that the rejection of isomorphic squares is a most demanding operation on modern automatic computers, and an approach is undertaken here which will permit rejection by more efficient means than any the authors have been able to devise for following the Tarry procedure. Generally, it is obviously true that the rejection of any array as isomorphic to one already studied must be based on some signature (or set of invariants) which is computable in less time than would be required for the rejection of the array without reference to earlier work. This is no easy requirement to meet with an automatic machine code. It requires, at least with machines available to us, a careful arithmetization of the rejection criterion so that the arithmetic properties of the machine can be utilized to good advantage.

The method we recommend, which still leaves a problem of seeming unacceptable extent if all possibilities must be examined, is based on techniques which are known to us to have been used by Esther Seiden [6] and by W. H. Munro [7]. However, it may be that neither intended the form of the calculation to be quite what we propose here.

The technique is to assume that an orthogonal square exists, but to reduce the number of marks, leaving the number of rows and columns unchanged, by identifying various marks. This leaves an array which is not an orthogonal pair of latin squares, but which has properties implied by its construction from such an array. Seiden identified pairs of marks so that there were five different marks and 25 different digraphs (ordered pairs of marks). Munro suggested identifying sets of five marks so that there are only four different digraphs; we shall follow Munro in this.

Prior to this calculation we have carried out several others intended to show that there is no immediately available solution or partial solution to the problem. These were routine searches using an electronic computer without rejection of isomorphs in the hope that the density of orthogonal pairs is so great that one would be discovered at a comparatively early stage of the calculation. One such search occupied many hours of time on a computer SWAC available to us, and no pair was found. However, it was crudely estimated that something like 10^{92} seconds would be required for an exhaustive search using this technique. This calculation was carried out mainly by Frank Meek. An additional search has been carried out by us with negative results reported in more detail in §3.

The technique we propose here has been applied by hand to the case of six marks, and it seems certainly no more efficient there than the Tarry approach was; there is a probability that it is less efficient. However, we

are momentarily convinced that it is more efficient for 10 marks than any other method we can presently devise, and we present it (along with a statement of our intention of carrying out such calculations) as a comparatively efficient method of seeking orthogonal pairs although not a method which will permit an exhaustive search with equipment currently available. It should be noted in this connection that the Tarry method is expanded in the step from six to ten marks not only by the large increase in the number of non-isomorphic squares but also by the difficulty of rejecting any one of these either by the criterion used by Tarry or by exhaustive search.

We suggest that the greatest contribution which might be made by the calculation we propose is the uncovering of some presently unknown invariant or method of counting which will permit later efficient calculation. We feel inadequately prepared to suggest such techniques without a great deal of experience which may be acquired by machine calculation but which seem too extensive to be undertaken by hand.

2. Definitions. One of the more troublesome aspects of a machine attack on the orthogonal latin square problem is the identification and rejection of isomorphic repetitions. Hence we shall carefully review this concept for orthogonal latin rectangles.

DEFINITION 1. Let the marks a_{ijk} be selected from the set of integers, $1, 2, \dots, n$, where

$$\begin{aligned} i &= 1, 2, \dots, \alpha \text{ is the Row Index;} \\ j &= 1, 2, \dots, \beta \text{ is the Column Index;} \\ k &= 1, 2, \dots, \gamma \text{ is the Rectangle Index;} \end{aligned}$$

further impose the restrictions

- (i) $a_{ijk} = a_{i\bar{j}k}$ implies $i = \bar{i}$;
- (ii) $a_{ijk} = a_{i\bar{j}\bar{k}}$ implies $j = \bar{j}$;
- (iii) $(a_{ijk}, a_{i\bar{j}k}) = (a_{ijk}, a_{i\bar{j}\bar{k}})$ implies $(i, j) = (\bar{i}, \bar{j})$.

Then these marks represent a set of γ orthogonal latin rectangles of α rows and β columns on n marks.

We denote such an array by $L(\alpha, \beta, \gamma, n)$.

LEMMA 1. The following transformations and hence any combination of them always transforms an $L(\alpha, \beta, \gamma, n)$ array into another of the same form:

- (i) Permute Rows,
- (ii) Permute Columns,
- (iii) Permute the k ,
- (iv) For a fixed k permute the value of the a_{ijk} by permuting the marks $1, 2, \dots, n$.

LEMMA 2. The interchange of rows and columns transforms an $L(\alpha, \beta, \gamma, n)$ into an $L(\beta\gamma\alpha)$.

DEFINITION 2. An $L(\alpha, \beta, \gamma, n)$ is said to be isomorphic with another if and only if one can be obtained from the other by a sequence of transformations of the type described in Lemma 1.

It is clear that the definition of isomorphism is an equivalence relation on the class of all sets of $L(\alpha, \beta, \gamma, n)$ and in any computation it will be desirable to retain only one set from each equivalence class. In §3, we indicate the magnitude of the computational problem when no attempt is made to remove these isomorphic possibilities.

3. Numerical results and time estimates for the 10×10 latin square problem. A simple machine method that might lead to information on the 10×10 latin square problem would be to proceed as follows: Choose a first square S_1 and then proceed to an exhaustive search for a square S_2 orthogonal to S_1 .

Certainly the 1st elements of the rows of S_2 may be chosen to be $1, 2, \dots, 10$ because an isomorphism of 2-graphs, as described in §2, allows this normalization. Without further regard to isomorphic repetitions this problem was coded for SWAC. In order to verify the coding and to obtain time estimates for the 10×10 latin square problem, this method was applied to the 17 non-isomorphic 6×6 latin squares of Tarry's listing [1]. For each of the 6×6 latin squares it required approximately 60 seconds to examine all possible squares and, of course, to find none orthogonal to it. These results are tabulated below with the following explanations: A tally of the number of times that the 1st, 2nd, \dots , 6th rows of the second square were completed was kept and these were printed out at the end of each computation. Thus, for the 6×6 square (Tarry listing #1),

1	2	3	4	5	6
2	1	4	3	6	5
3	5	1	6	2	4
4	6	2	5	1	3
5	3	6	1	4	2
6	4	5	2	3	1

it was possible to select the 1st row of the second square

1 * * * * *

in 44 ways without immediate contradiction. The first two rows

1 * * * * *
2 * * * * *

could be selected in 336 ways and so on for the 3rd, 4th and 5th rows.

It should be observed that as a check on the code the number of permissible first rows agrees with the correct number [8],

$$5! \left\{ 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} \right\} = 44.$$

TABLE 1
6 × 6 Latin Square Search

Tarry Square #	# of 1st rows completed	# of 2nd rows completed	# of 3rd rows completed	# of 4th rows completed	# of 5th rows completed
1	44	336	330	36	0
2	44	336	305	3	0
3	44	336	330	22	8
4	44	336	305	3	0
5	44	336	305	9	0
6	44	336	330	23	0
7	44	336	333	68	0
8	44	336	305	22	0
9	44	336	305	22	0
10	44	336	305	16	0
11	44	332	280	38	16
12	44	332	284	43	0
13	44	332	280	38	4
14	44	332	284	43	0
15	44	332	294	36	0
16	44	334	192	18	0
17	44	336	333	68	4

One further example was attempted. The number of latin squares orthogonal to the group table of the cyclic group of order 7 (with first column normalized to be 1, 2, 3, 4, 5, 6, 7) was computed. In this computation it was found that the number of 1st, 2nd, ..., 7th rows of the second square completed were respectively

$$265; 11,421; 62,983; 36,859; 1,498; 635; 635.$$

We made no further attempt to remove isomorphic repetitions except for the normalization of the first column as indicated above. The computing time for this example was approximately 4 hours 10 minutes.

In order to obtain an estimate of the number of machine hours to search for a pair of 10 × 10 orthogonal latin squares, the following 10 × 10 latin square was selected (almost at random).

0	1	2	3	4	5	6	7	8	9
1	8	3	2	5	4	7	6	9	0
2	9	5	6	3	0	8	4	7	1
3	7	0	9	8	6	1	5	2	4
4	6	7	5	2	9	0	8	1	3
5	0	9	4	7	8	3	1	6	2
6	5	4	7	1	3	2	9	0	8
7	4	1	8	0	2	9	3	5	6
8	3	6	0	9	1	5	2	4	7
9	2	8	1	6	7	4	0	3	5

Of course it might possibly have been that the 10×10 square selected obviously did not have a square orthogonal to it. The only general rejection criterion known to us is one due to H. B. Mann [9] and is the following:

H. B. Mann's criterion: If in a latin square of $4n + 2$ rows and columns the intersection of $2n + 1$ rows and $2n + 1$ columns contains fewer than $3n + 2$ distinct marks, then it can have no orthogonal complementary latin square.

For $n = 2$, the criterion requires each 5×5 subsquare to contain at least 8 distinct elements.

Our choice was tested, and it could not be rejected by Mann's criterion.

It is clear that the number of possible first rows for the second square orthogonal to the first is [8]

$$9! \left\{ 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} \right\} = 133,496,$$

and a reasonable approximation of the number of two line starts is $(133,496)^2/e^2$, about 2.4×10^9 .

The exhaustive search was started on SWAC with the thought in mind of completing one two-line start. However, this proved unfeasible and, extrapolating the time used (5 hours), a completion of 1 two-line start was estimated to be 200 machine hours. Consequently the total time necessary to be an exhaustive search for orthogonal squares orthogonal to our example would be approximately 4.8×10^{11} machine hours.

Clearly isomorphic repetitions must be removed more carefully than a mere normalization of the first column of the second square!

4. Identification-rejection of early isomorphs. We now introduce the following identifications :

any mark 1, 2, 3, 4 or 5 will be denoted by α ;

any mark 6, 7, 8, 9 or u (used for 10) will be denoted by β ;

the digraph $\alpha\alpha$ will be denoted by a ;

the digraph $\alpha\beta$ will be denoted by b ;

the digraph $\beta\alpha$ will be denoted by c ;

the digraph $\beta\beta$ will be denoted by d ;

either mark a or d will be denoted by x ;

either mark b or c will be denoted by y .

We shall study ten by ten arrays of these various identified marks; admissible arrays are those which would correspond to orthogonal pairs. We define admissibility below.

DEFINITION. A square array with ten rows and ten columns is an admissible α -array if and only if each element is a digraph made from the marks α and β with each of these marks occurring exactly five times as first and five times as second element of the digraphs on every row and every column and with the digraphs $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$ and $\beta\beta$ occurring twenty-five times each.

We now establish the obvious relationships between $L(10, 10, 2, 10)$ and admissible arrays.

LEMMA 3. *Under the identification of marks indicated above to give digraphs containing marks α and β , an $L(10, 10, 2, 10)$ gives an admissible α -array.*

The obvious proof to this lemma is omitted.

DEFINITION. *An admissible a -array is a square array of ten rows and ten columns, each element being one of the marks a , b , c or d , and so arranged that each row and each column has exactly as many marks d as it has marks a , and exactly as many marks c as it has marks b , and such that each of the marks occurs exactly 25 times.*

LEMMA 4. *Identification of marks of an admissible α -array yields an admissible a -array, and every admissible a -array may be generated by identification of marks of some admissible α -array.*

Proof. The proof is straightforward. Since each row (or column) has five initial marks α , and since the digraphs with these initial elements become either a 's or b 's on identification, we must have that the number of elements a plus the number of elements b on any row (or column) is five. Similarly, examining those digraphs whose second mark is β , we get that the number of b 's plus the number of d 's on any row is five. It follows immediately that the number of a 's in any row or column equals the number of d 's. Similarly, the number of b 's equals the number of c 's. The rest of the proof is trivial counting, and it will be omitted.

DEFINITION. *An admissible x -array is a square array of ten rows and ten columns, each element being one of the marks x or y , exactly fifty of the marks being x , and with an even number of marks of each kind on each row and each column.*

LEMMA 5. *Identification of marks of an admissible a -array yields an admissible x -array, and every admissible x -array may be generated by identification of marks of some admissible a -array.*

Proof. The proof of the first part of the lemma is immediate. The proof of the second part will be given here. To get an a -array from which an x -array could have been derived by identification, replace any x by a and another x on the same row (there must be another because the number of x 's in each row is even) by d .

Not counting a 's and d 's, there are still an even number of x 's in each row, but an odd number in two columns. In each of these columns fill in an a or a d to complement the one already filled in; there is now an even number of x 's in every column but possibly an odd number in two rows. The procedure is now continued, always correcting the two rows or two columns with an odd number of x 's and always keeping the same number of a 's and

d 's in each row or column with an even number of a, d marks. Eventually the process must end with a smaller number of x 's and with every row and every column having an even number of x 's (possibly zero) and an equal number of a 's and d 's. The process may be restarted if necessary, and then the y 's may be replaced by b 's and c 's analogously. This completes the proof.

DEFINITION. *The set of isomorphisms on a -arrays includes the following transformations:* (i) *any permutation of rows*; (ii) *any permutation of columns*; (iii) *replacement of a by d and of d by a* ; (iv) *replacement of b by c and of c by b* ; (v) *replacement of a by b , b by a , c by d and d by c* . *Isomorphisms of a -arrays and x -arrays are defined analogously.*

We state without proof the following lemma.

LEMMA 6. *If an a -array is obtained by identification from an $L(10, 10, 2, 10)$ then any isomorph of the a -array may be obtained by identification from some isomorph of the $L(10, 10, 2, 10)$.*

The converse of this lemma is presumably not true, for the class of isomorphic transformations admitted for latin squares permits many permutations of the marks which are not admitted among the isomorphic transformations specified for a -arrays. However, our ultimate problem is with latin squares, and we shall avail ourselves of any help which can be obtained by exploiting the full set of isomorphic transformations admitted for latin squares. In particular, we shall always be able to choose one row to have the form 11, 22, 33, 44, 55, 66, 77, 88, 99, uu .

We shall use isomorphs to restrict some of the forms admissible for a -arrays. In particular, we denote by r_σ , $\sigma = 0, 1, \dots, 5$, the number of rows with exactly σ marks a in an a -array. Immediate restrictions on a -arrays may be made on the basis of the following lemma.

LEMMA 7. *The numbers r_σ satisfy the two relations*

$$\sum r_\sigma = 10, \quad \sum \sigma r_\sigma = 25.$$

Furthermore given any a -array there always exists an isomorph with

$$r_5 \geq r_0, \quad \text{and} \quad r_{5-\tau} \geq r_\tau \\ \text{if } r_{5-\sigma} = r_\sigma \text{ for all } \sigma < \tau.$$

Proof. The first sum states simply that the array has ten rows, and the second states that the array has 25 marks a . The inequalities are arrived at by interchanging the roles of a and b and those of c and d (an allowed isomorphism) if necessary.

The conditions of Lemma 7 are not sufficient for existence of an a -array. Conditions must be added in order to give the required balance in both rows and columns; explicitly, these conditions are easiest to apply to the x -arrays,

where they demand that each column have an even number of x 's (each row has twice as many x 's as it does a 's).

For example, the array with $r_5 = r_0 = 5$ and $r_4 = r_3 = r_2 = r_1 = 0$ satisfies the conditions of Lemma 7; however, clearly all the a 's and d 's occur in five rows, which they completely fill. Hence there must be different numbers of a 's and d 's in each column, 5 being an odd number.

This adjustment may be made arithmetically. To this end consider the rectangular x -array consisting of the first k rows of a square x -array being built. We shall denote by σ_k the number of a 's in the k th row (half the number of x 's in the k th row) and by ξ_k the number of columns with an odd number of x 's in this 10 by k rectangle. Finally, denote by η_k the number of x 's in the k th row which fall in columns containing an odd number of x 's in the first $k - 1$ rows. It is clear that

$$\xi_k = \xi_{k-1} + 2\sigma_k - 2\eta_k$$

where ξ_{k-1} was the number of columns with an odd number of x 's in the first $k - 1$ rows, $\xi_{k-1} - \eta_k$ is the number of these which continue to have an odd number of x 's in k rows, and $2\sigma_k - \eta_k$ is the number of columns which change from having an even number of x 's in $k - 1$ rows to an odd number in k rows. Incidentally, it is apparent that ξ_k is always even. Actually we prove that ξ_k can always take on any even value between its maximal and minimal values (always permitting readjustment of the placing of x 's in earlier rows).

LEMMA 8. *There exist even numbers μ_k and ν_k depending on σ_j for $J \leq k$ such that for any arrangement of the x 's in the first k rows $\nu_k \leq \xi_k = \mu_k$, and such that ξ_k may take any even value in this range.*

Proof. The lemma may be proved by induction. It is obviously true for $k = 1$, where $\nu_k = \nu_1 = 2\sigma_1 = \mu_1$. The rest of the induction requires no more than an examination of the formula for ξ_k above and the restrictions which must be imposed in η_k . We omit this proof but give explicit formulas for computing μ_k and ν_k in a form suitable for machine calculation:

If $10 - \mu_{k-1} - 2\sigma_k$ and $10 - \nu_{k-1} - 2\sigma_k$ differ in sign, or if either is zero then $\mu_k = 10$; if both are negative, then $\mu_k = 20 - \nu_{k-1} - 2\sigma_k$, and if neither is negative then $\mu_k = \mu_{k-1} + 2\sigma_k$.

If $\mu_{k-1} - 2\sigma_k$ and $\nu_{k-1} - 2\sigma_k$ differ in sign or if either is zero, then $\gamma_k = 0$; if both are negative then $\nu_k = 2\sigma_k - \mu_{k-1}$; if neither is negative then $\gamma_k = \nu_{k-1} - 2\sigma_k$.

This rule has been used to calculate ν_{10} and to reject unsuitable combinations of r_o ; these are values for which $\nu_{10} \neq 0$.

We note here (as above) that we are ultimately interested in the study of orthogonal latin squares. Therefore we introduce one further restriction

which is not immediately introduced from the definition of isomorphisms of x -arrays. It is clear that by permuting the first and the second marks in an $L(10, 10, 2, 10)$ we can require that the last row (or any row) have the form 11, 22, 33, 44, 55, 66, 77, 88, 99, uu . Here for convenience we use the elementary mark u for ten. There is a row for which $\sigma_k = 5$, and therefore we note that every $L(10, 10, 2, 10)$ is isomorphic to one with $r_5 > 0$. We restrict the admissible sets of r_σ to those satisfying this condition.

The calculation of admissible sets of numbers r_σ was carried out on SWAC, rejecting obvious isomorphs. The time of the calculation was so short that no subtleties of coding were involved. All values from 0 to 5 were tried for each of the numbers r_0 through r_4 , r_5 was taken to be 10 minus the sum of the others, the set being rejected if r_5 was less than 1. Then the condition $\sum \sigma r_\sigma = 25$ was tested, and the other condition stated in Lemma 7. If all these tests were passed, the value of v_{10} for this set of r_σ was computed, and the set retained if $v_{10} = 0$. The calculation required 1 minute 50 seconds on SWAC, and 64 acceptable sets of r_σ values resulted. This calculation was repeated with the same results after some weeks, and the probability of error is presumably negligibly small.

It is true that at least one admissible x -array exists for each of these admissible sets, and usually many exist. Furthermore, these are the only admissible x -arrays. We now turn our attention to the construction of these arrays.

To this end, note that by permuting rows we can arrange for σ_k to be a monotone non-decreasing function of k . Because of the restrictions we have already imposed, $\sigma_{10} = 5$.

Given any admissible set r_σ there are usually several sets η_k which are admissible with it in the sense that each set η_k causes an even number of x 's to appear in each column (that is, implies $\xi_{10} = 0$). These are calculable using no more than $6^6 = 46656$ trials, as indicated below. We first note that values are known for $\xi_1 = 2\sigma_1$, $\xi_{10} = 0$, $\xi_9 = 10$ and $\xi_8 = 10 - 2\sigma_9$. The first of these follows from the fact that every x in the first row creates an odd column, the second is the admissibility condition on η_k sets, the third follows from the restriction already imposed that $\sigma_{10} = 5$, and the fourth from the fact that $\xi_9 = 10$ implies that $\eta_9 = 0$ and that $2\sigma_9 = 10 - \xi_8$; this last symbolizes the condition that every x in the ninth row is placed in a column with an even number of x 's above and that every such column receives an x in the ninth row.

The values of ξ_k in the intervening rows are certainly from one of the six values $2^{-1}\xi_k = 0, 1, \dots, 5$, and this gives our early estimate of extent of the calculation. From a formula already given, all η_k may be calculated from any set ξ_k .

$$\eta_k = \frac{1}{2}\xi_{k-1} - \frac{1}{2}\xi_k + \sigma_k.$$

Admissibility of any η_k value depends on four inequalities. These in order of application in the calculation are

- (i) $\eta_k \geq 0,$
- (ii) $\eta_k \leq 10 - \xi_k,$
- (iii) $\eta_k \leq 2\sigma_k,$
- (iv) $\eta_k = \xi_{k-1}.$

Of these (i), (iii) and (iv) are obvious, saying respectively that no negative number of odd columns are made even, that the number made even is limited by the number of x -marks available, and the number is limited by the number of odd columns available. Relation (ii) follows from the formula for η_k above and the relation

$$2\sigma_k - \eta_k \leq 10 - \xi_{k-1}$$

which states that the number of even columns made odd is limited by the number of even rows available. These four relations are the only independent ones required for admissible η_k .

Relations (i) and (ii) impose upper limits on ξ_k , $\xi_k \leq \xi_{k-1} + 2\sigma_k$ and $\xi_k \leq 20 - 2\sigma_k - \xi_{k-1}$ respectively. Conditions (iii) and (iv) impose lower limits on ξ_k . The sets ξ_k are tried exhaustively in order of increasing k with rejection of whole blocks in case of early failure.

5. Non-arithmetized procedure. The computation of the admissible values of η_k -sets for each r_σ -set completes the portion of the problem that precedes the possible construction of $L(10, 10, 2, 10)$. We contemplate the possible construction in three steps:

- (i) For each admissible η_k -set associated with a r_σ -set we assemble a complete set of non-isomorphic x -arrays.
- (ii) For each such x -array we construct a complete set of non-isomorphic a -arrays.
- (iii) For each such a -array the actual substitution of marks is attempted in essentially every possible way.

We will content ourselves with a fairly complete sketch of step (i) and leave the remainder undescribed, since it is done analogously. We note in the beginning that x -arrays corresponding to different r_σ -sets are non-isomorphic, at least under isomorphisms in which only x -arrays are considered (although they may ultimately correspond to isomorphic latin squares, where a wider class of isomorphisms has been described). The transformations which are admissible for the isomorphisms here are permutations of columns, permutations of sets of rows all having the same value σ_k , and in case $r_5 = r_0$, $r_4 = r_1$, $r_3 = r_2$, reversal of the roles of x and y . We shall ignore this last minor set of isomorphisms here.

Generally, the procedure is inductive, adding a row to a rectangle already assembled. At any stage there are between one and ten classes of columns in the rectangle already assembled; two columns are considered to be different if they are not identical in marks, including the rows in which the marks are located.

There seems to be no way around an exhaustive procedure here, but for any set of identical columns, any x -marks inserted in the next row may be placed as far to the right as possible among this set of columns. Thus, a square is built up in the following way:

(i) at the beginning all columns are equivalent, and the first row is built uniquely by placing $2\sigma_1$ x -marks in the last columns;

(ii) if $5 > \sigma_1 > 0$, then there are two classes of columns; of these $2\sigma_1$ are odd and the remainder are even. For the second row, exactly $2\eta_2$ x -marks are put as far to the right as possible in the odd columns and any remaining x -marks required are put as far to the right as possible in the even rows;

(iii) there may now be four classes of columns,

$$\begin{matrix} y & y & x & x \\ y & x & y & x \end{matrix}$$

For the first try in the third row, the odd columns are filled as much as possible from the right (up to the number permitted by η_3), and the even columns are filled from the right as much as possible with the remaining x -marks. No permutation within the various classes of columns is allowed later, but in later tries the allocation of the η_3 x -marks to odd columns of the two available classes and the allocation of the remaining x -marks to the classes of even columns will be changed;

(iv) in later tries, where there is no change in the value of σ_k from σ_{k-1} of the last row, a permutation of these rows must be attempted to determine whether after this permutation it is possible to permute columns to create a rectangular array already studied—that is one in which the first row changed has its left-most x -mark to the right of the position of the left-most mark in this row before the test.

In hand computed cases of step (i), it has been found that the number of different x -arrays possible for each fixed (η_k, r_σ) -set is large. However, the principal difficulty will arise in step (iii) where all possible permutations of marks must be attempted and (supposedly) rejected. Steps (i), (ii) and (iii) await the availability to us of a machine with large high-speed storage which we expect shortly.

BIBLIOGRAPHY

1. G. Tarry, *Le problème de 36 officiers*, C.R. de l'Ass. Fr. pour l'Av. de Sci. Nat. 29 (1900) Part I pp. 122–123; Part II pp. 170–203.
2. R. H. Bruck and H. J. Ryser, *The non-existence of certain finite projective planes*, Canad. J. Math. vol. 2 (1950) pp. 93–99.

3. H. B. Mann, *Analysis and design of experiments*, New York, 1949.
4. M. Hall, Jr., *Projective planes and related topics*, Lectures at California Institute of Technology, April, 1954.
5. H. W. Norton, *The 7 × 7 squares*, Ann. Engin. vol. 9 (1939) pp. 269–307.
6. Esther Seiden, Private communication.
7. W. Munro, Private communication.
8. E. Netto, *Lehrbuch der Combinatorik*, New York, Chelsea Press, pp. 66–68 (reprint of 1927 edition).
9. H. B. Mann, *On orthogonal Latin squares*, Bull. Amer. Math. Soc. vol. 50 (1944) pp. 249–257.

UNIVERSITY OF CALIFORNIA,
LOS ANGELES, CALIFORNIA

This page intentionally left blank

SOME COMBINATORIAL PROBLEMS ON PARTIALLY ORDERED SETS

BY

R. P. DILWORTH

1. Introduction. This paper is concerned with some combinatorial problems related to the following theorem on partially ordered sets:

The minimal number of chains in the representation of a finite partially ordered set P as a set union of chains is equal to the maximal number of mutually non-comparable elements of P .

The theorem was first formulated and proved in connection with a problem on subdirect union representations of distributive lattices (Dilworth [1]). It was well known that any distributive lattice could be represented as a subdirect union of chains and the problem concerned the minimal number of chains required for such a representation. Now it is easily shown that subdirect union representations of a finite distributive lattice in terms of chains correspond to the decomposition of the partially ordered set of join irreducibles into a set union of chains. On the other hand, a mutually non-comparable set of n join irreducibles leads directly to an element s of the distributive lattice which covers exactly n elements of the distributive lattice. These elements are the *cover set* of s . Conversely, every cover set of an element of the lattice produces a collection of non-comparable join irreducibles having the same number of elements. Thus the theorem on partially ordered sets gives the following theorem for distributive lattices.

The minimal number of chains in the representation of a finite distributive lattice as a subdirect union of chains is equal to the maximal number of elements in the cover sets of the lattice.

As a by-product of the investigation, it was observed that a wide variety of theorems on representatives of sets could be derived directly from the theorem by applying it to the partially ordered set obtained from a collection of subsets together with the elements themselves by ordering each subset to the elements which it contains. In particular, the Rado-Hall theorem on representatives of sets is an immediate consequence of the theorem.

The next development in connection with the theorem was the discovery by Dantzig and Hoffman [2] that the problem of finding the minimal decomposition of a finite partially ordered set into disjoint chains can be formulated as a transportation type linear programming problem and that the theorem follows from the duality theorem of linear inequality theory. If P_1, \dots, P_n are the elements of a finite partially ordered set Dantzig and Hoffman

consider the array $\{x_{ij}\}$, $i, j = 0, 1, \dots, n$ and require that x_{00} be a maximum subject to the restrictions

$$\sum_{i=0}^n x_{ij} = \begin{cases} n, & j = 0, \\ 1, & j \neq 0; \end{cases}$$

$$\sum_{j=0}^n x_{ij} = \begin{cases} n, & i = 0, \\ 1, & i \neq 0; \end{cases}$$

$$x_{ij} \geq 0; \quad x_{ij} = 0 \quad \text{if } P_i \not\geq P_j \quad \text{or if } i = j \neq 0.$$

There are always integers x_{ij} which give the required maximum. The maximum value for x_{00} is $n - m$ where m is the minimal number of chains in the representation of the partially ordered set as a set union of chains. A chain of the minimal representation may be obtained as follows: Select P_j so that $x_{0j} = 1$. Then there is exactly one element x_{jk} in the j th row which is equal to 1. If $k \neq 0$, $P_j \geq P_k$. Similarly there is exactly one element x_{kl} in the k th row which is equal to 1. If $l \neq 0$, then $P_j \geq P_k \geq P_l$. This process continues until an $x_{r0} = 1$ is obtained in which case P_r terminates the chain. One of the important consequences of the work of Dantzig and Hoffman is the fact that the techniques of linear programming may be used to construct the chains of a minimal representation.

Now it had earlier been observed by Birkhoff that there is a certain formal resemblance between the partially ordered set theorem and a theorem of König on linear graphs. If V denotes the set of vertices of a linear graph, let $V = V_1 + V_2$ be a partition of V into two disjoint subsets. A *cut* of the graph is a collection of vertices such that every edge joining a vertex of V_1 to a vertex of V_2 has at least one of its end points in the collection. A *join* is a collection of edges joining vertices of V_1 to vertices V_2 such that no two edges of the collection have a vertex in common. König's theorem asserts that *the minimal number of vertices forming a cut of the graph is equal to the maximal number of edges making up a join of the graph*.

The König theorem can also be obtained as an application of the duality theorem of linear programming. Furthermore, the formulation of the problem in linear programming terms is so closely similar to the transportation problem described above that Fulkerson [3] succeeded in deducing the partially ordered set theorem from König's theorem and conversely.

The method employed by Fulkerson is as follows: If P_1, \dots, P_n are the elements of the partially ordered set, let V be the set of $2n$ vertices $a_1, \dots, a_n, b_1, \dots, b_n$ where a_i is joined to b_j by an edge if and only if $P_i > P_j$. Let $V_1 = \{a_1, \dots, a_n\}$ and $V_2 = \{b_1, \dots, b_n\}$. If D is a decomposition of the partially ordered set into chains, the collection of edges $a_i b_j$ where P_i covers P_j in one of the chains clearly forms a join J of the graph. Conversely every join leads to a decomposition of the partially ordered set. It is easily shown that $n(D) + n(J) = n$. Furthermore if U is a maximal non-comparable subset of the partially ordered set, then each $P_i \notin U$ is such that

either $P_i > u$ or $u > P_i$ for some $u \in U$. Set $a_i \in C$ if $P_i > a$ and $b_i \in C$ if $u > P_i$. Then C is a cut of the graph and $n(U) + n(C) = n$. The equivalence of the two theorems can clearly be deduced from these formulas.

As noted above, the Rado-Hall theorem on representatives of sets is an immediate consequence of the decomposition theorem for partially ordered sets. On the other hand, it is well known that the König theorem on graphs can be easily derived from the Rado-Hall theorem. Thus, making use of Fulkerson's correspondence between partially ordered sets and graphs, the partially ordered set theorem can be proved from the Rado-Hall theorem. However such a derivation is clearly indirect and moreover requires the introduction of a somewhat artificial auxiliary partially ordered set. Actually, there is a much more intimate relation between the partially ordered set theorem and the Rado-Hall theorem. I will develop this relationship in the following section and will also show that there is still another connection between the theorem and properties of distributive lattices.

2. The theory of maximal non-comparable sets. Let us consider first a class of partially ordered sets for which the decomposition theorem is immediately equivalent to the Rado-Hall theorem. Let $A = \{a_1, \dots, a_n\}$, $B = \{b_1, \dots, b_n\}$ be two sets each of which contain n distinct elements. A class of pairs (a_i, b_j) such that $a_i, b_j \notin A \cap B$ is selected. We write $a_i > b_j$ if (a_i, b_j) belongs to the selected class. Denote this partially ordered set by $P(A, B)$. For each a_i let S_{a_i} denote the set of b_j such that $a_i \geq b_j$.

LEMMA 2.1. *The maximal number of mutually non-comparable elements in $P(A, B)$ is n if and only if the union of any k of the subsets S_{a_i} contains at least k elements.*

For if $S_{a_1} \vee \dots \vee S_{a_k} = S$ contains less than k elements, $(B - S) \vee \{a_1, \dots, a_k\}$ is a non-comparable subset of $P(A, B)$ with more than n elements. Conversely, if C is a non-comparable subset of $P(A, B)$ with more than n elements let $C \cap A = \{a_1, \dots, a_k\}$. Then $C \cap B$ contains more than $n - k$ elements. But since $(S_{a_1} \vee \dots \vee S_{a_k}) \wedge C = \emptyset$, it follows that $S_{a_1} \vee \dots \vee S_{a_k}$ contains less than k elements.

LEMMA 2.2. *The minimal number of chains whose union contains $P(A, B)$ is n if and only if there exists a set of distinct representatives for the sets S_{a_1}, \dots, S_{a_n} .*

For if $P(A, B)$ is the set union of n chains, then each chain contains exactly one element a_i of A and b_j of B . The b_j provide a set of distinct representatives for the sets S_{a_i} . Clearly a set of distinct representatives for the S_{a_i} gives a representation of $P(A, B)$ as a set union of n chains.

From Lemmas 2.1, 2.2, and the Rado-Hall theorem it follows that if the maximal number of mutually non-comparable elements in $P(A, B)$ is n , then $P(A, B)$ is the set union of n distinct chains. These chains, in turn, define a 1-1 mapping of A onto B , $a_i \rightarrow b_j$ such that $a_i \geq b_j$.

Now let P be a finite partially ordered set in which n is the maximal number of mutually non-comparable elements. The proof that P is the set union of n chains can be easily reduced to the case in which each element of P belongs to at least one non-comparable set having n elements. For let P' be the union of all such non-comparable sets in P . Then if P' is the set union of n chains, let C be one of these chains. Suppose $P - C$ contains a non-comparable subset with n elements. Then the elements of this set belong to P' and at least one of them belongs to C contrary to the assumption that the elements belong to $P - C$. Thus the maximal number of mutually non-comparable elements in $P - C$ is $n - 1$. Induction on n then completes the proof for the partially ordered set P .

We shall assume now that each element of P belongs to at least one non-comparable set of n elements.

Let L denote the collection of non-comparable subsets of P having n elements. We shall call these sets n -sets. If Q and Q' are n -sets, a partial ordering on L is defined by the relation $Q \leq Q'$ if and only if for each $q \in Q$, there exists $q' \in Q'$ such that $q \leq q'$. The following lemma shows that this relation is self dual.

LEMMA 2.3. $Q \leq Q'$ if and only if $q' \in Q'$ implies that there exists $q \in Q$ such that $q \leq q'$.

For let $Q \leq Q'$ and let q' be an arbitrary element of Q' . Then since $\{q'\} \vee Q$ is no longer non-comparable, there exists $q \in Q$ which is comparable with q' . If $q \leq q'$, the lemma follows. If $q' \leq q$, then it follows from $Q \leq Q'$ that $q'' \in Q'$ exists such that $q \leq q''$. But then $q' \leq q''$ and since Q' is a non-comparable set we must have $q' = q''$. Hence $q \leq q'$. Thus the necessity of the condition is proved and a dual argument gives the sufficiency.

If Q and Q' are arbitrary n -sets of P , let M denote the maximal elements of $Q \vee Q'$ and let N denote the minimal elements of $Q \vee Q'$. Clearly M and N are non-comparable subsets of P . We shall show that they are n -sets of P .

LEMMA 2.4. $M \vee N = Q \vee Q'$.

Clearly $M \vee N \leq Q \vee Q'$. Suppose that $s \in Q \vee Q'$ but $s \notin M \vee N$. If $s \in Q$, then since $s \notin N$ there exists $s_1 \in Q \vee Q'$ such that $s_1 < s$. Similarly since $s \notin M$ there exists $s_2 \in Q \vee Q'$ such that $s < s_2$. Since s_1 and s are comparable, it follows that $s_1 \in Q'$. Similarly we find that $s_2 \in Q'$. But $s_1 < s_2$ contrary to the non-comparability of Q' . If $s \in Q'$, a similar contradiction can be obtained for Q . Thus s must belong to $M \vee N$ and the lemma is proved.

LEMMA 2.5. $M \wedge N = Q \wedge Q'$.

For let $s \in M \wedge N$. Then if s is comparable with an element t of $Q \vee Q'$ we must have $s = t$ since s is both maximal and minimal. But s is comparable with at least one element of Q and hence $s \in Q$. Similarly $s \in Q'$

and thus $s \in Q \wedge Q'$. Conversely, if $s \in Q \wedge Q'$ then $s = q = q'$ where $q \in Q$ and $q' \in Q'$. Let $x \geq s$. Then if $x \in Q$ we have $x \geq q \Rightarrow x = q = s$. If $x \in Q'$, then $x \geq q' \Rightarrow x = q' = s$. Thus $x \geq s \Rightarrow x = s$ and hence $s \in M$. A dual argument shows that $s \in N$ and hence that $s \in M \wedge N$.

LEMMA 2.6. *M and N are n-sets of P.*

For let $n(A)$ denote the number of elements in A . Then $n(M) + n(N) = n(M \vee N) + n(M \wedge N) = n(Q \vee Q') + n(Q \wedge Q') = n(Q) + n(Q') = 2n$. But since M and N are sets consisting of mutually non-comparable elements, we have $n(M) \leq n$ and $n(N) \leq n$. But then $n(M) = n(N) = n$ and M and N are n -sets.

LEMMA 2.7. *M, N are respectively the l.u.b. and g.l.b. of Q and Q' under the partial ordering of L.*

For $q \in Q$ implies $q \leq s$ where s is a maximal element of $Q \vee Q'$. Hence $s \in M$ and thus $Q \leq M$. Similarly $Q' \leq M$. Now let $Q \leq R$, $Q' \leq R$ where R is an n -set. Let $s \in M$. Then $s \in Q$ or $s \in Q'$. If $s \in Q$ there exists $r \in R$ such that $s \leq r$. A similar argument holds if $s \in Q'$. Hence $M \leq R$ and M is the l.u.b. of Q and Q' in L . A dual argument shows that N is the g.l.b. and the proof of the lemma is thus complete.

It follows from Lemma 2.7, that L is a lattice under the partial ordering of n -sets. In fact, a much stronger result holds.

THEOREM 2.1. *L is a distributive lattice.*

Proof. Let Q_1 , Q_2 , and Q_3 be n -sets of P . Let $s \in Q_1 \cap (Q_2 \cup Q_3)$. Then s is a minimal element of $Q_1 \vee (Q_2 \cup Q_3)$. Let us suppose first that $s \in Q_1$. Then since $Q_1 \cap (Q_2 \cup Q_3) \leq Q_2 \cup Q_3$, it follows that $s \leq t$ where $t \in Q_2 \cup Q_3$. Then $t \in Q_2 \vee Q_3$ and, by symmetry it will suffice to consider the case $t \in Q_2$. Now suppose that s is not a minimal element of $Q_1 \vee Q_2$. Then $r < s$ where $r \in Q_1 \vee Q_2$. Since $s \in Q_1$ it follows that $r \notin Q_1$ and hence $r \in Q_2$. But then $r < t$ where $r, t \in Q_2$ contrary to the non-comparability of Q_2 . Thus $s \in Q_1 \cap Q_2$. But $Q_1 \cap Q_2 \leq (Q_1 \cap Q_2) \cup (Q_2 \cap Q_3)$ implies that $s \leq s'$ where $s' \in (Q_1 \cap Q_2) \cup (Q_2 \cap Q_3)$. Next let us suppose that $s \notin Q_1$. Then $s \in Q_2 \cup Q_3$ and s is a maximal element of $Q_2 \vee Q_3$. Since $Q_1 \cap (Q_2 \cup Q_3) \leq Q_1$, it follows that there exists $t \in Q_1$ such that $s \leq t$. Again, by symmetry, we may suppose that $s \in Q_2$. If s is not a minimal element of $Q_1 \vee Q_2$ then $r < s$ where $r \in Q_1$. But then $r < t$ contrary to the non-comparability of Q_1 . Thus $s \in Q_1 \cap Q_2$ and hence we again have $s \leq s'$ where $s' \in (Q_1 \cap Q_2) \cup (Q_2 \cap Q_3)$. Thus $Q_1 \cap (Q_2 \cup Q_3) \leq (Q_1 \cap Q_2) \cup (Q_2 \cap Q_3)$ and L is distributive.

For the proof of the next theorem the following lemma is required.

LEMMA 2.8. *If $q < q'$ where $q \in Q$ and $q' \in Q'$, then $q \in Q \cap Q'$ and $q' \in Q \cup Q'$.*

For if q is not a minimal element of $Q \vee Q'$, there exists $q'' \in Q \vee Q'$ such that $q'' < q$. But then $q'' \in Q'$ and $q'' < q'$ contrary to the non-comparability of Q' . Thus $q \in Q \cap Q'$ and a similar argument shows that $q' \in Q \cup Q'$.

THEOREM 2.2. *Let $Q_1 \leq Q_2 \leq \cdots \leq Q_m$ be a maximal chain of L . Then $P = Q_1 \vee Q_2 \vee \cdots \vee Q_m$.*

Proof. Let $q \in P$. Then q is comparable with at least one element of Q_i for each i . Now by hypothesis $q \in Q$ for some $Q \in L$. Since $Q_1 \leq \cdots \leq Q_m$ is a maximal chain, Q_1 is the minimal n -set of L and hence $Q_1 \leq Q$. Thus $q_1 \leq q$ for some $q_1 \in Q_1$. Let k be maximal such that $q_k \leq q$ for some $q_k \in Q_k$. If $k = m$, then since Q_m is the maximal n -set of L , we have $q \leq q'_m$ for some $q'_m \in Q_m$. But then $q_m \leq q \leq q'_m$ and thus $q = q_m = q'_m \in Q_m \subseteq Q_1 \vee \cdots \vee Q_m$. If $k < m$, then q is comparable with an element of Q_{k+1} and hence by the maximal property of k we must have $q < q_{k+1}$ where $q_{k+1} \in Q_{k+1}$. Now suppose that $q_k < q$. Then by Lemma 2.8, $q \in Q_k \cup Q$. Again by Lemma 2.8 since $q < q_{k+1}$ we have $q \in (Q_k \cup Q) \cap Q_{k+1}$. But $Q_k \leq (Q_k \cup Q) \cap Q_{k+1} \leq Q_{k+1}$ and by the maximal property of the chain we must have $Q_k = (Q_k \cup Q) \cap Q_{k+1}$ or $(Q_k \cup Q) \cap Q_{k+1} = Q_{k+1}$. Hence either $q \in Q_k$ or $q \in Q_{k+1}$. But $q_k < q$ contradicts $q \in Q_k$ and $q < q_{k+1}$ contradicts $q \in Q_{k+1}$. Thus we must have $q = q_k$. Hence $q \in Q_k \subseteq Q_1 \vee \cdots \vee Q_m$.

Making use of this theorem we can now apply the Rado-Hall theorem to give a simple direct construction of a representation of P as a set union of n -chains. For the maximal number of non-comparable elements in $P(Q_2, Q_1)$ is n . Thus, by the Rado-Hall theorem there is a one-to-one mapping of Q_1 onto Q_2 such that if $q_1 \rightarrow q_2$ then $q_1 \leq q_2$. Similarly there is a one-to-one mapping of Q_2 onto Q_3 such that if $q_2 \rightarrow q_3$ then $q_2 \leq q_3$. Continuing in this manner we get a chain $q_1 \leq q_2 \leq \cdots \leq q_m$. The n possible choices for q_1 give the n chains of the representation. Clearly the union of the chains is $Q_1 \vee \cdots \vee Q_m$ which by Theorem 2.2 is the partially ordered set P .

REFERENCES

1. R. P. Dilworth, *A decomposition theorem for partially ordered sets*, Ann. of Math. vol. 51 (1950) pp. 161–166.
2. G. B. Dantzig and A. Hoffman, *On a theorem of Dilworth*, Contributions to linear inequalities and related topics, Annals of Mathematics Studies, no. 38.
3. D. R. Fulkerson, *Note on Dilworth's decomposition for partially ordered sets*, Proc. Amer. Math. Soc. vol. 7 (1956) pp. 701–702.
4. P. Hall, *On representatives of subsets*, J. London Math. Soc. vol. 10 (1935) pp. 26–30.
5. D. König, *Theorie der Graphen*, New York, Chelsea, 1950.

CALIFORNIA INSTITUTE OF TECHNOLOGY,
PASADENA, CALIFORNIA

AN ENUMERATIVE TECHNIQUE FOR A CLASS OF COMBINATORIAL PROBLEMS

BY

R. J. WALKER¹

1. The general problem. Many combinatorial problems involve the construction of one or more (possibly all) ordered sets $A = \{a_1, \dots, a_n\}$, where the a_i are elements of a finite set U and the elements of a set A are subject to certain restrictions. The most common example of such a set A is a permutation; here n is the number of elements in U and the restriction is that $a_i \neq a_j$ if $i \neq j$. More complicated examples are the following:

(i) *The problem of the queens.* Place eight queens on a chess board so that no two attack each other. Since there must be one queen in each file (column) let the queen in the i th file be on the a_i th rank (row). Then $U = \{1, 2, \dots, 8\}$, $n = 8$, and the restrictions are

$$\begin{aligned} a_i &\neq a_j && \text{if } i \neq j, \\ a_i - i &\neq a_j - j && \text{if } i \neq j, \\ a_i + i &\neq a_j + j && \text{if } i \neq j. \end{aligned}$$

(ii) *Orthogonal latin squares.*² Here U is the set of ordered pairs $a = (b, c)$, $b, c = 1, \dots, N$. The natural ordering of elements of A is not linear but quadratic, i.e. $A = \{a_{ij}\}$, $i, j, = 1, \dots, N$. The restrictions are:

For $(i, j) \neq (i', j')$, $a_{ij} \neq a_{i'j'}$;

For $i \neq i'$, a_{ij} and $a_{i'j}$ do not have the same b ;

For $i \neq i'$, a_{ij} and $a_{i'j}$ do not have the same c ;

For $j \neq j'$, a_{ij} and $a_{ij'}$ do not have the same b ;

For $j \neq j'$, a_{ij} and $a_{i'j'}$ do not have the same c .

There are various ways in which a linear order can be imposed on the elements of A so as to reduce this formulation to the general one stated above.

One way of constructing a set A is to build it up element by element. Suppose that a_1, \dots, a_{k-1} have been chosen. The given restrictions will then require that a_k belong to some subset S_k of U . If S_k is not empty an a_k can be chosen and the building-up process continued; if S_k is empty one

¹ Most of this work was done on SWAC at the University of California at Los Angeles, and was sponsored by the Office of Naval Research.

² Cf. *Contributions to geometry*, Herbert Ellsworth Slaught Memorial Paper No. 4, Amer. Math. Monthly vol. 62 (1955) p. 21.

must back-track and change one of the previous a 's. To do this systematically we shall assume that the elements of U have been linearly ordered, and shall always choose a_k to be the least element of S_k . If S_k is empty we return to S_{k-1} and change a_{k-1} to the next larger element of S_{k-1} ; if this is impossible we back-track still further. The following condensed program contains the essential features of this process as it would be done by an automatic calculator. It is to be assumed that Step $s + 1$ will follow Step s unless otherwise specified.

General program.

1. Set $k = 1$.
2. Set $S_k = S_1$.
3. If S_k is empty jump to 9.
4. Set a_k equal to the smallest element of S_k .
5. If $k = n$ jump to 14.
6. Replace k by $k + 1$.
7. Compute S_k .
8. Jump to 3.
9. If $k = 1$, stop.
10. Replace k by $k - 1$.
11. Compute S_k .
12. Replace S_k by $S_k \cap \{a | a > a_k\}$.
13. Jump to 3.
14. Record $A = \{a_1, \dots, a_n\}$.
15. Jump to 12.

This program is set up to record all possible sets A . If, for example, they are merely to be counted, Step 14 may be modified accordingly. Other modifications might involve the use of other criteria than a fixed value of n for determining when a set A is completed (Step 5), or other criteria for stopping the process (Step 9).

2. Manipulation of sets. The speed of a program of this type is largely dependent on the method used to construct and store the set S_k in Step 7, the method of storage being important in Step 4. In problems like the examples given above this can be done quite efficiently.

To illustrate the method consider Example (i). We define $S_1 = U$, and, for $k = 2, \dots, n$, $S_k = \bar{A}_k \cap \bar{B}_k \cap \bar{C}_k$, where \bar{A}_k , \bar{B}_k , and \bar{C}_k are the complements, with respect to U , of the sets

$$\begin{aligned} A_k &= \{a_i | i < k\}, \\ B_k &= \{a_i - i + k | i < k\} \cap U, \\ C_k &= \{a_i + i - k | i < k\} \cap U. \end{aligned}$$

The essential point is that the effect on any of these last sets of increasing k by 1 is very simple. For example, B_{k+1} is obtainable from B_k by enlarging B_k by the element a_k , adding 1 to each element, and intersecting with U .

If elements of sets are represented by 1's in appropriate positions in a binary word (the "dual" representation) all these operations, and also the determining of the least element of a non-empty set, can be carried out easily and rapidly.

Example (ii) can be treated similarly, for various linear orderings, with suitable, but not serious, complications.

One can also compute the sets S_k in Step 11 by a similar process, passing from S_k to S_{k-1} . However this can be avoided by saving the sets S_k as they are built up in Step 7. The only objection to this might be a shortage of storage space. This was the case with the SWAC solution of Example (ii) for $N = 8$, since SWAC has a small high speed memory, but with most machines the recomputation of S_k would not be necessary.

3. Some applications.

(i) An incredible amount of hand computation has been done on the generalized Queens Problem. Kraitchik (*La Mathématiques des Jeux*, p. 316) gives the figures in Table I for the number N of solutions for $n = 1, \dots, 12$. These were all checked by SWAC, and the additional value for $n = 13$ was obtained and checked in about two hours.

Kraitchik also gives the number of symmetric solutions up to $n = 15$ (Table II). SWAC checked these up to $n = 12$, disagreed (on two separate programmings) at $n = 13, 14, 15$, and gave the indicated results (not checked) up to $n = 19$.

TABLE I
Total Solutions

n	N
1	1
2	0
3	0
4	2
5	10
6	4
7	40
8	92
9	352
10	724
11	2680
12	14200
13	73712

TABLE II
Symmetric Solutions

n	Kraitchik	SWAC
1	1	1
2	0	0
3	0	0
4	2	2
5	2	2
6	4	4
7	8	8
8	4	4
9	16	16
10	12	12
11	48	48
12	80	80
13	132	136
14	412	420
15	1192	1240
16		2872
17		7652
18		18104
19		50184

(ii) As a preliminary to attacking the 10×10 orthogonal latin squares the case $N = 8$ was tried, with discouraging results.

a. Ordering the elements lexicographically by rows and columns no solutions were obtained in an hour's run. In fact, no elements in the seventh row ($k > 48$) were ever obtained, nor were the first three rows ever changed from their initial values. About 95% of the time was spent changing the fifth and sixth rows.

b. The program was changed so as to enumerate by rows for the first h rows and then by columns. The best value of h seemed to be 4; this procedure was about 14 times as fast as the first one. No solutions were obtained.

c. Rough and optimistic extrapolation from the data obtained gives an estimate of the order of three million hours of computation for the whole problem.

(iii) This method was used in the enumerative part of the proof of the *Uniqueness of the projective plane of order eight* (M. Hall, J. D. Swift, R. J. Walker, MTAC vol. 10 (1956) pp. 186–194).

CORNELL UNIVERSITY,
ITHACA, NEW YORK

THE CYCLOTOMIC NUMBERS OF ORDER TEN¹

BY

ALBERT LEON WHITEMAN

1. Introduction. A central problem in the theory of cyclotomy is to obtain precise formulas for the numbers $(h,k) = (h,k)_0$ defined in terms of a prime $p = ef + 1$ and a primitive root g of p as the number of solutions s, t of the trinomial congruence

$$(1.1) \quad g^{es+h} + 1 \equiv g^{et+k} \pmod{p},$$

where the values of s and t are each selected from the integers $0, 1, \dots, f - 1$. In volume VI of these Proceedings R. H. Bruck [2] has given an interesting account of the current status of the problem for the specific values $e = 3, 4, 5, 6, 8, 9, 10, 12, 16, 20$. The purpose of the present paper is to give the first complete solution in the case $e = 10$.

The essential groundwork for this solution has already been laid by Dickson in the first [3] of a series of three memoirs. In the first place, Dickson showed that if p is a prime $\equiv 1 \pmod{5}$, then there are exactly four integral simultaneous solutions of the pair of diophantine equations

$$16p = x^2 + 50u^2 + 50v^2 + 125w^2, \quad xw = v^2 - 4uv - u^2,$$

with x uniquely determined by the condition $x \equiv 1 \pmod{5}$. The four solutions are given by (x, u, v, w) , $(x, -u, -v, w)$, $(x, v, -u, -w)$, $(x, -v, u, -w)$. Secondly, he expressed the numbers $(0,0)_0$ and $(0,5)_0$ as linear combinations with integral coefficients of p, x, u, v and w . There are ten sets of formulas depending on the parity of f and the quintic residue character of 2 modulo p . For example, when 2 is a quintic residue of $p = 10f + 1$, f even, the formulas are

$$100(0,0)_0 = p - 29 + 18x, \quad 100(0,5)_0 = p - 9 - 2x.$$

In [2] Bruck showed that all of the numbers $(h,k)_0$ can be expressed in terms of p, x, u, v, w . He also set up the problem of calculating the precise formulas in a form which could be handled by computing machines. The method of the present paper is along completely different lines (compare [12]). It is best suited for hand computation and differs from Dickson's method in many significant details. The principal new tools are Theorem 1 in §3 and Theorem 3 in §5. Complete tables of formulas for the cyclotomic numbers of order ten are given in §7.

The problem of determining the number of solutions of the congruence

$$(1.2) \quad ax^e + by^e \equiv c \pmod{p} \quad (xy \not\equiv 0 \pmod{p})$$

¹ The work presented in this paper was sponsored (in part) by the National Science Foundation under contract NSF G2791 with the University of Southern California.

for each integer c can be reduced to the problem of determining the number of solutions of (1.1). It is natural to take $a = 1$, $b = -1$ and ask for what values of p is it true that the number of solutions of (1.2) is the same for every $c \not\equiv 0 \pmod{p}$. This is the problem of residue difference sets and it is solved in §8 in the case $e = 10$.

2. Cyclotomy. In this section some results from the theory of cyclotomy are presented for convenient reference. Let p be an odd prime and g a fixed primitive root of p . Let e be a divisor of $p - 1$ and write $p - 1 = ef$. The cyclotomic numbers (h,k) are periodic in both h and $k \pmod{e}$. They also have the following well known properties [1, pp. 202–203]:

$$(2.1) \quad (h,k) = (e - h, k - h);$$

$$(2.2) \quad (h,k) = \begin{cases} (k,h) & (f \text{ even}), \\ \left(k + \frac{1}{2}e, h + \frac{1}{2}e \right) & (f \text{ odd}). \end{cases}$$

Let $\beta = \exp(2\pi i/e)$ be a primitive e th root of unity. For an integer a not divisible by p let $\text{ind } a$ be defined by means of the congruence $g^{\text{ind } a} \equiv a \pmod{p}$. In the theory of cyclotomy the so-called Jacobi sum [1, p. 122] plays a fundamental role. For each pair of integers m,n this sum is defined by

$$(2.3) \quad \psi(\beta^m, \beta^n) = \sum_{a+b \equiv 1 \pmod{p}} \beta^{m \text{ ind } a + n \text{ ind } b},$$

where a,b runs over all pairs of integers in the range $1 \leq a, b \leq p - 1$ satisfying the summation condition. It follows from the definitions that

$$(2.4) \quad \psi(\beta^m, \beta^n) = \psi(\beta^n, \beta^m) = (-1)^{nf} \psi(\beta^{-m-n}, \beta^n).$$

The most important property of the function $\psi(\beta^m, \beta^n)$ is the formula [1, p. 123]

$$(2.5) \quad \psi(\beta^m, \beta^n) \psi(\beta^{-m}, \beta^{-n}) = p,$$

provided no one of the integers $m, n, m + n$ is divisible by e .

Clearly $\psi(\beta^m, \beta^n)$ is periodic in both m and n with respect to the modulus e . In terms of the cyclotomic numbers it may be expanded into the double finite Fourier series [9, p. 682]

$$(2.6) \quad \psi(\beta^m, \beta^n) = (-1)^{mf} \sum_{h,k=0}^{e-1} (h,k) \beta^{mh+nk}.$$

The expansion (2.6) may also be written in its inverted form

$$(2.7) \quad e^2(h,k) = \sum_{m,n=0}^{e-1} (-1)^{mf} \psi(\beta^m, \beta^n) \beta^{-mh-nk}.$$

In (2.6) replace m by vn , where v is an integer. Collecting the exponents

of β which are in the same residue class $(\bmod e)$ we get an alternative form [11, p. 366] of the finite Fourier series expansion

$$(2.8) \quad \psi(\beta^{vn}, \beta^n) = (-1)^{vnf} \sum_{i=0}^{e-1} B(i, v) \beta^{ni},$$

where the Fourier coefficients $B(i, v)$ are Dickson-Hurwitz sums [10, p. 90] defined by

$$(2.9) \quad B(i, v) = \sum_{h=0}^{e-1} (h, i - vh).$$

Putting $n = 0$ in (2.8) we obtain the special case

$$(2.10) \quad \sum_{i=0}^{e-1} B(i, v) = p - 2.$$

The inverted form of (2.8) is

$$(2.11) \quad eB(i, v) = \sum_{n=0}^{e-1} (-1)^{vnf} \psi(\beta^{vn}, \beta^n) \beta^{-ni}.$$

It follows from (2.3) and (2.11) that

$$(2.12) \quad B(i, v) = B(i, e - v - 1),$$

and in particular that

$$(2.13) \quad B(i, 0) = \begin{cases} f - 1 & (i = 0), \\ f & (1 \leq i \leq e - 1). \end{cases}$$

Of course (2.10), (2.12) and (2.13) may also be deduced directly from (2.9) in conjunction with (2.1) and (2.2).

There is also a finite Parseval relation [10, p. 79] associated with (2.8). For a fixed integer k this is given by

$$(2.14) \quad e \sum_{i=0}^{e-1} B(i, v) B(i + k, v) = \sum_{j=0}^{e-1} |\psi(\beta^{jv}, \beta^j)|^2 \beta^{jk}.$$

We next derive a lemma which will be useful in later sections.

LEMMA 1. *If v is relatively prime to e , then*

$$(2.15) \quad B(i, v) = B(\bar{v}i, \bar{v}),$$

where \bar{v} is any solution of the congruence $v\bar{v} \equiv 1 (\bmod e)$.

To prove (2.15) we replace h by $\bar{v}(i - h)$ in the right member of (2.9). This is permissible since $\bar{v}(i - h)$ runs over a complete residue system $(\bmod e)$ whenever h does. Making an application of (2.2) we find that the right member of (2.9) reduces to the sum for $B(\bar{v}i, \bar{v})$. This completes the proof of the lemma.

Let now α denote a root of the equation $\alpha^{p-1} = 1$ and put $\zeta = \exp(2\pi i/p)$. Closely related to the Jacobi sum $\psi(\beta^m, \beta^n)$ is the resolvent of Lagrange [1, p. 83]

$$(2.16) \quad \tau(\alpha) = \sum_{a=1}^{p-1} \alpha^{\text{ind } a} \zeta^a.$$

Indeed we have the formula [1, p. 86]

$$(2.17) \quad \psi(\beta^m, \beta^n) = \tau(\beta^m)\tau(\beta^n)/\tau(\beta^{m+n})$$

when $m + n$ is not divisible by e . A deeper property of (2.16) is given by the identity

$$(2.18) \quad \tau(-1)\tau(\alpha^2) = \alpha^{2m}\tau(\alpha)\tau(-\alpha) \quad (g^m \equiv 2 \pmod{p}).$$

The last result was stated without proof by Jacobi [6]. In [3, p. 407] Dickson gives a proof attributed to H. H. Mitchell. Another proof is given by Hasse in [5, p. 442].

In the rest of this section we assume that e is even and write $e = 2E$. We also define the functions

$$(2.19) \quad s(h, k) = (h, k) - (h, k + E), \quad t(h, k) = (h, k) - (h + E, k).$$

It follows at once from (2.2) that

$$(2.20) \quad t(h, k) = \begin{cases} s(k, h) & (f \text{ even}), \\ s(k + E, h + E) & (f \text{ odd}). \end{cases}$$

We shall employ two additional lemmas.

LEMMA 2. *If e is even and $E = e/2$, then*

$$(2.21) \quad 4(h, k)_e = (h, k)_E + s(h, k) + s(h + E, k) + 2t(h, k).$$

This lemma is an immediate consequence of the formula $(h, k)_E = (h, k) + (h + E, k) + (h, k + E) + (h + E, k + E)$. For details of the proof see [12]. Lemma 2 will be used in §6 to deduce the value of $(h, k)_{10}$ from the value of $(h, k)_5$.

LEMMA 3. *If $e = 2E$ and E is odd, then the Jacobi sums $\psi(\beta^m, \beta^n)$ satisfy the relations*

$$(2.22) \quad \psi(\beta, \beta) = (-1)^f \beta^{(e-2)m} \psi(\beta^{E-1}, \beta),$$

$$(2.23) \quad \psi(\beta^{(E-1)/2}, \beta) = (-1)^{(E-1)f/2} \beta^{(E+1)m} \psi(\beta^{E-1}, \beta),$$

$$(2.24) \quad \psi(\beta^{E-1}, \beta^{E-1}) = \beta^{2m} \psi(\beta^{E-1}, \beta),$$

where the integer m is defined by the congruence $g^m \equiv 2 \pmod{p}$.

The hypothesis that E is odd is not required in the derivation of (2.22). To prove the lemma we make three applications of (2.18) with $\alpha = \beta^{E-1}$, $\beta^{(E-1)/2}$, β . In the resulting three equations we multiply both members by

$\tau(\beta)/\tau(\beta^E)\tau(\beta^{e-1})$, $1/\tau(\beta^{e-1})$, $\tau(\beta^{E-1})/\tau(\beta^E)\tau(\beta^{E+1})$ respectively. Then using (2.17) and (2.4) we get (2.22), (2.23) and (2.24).

It will be seen in §6 that Lemma 3 serves the purpose of dividing the primes into classes depending on the residue character of 2 modulo p . For each class of primes there is a separate set of formulas of the cyclotomic constants.

3. Cyclotomy when e is an odd prime. Our first objective in this section is to bring the right member of (2.7) into a form which is useful for calculating (h,k) . To simplify matters we shall take e to be an odd prime. The number f is then of necessity even. In (2.7) we separate the terms in which $n = 0$ from the terms in which $n \neq 0$. For a fixed nonzero value of n the numbers vn run over a complete residue system $(\bmod e)$ whenever v does. Thus we obtain

$$(3.1) \quad e^2(h,k) = \sum_{m=0}^{e-1} \psi(\beta^m, \beta^0) \beta^{-mh} + \sum_{v=0, n=1}^{e-1} \psi(\beta^{vn}, \beta^n) \beta^{-vh - nk}.$$

By (2.3) $\psi(\beta^m, \beta^0) = p - 2$ or -1 according as $m = 0$ or $m \neq 0$. The first sum in (3.1) therefore reduces to $p - e - 1$ or $p - 1$ according as $h = 0$ or $h \neq 0$. To evaluate the second sum in (3.1) we introduce into it the expansion (2.8) of $\psi(\beta^{vn}, \beta^n)$. After inverting the orders of summation we sum over $i, v = 0, 1, \dots, e - 1$ and separate the pairs i, v into two sets S and T according as $i - vh - k \equiv 0 \pmod{e}$ or not. The second sum in (3.1) thus becomes

$$(3.2) \quad \begin{aligned} \sum_{i, v \in S} B(i, v) \sum_{n=1}^{e-1} \beta^{n(i-vh-k)} &+ \sum_{i, v \in T} B(i, v) \sum_{n=1}^{e-1} \beta^{n(i-vh-k)} \\ &= e \sum_{v=0}^{e-1} B(vh + k, v) - \sum_{i, v=0}^{e-1} B(i, v). \end{aligned}$$

In (2.10) sum over $v = 0, 1, \dots, e - 1$. We deduce that the second sum in the right member of (3.2) is equal to $e(p - 2)$. Combining the results in this paragraph we have the following theorem.

THEOREM 1. *Let e be an odd prime. Then*

$$(3.3) \quad e^2(h, k) = (p - 1)(1 - e) + \epsilon + e \sum_{v=0}^{e-1} B(vh + k, v),$$

where $\epsilon = 0$ if $e|h$ and $\epsilon = e$ if $e \nmid h$.

We next evaluate the right member of (2.14). The result is an orthogonality relation for the sums $B(i, v)$. We require the decomposition formula (2.5). When v is equal to 0 or $e - 1$ the value of $B(i, v)$ is given in (2.12) and (2.13). By (2.3) the complex conjugate of $\psi(\beta^m, \beta^n)$ is $\psi(\beta^{-m}, \beta^{-n})$. When $1 \leq v \leq e - 2$ we deduce from (2.3) and (2.5) that $|\psi(\beta^{jv}, \beta^j)|^2 = (p - 2)^2$ or p according as $j = 0$ or $1 \leq j \leq e - 1$. We thus obtain the following theorem.

THEOREM 2. *Let e be an odd prime and let $1 \leq v \leq e - 2$. Then*

$$(3.4) \quad \sum_{i=0}^{e-1} B(i, v) B(i + k, v) = f(p - 4) + \delta,$$

where $\delta = p$ if $e|k$ and $\delta = 0$ if $e\nmid k$.

Theorems 1 and 2 supply the machinery for calculating the formulas for $(h, k)_5$. This is accomplished in the next section.

4. The quintic case. The identity of Dickson stated in the introduction is a consequence of the orthogonality relations (3.4). We take $e = 5$, $v = 1$ and for brevity put

$$(4.1) \quad b_i = B(i, 1).$$

Making three applications of (3.4) with $k = 0, 1$ and 3 we get

$$(4.2) \quad \begin{aligned} p &= b_0^2 + b_1^2 + b_2^2 + b_3^2 + b_4^2 - B, \\ B &= b_0 b_1 + b_1 b_2 + b_2 b_3 + b_3 b_4 + b_4 b_0 \\ &\quad = b_0 b_3 + b_1 b_4 + b_2 b_0 + b_3 b_1 + b_4 b_2. \end{aligned}$$

The identity (4.2) may now be written in the equivalent form

$$(4.3) \quad 16p = x^2 + 50u^2 + 50v^2 + 125w^2, \quad xw = v^2 - 4uv - u^2.$$

A straight-forward computation shows that the numbers x, u, v, w in (4.3) satisfy the linear relations

$$(4.4) \quad \begin{aligned} -x &= b_1 + b_2 + b_3 + b_4 - 4b_0, \\ 5u &= b_1 + 2b_2 - 2b_3 - b_4, \\ 5v &= 2b_1 - b_2 + b_3 - 2b_4, \\ 5w &= b_1 - b_2 - b_3 + b_4. \end{aligned}$$

In view of (2.10) the first formula in (4.4) becomes

$$(4.5) \quad -x = p - 2 - 5b_0 \quad (x \equiv 1 \pmod{5}),$$

which is in keeping with Dickson's choice in the selection of the sign of x . On the other hand the numbers u, v, w in (4.3) are uniquely determined by (4.4) except for sign. This ambiguity arises because the numbers (h, k) are themselves indeterminate and depend on the choice of the primitive root g . In order to control the signs of u, v, w we transform the formulas (4.4) into the equivalent set

$$(4.6) \quad \begin{aligned} 3x &= -p + 14 + 25(0, 0), \\ u &= (0, 2) - (0, 3), \quad v = (0, 1) - (0, 4), \quad w = (1, 3) - (1, 2). \end{aligned}$$

In the derivation of the formulas (4.6) we express each of the numbers b_i in terms of the numbers (h, k) with the aid of (4.1), (2.9), (2.1) and (2.2). We

now agree to choose the numbers x, u, v, w so that the formulas (4.6) are satisfied.

The equations in (4.4) together with the equation $p - 2 = b_0 + b_1 + b_2 + b_3 + b_4$ of (2.10) constitute five independent linear relations involving b_0, b_1, b_2, b_3, b_4 . Consequently the numbers $b_i = B(i,1)$ are uniquely determined by x, u, v, w . Solving this system of linear equations we get

$$(4.7) \quad \begin{aligned} 5B(0,1) &= p - 2 + x, \\ 20B(1,1) &= 4p - 8 - x + 10u + 20v + 25w, \\ 20B(2,1) &= 4p - 8 - x + 20u - 10v - 25w, \\ 20B(3,1) &= 4p - 8 - x - 20u + 10v - 25w, \\ 20B(4,1) &= 4p - 8 - x - 10u - 20v + 25w. \end{aligned}$$

We next utilize Theorem 1 to yield formulas which express the numbers $(h,k)_5$ in terms of the numbers $B(i,1)$. We first note that the value of $B(i,0) = B(i,4)$ is given in (2.13). Again by (2.12) we have $B(i,3) = B(i,1)$. Finally by Lemma 1 we have $B(i,2) = B(3i,1)$. Using (3.3) we obtain the formulas

$$(4.8) \quad \begin{aligned} 25(0,0) &= -2p - 8 + 15B(0,1), \\ 100(0,1) &= -8p + 8 + 40B(1,1) + 20B(3,1), \\ 100(0,2) &= -8p + 8 + 20B(1,1) + 40B(2,1), \\ 100(0,3) &= -8p + 8 + 40B(3,1) + 20B(4,1), \\ 100(0,4) &= -8p + 8 + 20B(2,1) + 40B(4,1), \\ 100(1,2) &= -8p + 28 + 20B(0,1) + 20B(2,1) + 20B(3,1), \\ 100(1,3) &= -8p + 28 + 20B(0,1) + 20B(1,1) + 20B(4,1). \end{aligned}$$

Substituting from (4.7) into (4.8) we now get

$$(4.9) \quad \begin{aligned} 25(0,0) &= p - 14 + 3x, \\ 100(0,1) &= 4p - 16 - 3x + 50v + 25w, \\ 100(0,2) &= 4p - 16 - 3x + 50u - 25w, \\ 100(0,3) &= 4p - 16 - 3x - 50u - 25w, \\ 100(0,4) &= 4p - 16 - 3x - 50v + 25w, \\ 100(1,2) &= 4p + 4 + 2x - 50w, \\ 100(1,3) &= 4p + 4 + 2x + 50w. \end{aligned}$$

By means of (2.1) and (2.2) the twenty-five constants $(h,k)_5$ with $h, k \equiv 0, 1, \dots, 4 \pmod{5}$ can be expressed in terms of the seven constants appearing in (4.9). These relations are exhibited in the following table in which (h,k) is in row h and column k .

	00	01	02	03	04
01	04	12	13	12	
02	12	03	13	13	
03	13	13	02	12	
04	12	13	12	01	

Formulas (4.9) together with the table in (4.10) constitute a complete solution of the cyclotomic problem in the case $e = 5$.

5. **Cyclotomy when e is twice an odd prime.** The problem of expressing the numbers $(h,k)_10$ linearly in terms of p, x, u, v, w can be solved by extending the technique used in the quintic case. However, the procedure turns out to be unwieldy. We prefer instead to follow another method (compare [12, Lemma 2]). Our major tool is Theorem 3 and its corollary.

In this section we assume that e is even and put $E = e/2$. We first find it convenient to introduce the function $S(h,k) = S_1(h,k) + S_2(h,k)$, where

$$(5.1) \quad S_1(h,k) = (-1)^k(B(h,E) - B(h+E,E)) \\ + (-1)^{h+k}(B(-h,E) - B(-h+E,E)),$$

and

$$(5.2) \quad S_2(h,k) = \begin{cases} (-1)^k 2 & (h \text{ even}), \\ 0 & (h \text{ odd}). \end{cases}$$

In the application of Lemma 2 the value of $s(h,k)$ is needed. This is furnished by the following theorem.

THEOREM 3. *Let $e = 2E$, where E is an odd prime. If h and k are arbitrary integers, then*

$$(5.3) \quad es(h,k) = S(h,k) + \sum_{v=0}^{e-1} (B(k+hv,v) - B(k+hv+E,v)).$$

The starting point of the proof is the relation

$$(5.4) \quad B(k+hv,v) = (h,k) + \sum_{i=1}^{e-1} (i+h, k-vi)$$

which follows from (2.9). In (5.4) sum over $v = 0, 1, \dots, e-1$ and then replace k by $k+E$. Using (2.19) we obtain after subtraction

$$(5.5) \quad es(h,k) = S + \sum_{v=0}^{e-1} (B(k+hv,v) - B(k+hv+E,v)),$$

where we have put

$$(5.6) \quad S = \sum_{i=1}^{e-1} \sum_{v=0}^{e-1} ((i+h, k+E-v) - (i+h, k-vi)).$$

Comparing (5.3) and (5.5) we see that the remainder of the proof consists in showing that $S = S(h,k)$. Consider first a fixed value of i in the outer sum in (5.6). When i is odd and different from E , the numbers vi run over a complete residue system $(\bmod e)$ whenever v does. But when i is equal to E , the terms of the sequence vi $(\bmod e)$ are alternately 0 and E . In either event the corresponding contribution of the right member of (5.6) to the value of S is zero. Note that essential use has been made of the hypothesis that E is an odd prime. Finally, suppose that i is even and different from zero. Then the least positive residues of the numbers vi $(\bmod e)$ run twice

over the even integers from 0 to $e - 2$ in some order. Taking into account the fact that $i = 0$ is excluded from the outer sum in (5.6) we may now write $S = S_1 + S_2$, where

$$(5.7) \quad S_1 = 2 \sum_{m=0}^{E-1} ((h, k + 2m) - (h, k + 2m + 1)),$$

and

$$(5.8) \quad S_2 = 2 \sum_{i \text{ even}} \sum_{m=0}^{E-1} ((i + h, k + 2m + 1) - (i + h, k + 2m)).$$

In the outer sum of (5.8) i runs over all even integers in the interval $0 \leq i \leq e - 2$.

To complete the proof we shall show that (i) $S_1 = S_1(h, k)$ and (ii) $S_2 = S_2(h, k)$. In the definition (2.9) of the sum $B(h, v)$ take $v = E$ and separate the even values of the summation index from the odd. Then using (2.2) we get

$$(5.9) \quad B(h, E) = \sum_{i=0}^{E-1} (h, 2i) + \sum_{i=0}^{E-1} (h + E, 2i + 1).$$

We may also write down three formulas corresponding to (5.9) with h replaced by $h + E$, $-h$ and $-h + E$. By (2.1) the first sum in (5.9) is equal to $\sum_{i=0}^{E-1} (-h, 2i + h)$, and the second sum is equal to

$$\sum_{i=0}^{E-1} (-h, 2i + 1 + h).$$

The result stated in (i) may now be verified by substituting from (5.9) into each term in the right member of (5.1). There are four cases to consider depending upon the parity of h and k .

The proof of (ii) is based on the formulas

$$(5.10) \quad \sum_{i \text{ even}} B(i, v) = \frac{p - 3}{2}, \quad \sum_{i \text{ odd}} B(i, v) = \frac{p - 1}{2},$$

in which i ranges over integers from 0 to $e - 1$ satisfying the summation condition. The first formula in (5.10) may be proved by summing over even i in both members of (2.11). Alternatively, it may be deduced in the following way from (2.9). In view of (2.12) there is no loss of generality in assuming that v is even. Summing over even values of i in (2.9) and interchanging the signs of summation we find that the first sum in (5.10) reduces to $\sum_{i=0}^{E-1} B(2i, 0)$ and is therefore equal to $(p - 3)/2$ because of (2.13). The second formula in (5.10) can be proved in a similar way.

We are now in the position to evaluate the right member of (5.8). Again, there are four cases depending on the parity of h and k . Employing (5.9) we may combine these four cases into the single formula

$$(5.11) \quad (-1)^k S_2 = (1 + (-1)^h) \sum_{i \text{ odd}} (B(i, E) - B(i + E, E))$$

where i runs over the odd integers in the interval from 1 to $e - 1$. Introducing (5.10) into (5.11) we find that S_2 is equal to the right member of (5.2). This completes the proof of Theorem 3.

For the application of Theorem 3 to Lemma 2 it will be convenient to combine $s(h, k)$ and $s(h + E, k)$. Accordingly we derive the following corollary of Theorem 3.

COROLLARY. *If the hypotheses of Theorem 3 are satisfied, then*

(5.12)

$$\begin{aligned} E(s(h, k) + s(h + E, k)) &= (-1)^k + (-1)^{h+k}(B(-h, E) - B(-h + E, E)) \\ &\quad + \sum_{v=0}^{E-1} (B(k + 2hv, 2v) - B(k + 2hv + E, 2v)). \end{aligned}$$

The sum of the first two terms in the right member of (5.12) is equal to $(S(h, k) + S(h + E, k))/2$. To establish (5.12) we replace h by $h + E$ in (5.3). This has the effect of multiplying the v th term in the sum in (5.3) by $(-1)^v$. Thus the sum of $Es(h, k)$ and $Es(h + E, k)$ reduces to the right member of (5.12).

6. The decimic case. Let $e = 2E$, where E is an odd prime. In computing $es(h, k)$ by means of Theorem 3 we require the values of the successive terms in the sum in (5.3). The particular case $e = 10$, $E = 5$ is considered in this section. We shall derive formulas for the values of $B(i, v) - B(i + 5, v)$, $i = k + hv$, $v = 0, 1, \dots, 9$.

To begin with we note that (2.13) may be written in the form

$$(6.1) \quad B(i, 0) - B(i + 5, 0) = \begin{cases} -1 & (i = 0), \\ 0 & (i = 1, 2, 3, 4). \end{cases}$$

We now deduce additional results from Lemma 3. Employing (2.22), (2.23) and the expansion (2.8) we deduce

$$(6.2) \quad \beta^{2m} \sum_{i=0}^4 a_{2i} \beta^{2i} = \beta^{4m} \sum_{i=0}^4 b_{2i} \beta^{2i} = \sum_{i=0}^4 c_{2i} \beta^{2i},$$

where we have put $a_i = B(i, 1) - B(i + 5, 1)$, $b_i = B(i, 2) - B(i + 5, 2)$, $c_i = B(i, 4) - B(i + 5, 4)$, $i = 0, 1, \dots, 9$. Clearly the coefficients c_i satisfy the relations

$$(6.3) \quad c_5 = -c_0, \quad c_6 = -c_1, \quad c_7 = -c_2, \quad c_8 = -c_3, \quad c_9 = -c_4.$$

It also follows from (5.10) that

$$(6.4) \quad \sum_{i=0}^4 a_{2i} = \sum_{i=0}^4 b_{2i} = \sum_{i=0}^4 c_{2i} = -1.$$

Using the relation $1 + \beta^2 + \beta^4 + \beta^6 + \beta^8 = 0$ we next eliminate the constant term from each of the sums in (6.2). Since β^2 is a primitive fifth root of unity the numbers $\beta^2, \beta^4, \beta^6, \beta^8$ are linearly independent over the field of rational numbers. We may therefore equate the coefficients of like powers of β in the resulting equations. We get thus

$$(6.5) \quad a_{2i} - a_0 = c_{2i+2m} - c_{2m}, \quad b_{2i} - b_0 = c_{2i+4m} - c_{4m},$$

where the indices are to be taken modulo 10. Summing over $i = 0, 1, 2, 3, 4$ in (6.5) and applying the side condition (6.4) we obtain the formulas

$$(6.6) \quad B(i,1) - B(i + 5, 1) = c_{i+2m},$$

$$(6.7) \quad B(i,2) - B(i + 5, 2) = c_{i+4m}.$$

The integer m satisfies the congruence $g^m \equiv 2 \pmod{p}$. Hence there are five sets of formulas depending upon the quintic residue character of 2 (\pmod{p}) .

Returning to Lemma 1 we take $v = 3, \bar{v} = 7$ with $e = 10$. By (2.12) we get $B(i,3) = B(7i,2)$. Combining this result with (6.7) we have

$$(6.8) \quad B(i,3) - B(i + 5, 3) = c_{7i+4m}.$$

For $0 \leq v \leq 3$ formulas (6.1), (6.6), (6.7), (6.8) express the difference $B(i,v) - B(i + 5, v)$ in terms of the difference $B(i,4) - B(i + 5, 4)$. For $5 \leq v \leq 9$ we use the formula $B(i,v) = B(i,9-v)$ which follows from (2.12). Thus Theorem 3 and the corollary express $10t(h,k)$ and $5(s(h,k) + s(h+5,k))$ as linear combinations of c_0, c_1, c_2, c_3, c_4 . It should be noted that in some instances the calculation of $t(h,k)$ is especially simple. For it follows from (2.2), (2.19) and (2.20) that $t(h,k) = 0$ when f is even and $k = 5$ or when f is odd and $k = 0$.

To serve as an example of the preceding remarks we take $h = 4, k = 3, m \equiv 1 \pmod{5}, f$ odd. Then we have

$$(6.9) \quad \begin{aligned} 10t(4,3) &= -2 - 2c_0 + c_1 - c_2 + 2c_3 - 4c_4, \\ 5(s(4,3) + s(9,3)) &= -1 - c_0 + c_1 - c_2 + c_3 - c_4. \end{aligned}$$

We now give the details of the computation of (6.9). By (2.20) we have $t(4,3) = s(8,9)$. The ten consecutive terms of the sum in (5.3) for $h = 8, k = 9$ are given by $0, -c_4, -c_4, -c_0, c_1, -c_4, c_3, -c_4, -c_0, 0$. Also by (5.1) and (5.2) we get $S_1(8,9) = c_3 - c_2$ and $S_2(8,9) = -2$. The first formula in (6.9) now follows at once. Again, the five consecutive terms of the sum in (5.12) for $h = 4, k = 3$ are given by $0, -c_0, -c_4, c_3, -c_2$. The remaining portion of the right member of (5.12) is equal to $-1 + c_1$. This establishes the second formula in (6.9).

We next prove that each of the numbers c_0, c_1, c_2, c_3, c_4 is a linear combination of the numbers x, u, v, w . Returning to the expansion (2.8) we

replace β by $\beta^2 = \exp(2\pi i/5)$ and take $v = 1$, $n = 2$. We may now write $\psi(\beta^4, \beta^4) = \sum_{i=0}^4 B_5(i,1)\beta^{4i}$, where $B_5(i,1)$ is the Dickson-Hurwitz sum (2.9) corresponding to $e = 5$. It follows from (2.24) of Lemma 3 that

$$(6.10) \quad \sum_{i=1}^4 (B_5(i,1) - B_5(0,1))\beta^{4i} = \beta^{2m} \sum_{i=1}^4 (c_{2i} - c_0)\beta^{2i}.$$

Equating coefficients of like powers of β in both members of (6.10) we obtain

$$(6.11) \quad B_5(i,1) - B_5(0,1) = c_{4i+8m} - c_{8m}.$$

Summing over $i = 0, 1, 2, 3, 4$ in (6.11) and applying the side conditions (2.10) and (6.4) we get

$$(6.12) \quad B_5(i,1) = c_{4i+8m} + 2f.$$

The five identities in (4.7) express each of the numbers $B_5(i,1)$ in terms of the numbers p, x, u, v, w . Substituting (6.12) into (4.7) we get the identities

$$(6.13) \quad \begin{aligned} 20c_{8m+0} &= -4 + x, \\ 20c_{8m+1} &= 4 + x + 10u + 20v - 25w, \\ 20c_{8m+2} &= -4 - x - 20u + 10v - 25w, \\ 20c_{8m+3} &= 4 + x - 20u + 10v + 25w, \\ 20c_{8m+4} &= -4 - x + 10u + 20v + 25w. \end{aligned}$$

Lemma 2 provides a technique for expressing each cyclotomic number $(h,k)_{10}$ as a linear combination of p, x, u, v, w . We now have on hand all the necessary machinery for making the computations. To illustrate the method we give the derivation of the formula

$$(6.14) \quad 200(4,3) = 2p + 2 + x + 25u + 25v - 50w,$$

which is valid for $m \equiv 1 \pmod{5}$, f odd. Using Lemma 2 we get $400(4,3)_{10} = 100(4,3)_5 + 100s(4,3) + 100s(9,3) + 200t(4,3)$. By (6.9) the sum of the last three terms is $-60 - 60c_0 + 40c_1 - 40c_2 + 60c_3 - 100c_4$. By (4.10) we get $(4,3)_5 = (1,2)_5$. Substituting from (4.9) and (6.13) and simplifying we get (6.14).

7. Tables of the cyclotomic constants of order ten. The 100 constants $(h,k)_{10}$ with $h, k \equiv 0, 1, \dots, 9 \pmod{10}$ have at most 22 different values for a given p . Tables A and B, which follow, summarize the relations between the constants. In these tables the entry in row h and column k is equal to (h,k) .

The author has employed the method described in §6 to calculate the values of the 100 constants $(h,k)_{10}$. These values are expressible in terms of p, x, u, v, w , where the sign of x is such that $x \equiv 1 \pmod{5}$. There are ten sets of formulas depending on the parity of f and the quintic residue character of 2 modulo p . The 22 essentially different formulas of each set are given in the accompanying four tables.

TABLE A. f even

	0	1	2	3	4	5	6	7	8	9
0	00	01	02	03	04	05	06	07	08	09
1	01	09	12	13	14	15	16	17	18	12
2	02	12	08	18	24	25	26	27	24	13
3	03	13	18	07	17	27	36	36	25	14
4	04	14	24	17	06	16	26	36	26	15
5	05	15	25	27	16	05	15	25	27	16
6	06	16	26	36	26	15	04	14	24	17
7	07	17	27	36	36	25	14	03	13	18
8	08	18	24	25	26	27	24	13	02	12
9	09	12	13	14	15	16	17	18	12	01

TABLE B. f odd

	0	1	2	3	4	5	6	7	8	9
0	00	01	02	03	04	05	06	07	08	09
1	10	11	12	13	14	06	04	14	18	19
2	20	21	22	23	18	07	14	03	13	23
3	22	31	31	20	19	08	18	13	02	12
4	11	21	31	21	10	09	19	23	12	01
5	00	10	20	22	11	00	10	20	22	11
6	10	09	19	23	12	01	11	21	31	21
7	20	19	08	18	13	02	12	22	31	31
8	22	23	18	07	14	03	13	23	20	21
9	11	12	13	14	06	04	14	18	19	10

When $m \equiv 2, 3$ or $4 \pmod{5}$ a table of values for (h,k) may be deduced from Table II corresponding to $m \equiv 1 \pmod{5}$. Thus to find (h,k) when $m \equiv 2 \pmod{5}$ look up $(3h, 3k)$ in Table II and then replace u, v, w by $-v, u, -w$ respectively. To find (h,k) when $m \equiv 3 \pmod{5}$ look up $(7h, 7k)$ in Table II and then replace u, v, w by $v, -u, -w$ respectively. Finally, to find (h,k) when $m \equiv 4 \pmod{5}$ look up $(-h, -k)$ in Table II and replace u, v by $-u, -v$.

TABLE Ia. $m \equiv 0 \pmod{5}$, f even

$$\begin{aligned}
100(0,0) &= p - 29 + 18x \\
100(0,1) &= p - 9 - 2x + 25u + 50v - 25w \\
100(0,2) &= p - 9 - 2x + 25v - 25w \\
100(0,3) &= p - 9 - 2x - 50u + 25v + 25w \\
100(0,4) &= p - 9 - 2x + 25u + 25w \\
100(0,5) &= p - 9 - 2x \\
100(0,6) &= p - 9 - 2x - 25u + 25w \\
100(0,7) &= p - 9 - 2x + 50u - 25v + 25w \\
100(0,8) &= p - 9 - 2x - 25v - 25w \\
100(0,9) &= p - 9 - 2x - 25u - 50v - 25w \\
200(1,2) &= 2p + 2 + x + 25w \\
200(1,3) &= 2p + 2 + x + 75w \\
200(1,4) &= 2p + 2 + x - 75w \\
200(1,5) &= 2p + 2 + x + 25w \\
200(1,6) &= 2p + 2 + x + 25w \\
200(1,7) &= 2p + 2 + x - 75w \\
200(1,8) &= 2p + 2 + x + 75w \\
200(2,4) &= 2p + 2 + x - 25w \\
200(2,5) &= 2p + 2 + x - 25w \\
200(2,6) &= 2p + 2 + x + 25w \\
200(2,7) &= 2p + 2 + x - 25w \\
200(3,6) &= 2p + 2 + x - 25w
\end{aligned}$$

The following application of the tables is of interest. Dickson [3, p. 398] has given a formula expressing the number $N_e(a,b)$ of solutions of the congruence $ax^e + by^e \equiv 1 \pmod{p}$ in terms of the cyclotomic numbers $(h,k)_e$. Hence Tables I and II may be used to compute corresponding tables for the numbers $N_{10}(a,b)$. We confine ourselves to a single illustrative example. In the special case $a = b = 1$ Dickson's formula states that the number of solutions prime to $p = ef + 1$ is $e^2(0,0)$ and the number of all solutions is

$2e + e^2(0,0)$. Using Table Ib we find, for instance, when 2 is a quintic residue of p , that the congruence

$$x^{10} + y^{10} \equiv 1 \pmod{p = 10f + 1} \quad (f \text{ odd}),$$

has exactly $p - 19 + 8x$ solutions prime to p , and $p + 1 + 8x$ solutions in all.

TABLE Ib. $m \equiv 0 \pmod{5}$, f odd

$100(0,0) = p - 19 + 8x$
$200(0,1) = 2p + 2 + x + 50u + 50v - 25w$
$200(0,2) = 2p + 2 + x - 50u + 50v - 75w$
$200(0,3) = 2p + 2 + x - 50u + 50v + 25w$
$200(0,4) = 2p + 2 + x + 50u + 50v + 75w$
$100(0,5) = p + 1 - 12x$
$200(0,6) = 2p + 2 + x - 50u - 50v + 75w$
$200(0,7) = 2p + 2 + x + 50u - 50v + 25w$
$200(0,8) = 2p + 2 + x + 50u - 50v - 75w$
$200(0,9) = 2p + 2 + x - 50u - 50v - 25w$
$100(1,0) = p - 9 - 2x + 25v$
$100(1,1) = p - 9 - 2x - 25v$
$200(1,2) = 2p + 2 + x + 25w$
$200(1,3) = 2p + 2 + x - 25w$
$200(1,4) = 2p + 2 + x - 75w$
$200(1,8) = 2p + 2 + x - 25w$
$200(1,9) = 2p + 2 + x + 25w$
$100(2,0) = p - 9 - 2x + 25u$
$200(2,1) = 2p + 2 + x - 75w$
$100(2,2) = p - 9 - 2x - 25u$
$200(2,3) = 2p + 2 + x + 75w$
$200(3,1) = 2p + 2 + x + 75w$

TABLE IIa. $m \equiv 1 \pmod{5}$, f even

$400(0,0) = 4p - 116 - 3x - 150u + 75w$
$100(0,1) = p - 9 - 2x + 50w$
$400(0,2) = 4p - 36 + 17x + 50u - 25w$
$200(0,3) = 2p - 18 - 4x + 25u - 25v + 25w$
$200(0,4) = 2p - 18 - 4x + 25u - 25v + 25w$
$400(0,5) = 4p - 36 + 17x + 50u - 25w$
$100(0,6) = p - 9 - 2x - 50w$
$400(0,7) = 4p - 36 + 17x + 50u - 25w$
$200(0,8) = 2p - 18 - 4x - 75u + 75v - 75w$
$200(0,9) = 2p - 18 - 4x + 25u - 25v + 25w$
$200(1,2) = 2p + 2 + x + 25u + 25v - 50w$
$200(1,3) = 2p + 2 + x - 50v - 75w$
$200(1,4) = 2p + 2 + x - 25u - 25v$
$200(1,5) = 2p + 2 + x + 50v + 25w$
$200(1,6) = 2p + 2 + x - 25u - 25v$
$200(1,7) = 2p + 2 + x + 25u + 25v - 50w$
$200(1,8) = 2p + 2 + x - 50u + 75w$
$200(2,4) = 2p + 2 + x + 50u + 75w$
$400(2,5) = 4p + 4 - 23x + 50u - 25w$
$200(2,6) = 2p + 2 + x - 25u - 25v$
$200(2,7) = 2p + 2 + x - 25u - 25v$
$200(3,6) = 2p + 2 + x + 50v + 25w$

TABLE IIb. $m \equiv 1 \pmod{5}$, f odd

400(0,0) =	$4p - 76 + 7x - 50u + 25w$
200(0,1) =	$2p + 2 + x + 50v + 125w$
400(0,2) =	$4p + 4 - 23x + 50u - 25w$
200(0,3) =	$2p + 2 + x + 25u - 75v + 50w$
200(0,4) =	$2p + 2 + x - 25u - 25v$
400(0,5) =	$4p + 4 + 27x + 150u - 75w$
200(0,6) =	$2p + 2 + x + 50v - 75w$
400(0,7) =	$4p + 4 - 23x + 50u - 25w$
200(0,8) =	$2p + 2 + x - 75u + 25v - 50w$
200(0,9) =	$2p + 2 + x - 25u - 25v$
100(1,0) =	$p - 9 - 2x$
200(1,1) =	$2p - 18 - 4x + 25u - 25v + 25w$
200(1,2) =	$2p + 2 + x - 25u - 25v$
200(1,3) =	$2p + 2 + x + 50u - 25w$
200(1,4) =	$2p + 2 + x - 25u - 25v$
200(1,8) =	$2p + 2 + x + 50v + 125w$
200(1,9) =	$2p + 2 + x + 25u + 25v - 50w$
400(2,0) =	$4p - 36 + 17x + 50u - 25w$
200(2,1) =	$2p + 2 + x + 25u + 25v - 50w$
200(2,2) =	$2p - 18 - 4x - 25u + 25v - 25w$
200(2,3) =	$2p + 2 + x - 50u - 25w$
200(3,1) =	$2p + 2 + x - 50v + 25w$

8. **Application to residue difference sets.** By a difference set of order k and multiplicity λ is meant a set of k distinct residues $r_1, r_2, \dots, r_k \pmod{v}$ such that the congruence $r_i - r_j \equiv d \pmod{v}$ has exactly λ solutions for each $d \not\equiv 0 \pmod{v}$. In [4] there is given a survey of all known difference sets. Residue difference sets are difference sets composed of e th power residues modulo a prime p . The results in §7 provide a useful tool for investigating the existence of difference sets composed of tenth power residues modulo p . We shall employ the following criterion due to E. Lehmer [8]. If e is even and $f = (p - 1)/e$ is odd, then a necessary and sufficient condition for the set of e th power residues modulo p to form a difference set is that $(i,0) = (f - 1)/e$, $i = 0, 1, \dots, (1/2)e - 1$, where $(f - 1)/e = \lambda$ is the multiplicity of the difference set.

For the application of this criterion we take $e = 10$, f odd. The values of the cyclotomic constants $(i,0)_10$ are tabulated in Tables Ib and IIb of §7. In the first place, if $m \equiv 0 \pmod{5}$ we have $100(0,0) = p - 19 + 8x$, and the condition $(0,0) = (p - 11)/100$ now implies that $x = 1$. It has been proved in [7] that if 2 is a quintic residue of p , x must be even. Hence we have arrived at a contradiction, and there is no difference set in this case. Secondly, if $m \equiv 1 \pmod{5}$ we have $100(1,0) = p - 9 - 2x$. The condition $(1,0) = (p - 11)/100$ again implies that $x = 1$. This time there is no contradiction at this point. Equating the values of $(0,0)$ and $(2,0)$ we obtain the relation $2u - w + 1 = 0$. Equating the values of $(3,0)$ and $(4,0)$ we get the equation $u - v + w = 0$. Therefore $v = 3u + 1$. Combining these results with the condition $xw = v^2 - 4uv - u^2$ in (4.3) we find at once that $u = 0$, $v = w = 1$, which yields $p = 11$. This is a trivial example since the only tenth power residue modulo 11 is $r = 1$ so that

$\lambda = 0$. There is no need to examine further cases. For if g' is any primitive root of p other than the fixed primitive root g , then $g' \equiv g^t \pmod{p}$ for some integer t prime to $p - 1$. The cases $m \equiv 2, 3$ or $4 \pmod{5}$ may therefore be transformed into the case $m \equiv 1 \pmod{5}$ by making an appropriate choice of g . We have completed the proof of the following theorem.

THEOREM 4. *The set of tenth power residues modulo a prime $p = 10f + 1$ cannot form a difference set.*

It is not necessary to give a separate proof of this theorem for the case f even. For it is known [8] that there exists no residue difference set for e odd, or for e even and f even. In the case $m \equiv 0 \pmod{5}$ Theorem 4 has previously been proved by E. Lehmer [8].

A modified residue difference set is one in which zero is counted as a residue. It is known [8] that such difference sets cannot exist for e odd or for e even and f even. Emma Lehmer [8] has proved that if e is even and $f = (p - 1)/e$ is odd, then a necessary and sufficient condition for the set of e th power residues and zero to be a difference set is that $1 + (0,0) = (i,0) = (f+1)/e, i = 1, 2, \dots, (1/2)e - 1$, where $(f+1)/e = \lambda$ is the multiplicity of the set. To apply this criterion we again take $e = 10$, f odd. Proceeding exactly as in the proof of Theorem 4 we may prove the following result.

THEOREM 4'. *The set of tenth power residues and zero modulo a prime $p = 10f + 1$ cannot form a difference set.*

Theorem 4 may be compared with a similar theorem [8] which states that there exists no difference set of sextic residues. On the other hand Marshall Hall [4] has shown that the set of residues with indices congruent to 0, 1 or 3 modulo 6 (for an appropriate choice of primitive root) constitute a difference set modulo a prime $p = 6f + 1$ of the form $4x^2 + 27$.

The method used to prove Theorem 4 may be extended so as to yield additional non-existence results. One such result is suggested by the difference set of order $k = 6$ and multiplicity $\lambda = 1$ composed of the residues $3^0, 3^{10}, 3^{20}, 3^3, 3^{13}, 3^{23} \pmod{31}$. In this example the number 3 is a primitive root of $p = 31$. The first three residues have indices $\equiv 0 \pmod{10}$, whereas the second three residues have indices $\equiv 3 \pmod{10}$. It is natural to ask if there are other primes $p = 10f + 1$, f odd, for which the set of residues congruent to 0 or 3 modulo 10 constitute a difference set. To answer this question we follow the procedure employed by Marshall Hall to establish his result.

It is convenient to say that a number x is in class i if $\text{ind } x \equiv i \pmod{10}$. Then for a fixed integer d in class k the number of solutions of the congruence $y - x \equiv d \pmod{p}$ with y in class j and x in class i is given by $(i - k, j - k)$. This enables us to tell how often each difference arises from sets composed of classes of given tenth power character. With a set whose

characters are 0,3 we find that for $y - x \equiv d \pmod{p}$ with d in class k the number N_k of solutions is

$$N_k = (-k, -k) + (3 - k, -k) + (-k, 3 - k) + (3 - k, 3 - k).$$

The values of N_k for $k = 5, 6, 7, 8, 9$ are the same as the values for $k = 0, 1, 2, 3, 4$ respectively. Formulas for the values of N_k , $k = 0, 1, 2, 3, 4$ may now be computed with the aid of Tables Ib and IIb in §7. We omit the actual formulas. Equating the five values of N_k in the case $m \equiv 0 \pmod{5}$ we find readily that $u = v = w = 0$ so that the decomposition (4.3) is impossible. Similarly, in the case $m \equiv 1 \pmod{5}$ we find that $x = 11$, $u = 2$, $v = 1$, $w = -1$ so that $p = 31$. The three remaining cases also yield $p = 31$. Hence the difference set corresponding to $p = 31$ is the only one of the prescribed type.

Altogether there are 1023 combinations of tenth power classes modulo 10. Whether any of these combinations when calculated as above lead to new difference sets is an open question. The amount of hand computation required to solve this problem appears prohibitive. Perhaps the best approach would involve an interplay of theoretical and machine methods.

BIBLIOGRAPHY

1. P. Bachmann, *Die Lehre von der Kreisteilung*, 2nd ed., B. G. Teubner, Leipzig and Berlin, 1921.
2. R. H. Bruck, *Computational aspects of certain combinatorial problems*, Proceedings of Symposia in Applied Mathematics, New York, McGraw-Hill Book Company, Inc., vol. 6, 1956, pp. 31–43.
3. L. E. Dickson, *Cyclotomy, higher congruences and Waring's problem*, Amer. J. Math. vol. 57 (1935) pp. 391–424.
4. Marshall Hall, Jr. *A survey of difference sets*, Proc. Amer. Math. Soc. vol. 7 (1956) pp. 975–986.
5. H. Hasse, *Vorlesungen über Zahlentheorie*, Berlin, Göttingen and Heidelberg, Springer, 1950.
6. C. G. Jacobi, *Ueber die Kreisteilung und ihre Anwendung auf die Zahlentheorie*, J. Reine Angew. Math. vol. 30 (1846) pp. 166–182; or *Gesammelte Werke* vol. 6 (1891) pp. 254–274.
7. Emma Lehmer, *The quintic character of 2 and 3*, Duke Math. J. vol. 18 (1951) pp. 11–18.
8. ———, *On residue difference sets*, Canad. J. Math. vol. 5 (1953) pp. 425–432.
9. H. S. Vandiver, *Quadratic relations involving the number of solutions of certain types of equations in a finite field*, Proc. Nat. Acad. Sci. U.S.A. vol. 35 (1949) pp. 681–685.
10. A. L. Whiteman, *Finite Fourier series and equations in finite fields*, Trans. Amer. Math. Soc. vol. 74 (1953) pp. 78–98.
11. ———, *The sixteenth power residue character of 2*, Canad. J. Math. vol. 6 (1954) pp. 364–373.
12. ———, *The cyclotomic numbers of order sixteen*, Trans. Amer. Math. Soc. vol. 86 (1957) pp. 401–413.

This page intentionally left blank

SOME RECENT APPLICATIONS OF THE THEORY OF LINEAR INEQUALITIES TO EXTREMAL COMBINATORIAL ANALYSIS

BY

ALAN J. HOFFMAN

1. Introduction. The purpose of this talk is to give an account of some aspects of recent research on the interplay between the theory of linear inequalities and a certain class of combinatorial problems. The kind of problem to be considered can be illustrated by the following example:

Let $A = (a_{ij})$ be a square incidence matrix of order v such that every row contains $k > 0$ ones and every column contains $k > 0$ ones. All other entries of A are 0. Mann and Ryser [1] have observed that A can then be expressed in the form

$$(1.1) \quad A = P_1 + \cdots + P_k,$$

where the P_i are permutation matrices. Obviously, an inductive argument will suffice to prove (1.1) if it can be shown that the hypotheses imply the existence of a permutation matrix $P = (p_{ij})$ such that

$$(1.2) \quad p_{ij} = 1 \quad \text{only if } a_{ij} = 1.$$

Mann and Ryser establish the existence of such a P by exploiting the Egerváry-Hall-König (see [2; 3; 4] and the discussions below) theorem on systems of distinct representatives. An alternative approach to (1.2) is to consider the convex set of all vectors with kv components $X = (\dots, x_{ij}, \dots)$ (where the subscript “ ij ” appears if and only if $a_{ij} = 1$), satisfying

$$(1.3) \quad \sum_i x_{ij} = \sum_j x_{ij} = 1, \quad x_{ij} \geq 0.$$

This convex set is not empty, since the hypotheses on A imply that setting $x_{ij} = 1/k$ satisfies (1.3). The set is also bounded, so it admits a vertex. The co-ordinates of the vertex may be obtained by solving a certain set of equations contained in the system of equations and inequalities (1.3), and it is easy to show ([5; 6] and many other places) that the determinant associated with this set of equations is ± 1 . Since the right hand side is integral, it follows from Cramer's rule that the co-ordinates of the vertex are integers. Conditions (1.3) imply that for each i , $x_{ij} = 0$ for all j with exactly one exception, for which $x_{ij} = 1$. Similarly, every column consists entirely of 0 entries except for exactly one entry which is 1. But this means that our vertex of (1.3) is the desired permutation P satisfying (1.2). (A slight generalization of the result is contained in [8].)

Observe that we have replaced a combinatorial argument—to wit, the

appeal to Egervary-Hall-König—by a quasi-geometric discussion involving polyhedra and vertices, to prove the combinatorial result (1.1). This suggests that invocation of concepts from the theory of linear inequalities may be useful in studying certain kinds of combinatorial situations. As a matter of fact, about a dozen mathematicians have chewed on this bone during the past five or six years, and this talk will summarize their findings.

2. Systems of representatives. The first result in this direction of using linear inequalities on combinatorial problems seems to be due to Rado [9], but his paper was regrettably overlooked by later workers until 1956. The more recent work began with the observation that the Hall theorem on systems of distinct representatives has itself an easy proof through the theory of linear inequalities. That theorem is:

Let R be a finite set with elements P_1, \dots, P_n ,

$$(2.1) \quad \begin{aligned} \mathcal{S} &= \{S_1, \dots, S_m\} \text{ a family of subsets of } R. \\ \text{A subset } S &= \{P_{t_1}, \dots, P_{t_m}\} \subset R \end{aligned}$$

of m distinct elements is called a system of representatives of \mathcal{S} if

$$(2.2) \quad P_{t_k} \in S_k, \quad k = 1, \dots, m.$$

In order that \mathcal{S} admit a system of distinct representatives, it is necessary and sufficient that, for any $I \subset \{1, \dots, m\}$,

$$(2.3) \quad \bar{I} \leq \overline{\bigcup_{i \in I} S_i}.$$

(Here and elsewhere \bar{M} = the number of elements in the set M .)

Proof by linear inequalities : Let C be the $n \times m$ matrix given by :

$$c_{ij} = \begin{cases} 1 & \text{if } P_i \in S_j, \\ 0 & \text{if } P_i \notin S_j \end{cases}$$

Consider the linear programming problem :

$$\text{minimize } \sum_{ij} c_{ij} x_{ij}, \text{ where } (x_{ij}) \text{ varies over all } n \text{ by } m$$

$$\text{matrices satisfying: } x_{ij} \geq 0, \quad \sum_i x_{ij} \leq 1, \quad \sum_j x_{ij} \leq 1.$$

It is easy to prove, using the unimodular property (see [5]) exploited in §1, that the maximum is m if and only if R contains a set S satisfying (2.2). But the maximum of the primal linear program equals the minimum in the dual program :

$$\text{minimize } \sum_i u_i + \sum_j v_j \quad u_i \geq 0, \quad v_j \geq 0, \quad u_i + v_j \geq c_{ij}.$$

Again by the unimodular property, it is sufficient to consider only the case where each u_i and each v_j is 0 or 1. So (2.2) holds for some S if and only if the smallest number of rows and columns collectively comprising all 1's in the matrix C is m , and it is easy to show that this condition is a consequence of (2.3). The necessity of (2.3) is, of course, trivial.

This method of proof, noted independently by several people (Motzkin [10], Kuhn [11], Hoffman [12], and probably others less vocal), while not as brief as the elegant induction of Halmos and Vaughan [13], has the redeeming feature that it fits the theorem into a larger context that enables us to know it better. We can now recognize it as a special case of the duality theorem of linear programming. Further, this recognition permits a facility in generalization.

One such direction was inspired by a result of Mann and Ryser (see [1; 14], also Hoffman and Kuhn [15; 16]). M. Tinsley and R. Rado have privately communicated alternative proofs of the main results of [1] and [15].

Let R and \mathcal{S} be as in (2.1). Let $\mathcal{T} = \{T_1, \dots, T_p\}$ be a partition of R ; i.e., $\bigcup T_i = R$, $T_i \cap T_j = \emptyset$ if $i \neq j$:

Let $c_k \leq d_k$ ($k = 1, \dots, m$) be given integers. In order that there exist a set $S \subset R$ satisfying (2.2) and

$$(2.4) \quad c_k \leq \overline{S \cap T_k} \leq d_k, \quad k = 1, \dots, m$$

it is necessary and sufficient that, for all $A \subset \{1, \dots, m\}$ and $B \subset \{1, \dots, p\}$,

$$(2.5) \quad \left(\overline{\bigcup_{i \in A} S_i} \right) \cap \left(\overline{\bigcup_{k \in B} T_k} \right) \geq \overline{A} - m + \sum_{k \in B} c_k$$

and

$$(2.6) \quad \left(\overline{\bigcup_{i \in A} S_i} \right) \cap \left(\overline{\bigcup_{k \notin B} T_k} \right) \geq \overline{A} - \sum_{k \in B} d_k.$$

As before, the necessity of these conditions is trivial. The proof of their sufficiency is given in [16].

Another generalization is given by Ford and Fulkerson [17]:

Let R and \mathcal{S} be as in (2.1). Let $a_i \leq b_i$ ($i = 1, \dots, n$) be integers associated with the elements of R . A subset $S \subset R$ of not necessarily distinct elements satisfying (2.2) in which the number of times P_i is used is at least a_i and at most b_i is called a system of restricted representatives. Such a set S exists if and only if, for any $X \subset \{1, \dots, n\}$, we have

$$(2.7) \quad \bar{X} \leq \min \left(m - \sum_{\substack{P_i \notin X \\ j \in S_i}} a_i, \sum_{\substack{P_i \in X \\ j \in S_i}} b_i \right).$$

The proof given in [17] depends on a result on network flow called the min-cut max-flow theorem [18; 19], about which we will say more in the next section. But it is equally possible to prove it via the theory of linear

inequalities, or as a direct corollary of the theorem just quoted. Let $b = \max b_i$, and construct the set consisting of b copies of R , whose points are each P_i repeated b times. Then, if we let the k th summand of the partition $\mathcal{T} = \{T_1, \dots, T_n\}$ be the b copies of P_k , we have a situation to which the previous hypotheses apply. Then (2.5) and (2.6) together are equivalent to (2.7).

Ford and Fulkerson have also considered the question of finding a subset S which is not only a system of restricted representatives for $\mathcal{S} = \{S_1, \dots, S_m\}$, but also for another family $\mathcal{T} = \{T_1, \dots, T_m\}$ (note \mathcal{T} is generally *not* a partition). They have shown that such a system of restricted representatives exists if and only if, for every $X, Y \subset \{1, \dots, m\}$, we have

$$\bar{X} + \bar{Y} \leq m - \sum_{\substack{P_i \notin \bigcup_{j \in X} S_j; P_i \notin \bigcup_{j \in Y} T_j}} a_i + \sum_{\substack{P_i \in \bigcup_{j \in X} S_j; P_i \in \bigcup_{j \in Y} T_j}} b_i.$$

See [17] for the proof, which is based on consideration of flows in networks.

Another result [20] on two families of sets deals with the problem of choosing a subset S of the given set R such that the intersection of S with each set of each family has a number of elements lying within prescribed bounds. Note that the previous theorems deal with the *assignment* of elements to sets which contain them, ignoring the fact that an element assigned to one set may be contained in others. That consideration is not ignored in the present case, so it is not astonishing that fairly stringent conditions are imposed on the two families.

Let R be a set, $\mathcal{S} = \{S_1, \dots, S_m\}$ and $\mathcal{T} = \{T_1, \dots, T_n\}$ two families of subsets of R such that $S_i \cap S_j \neq \emptyset$ implies $S_i \subset S_j$ or $S_j \subset S_i$; $T_i \cap T_j \neq \emptyset$ implies $T_i \subset T_j$ or $T_j \subset T_i$. Let $a_i \leq b_i$ ($i = 1, \dots, m$) and $c_j \leq d_j$ ($j = 1, \dots, n$) be prescribed integers. In order that there exist a set $S \subset R$ such that

$$a_i \leq \overline{S \cap S_i} \leq b_i \quad i = 1, \dots, m,$$

and

$$c_j \leq \overline{S \cap T_j} \leq d_j \quad j = 1, \dots, n,$$

it is necessary and sufficient that, for all $I \subset \{1, \dots, m\}$ and $J \subset \{1, \dots, n\}$, we have

$$\sum_{i \in I_o} a_i + \sum_{j \in J_e} c_j \leq \sum_{i \in I_e} b_i + \sum_{j \in J_o} d_j + \overline{S^o \cap T^e},$$

and

$$\sum_{j \in J_o} c_j + \sum_{i \in I_e} a_i \leq \sum_{j \in J_e} d_j + \sum_{i \in I_o} b_i + \overline{S^e \cap T^o}.$$

Now to explain the notation. By virtue of the conditions on S , it is possible to count the number of sets in $\{S_i\}_{i \in I}$ containing a given S_i ($i \in I$). We count the set S_i itself, so that, for example, the number associated with a maximal S_i —maximal in the set $\{S_i\}_{i \in I}$ —is 1. If the number associated with S_i is odd, we assign i to I_o ; if even, we assign i to I_e . This explains the symbols I_o and I_e , and a similar discussion serves to define J_o and J_e .

Further, S^o is the set of all elements of R contained in an odd number of sets $\{S_i\}_{i \in I}$, S^e is the set of all elements of R contained in an even number of sets $\{S_i\}_{i \in I}$. T^o and T^e are defined similarly.

The original proof was a fairly elaborate deduction based on the fact that the incidence matrix of elements of R versus sets in both families had the unimodular property, and exploitation of the well-known result in the theory of linear inequalities that

$$(2.8) \quad \begin{aligned} Ax \leq b \text{ is consistent if and only if} \\ y \geq 0, \quad yA = 0 \text{ implies } (y, b) \geq 0. \end{aligned}$$

But a simpler proof based on flow considerations was subsequently discovered.

3. Flows in networks. The discussion will concentrate on directed or oriented graphs, although most of what is said applies equally well to unoriented graphs.

The prototype of theorems of this type is the well-known result of Menger (see [21; 22]): If A and B are disjoint subsets of the nodes of a graph, the largest number of nodewise disjoint paths from A to B is the smallest number of nodes in any set of nodes intersecting each path. An analogous result is that the largest number of arcwise disjoint paths from A to B is the smallest number of arcs in any set of arcs intersecting each path. One can combine and generalize these two statements as follows:

Let G be a directed graph with capacities c_{ij} on arcs from node i to node j , and capacities c_{ii} on the nodes i . Let A and B be distinct nodes, designated source and sink respectively. A flow is an assignment of numbers x_{ij} to the arcs satisfying

$$\begin{aligned} 0 \leq x_{ij} \leq c_{ij}, & \quad i \neq j, \\ \sum_j x_{ij} = \sum_j x_{ji} \leq c_{ii}, & \quad i \neq A, B. \end{aligned}$$

Then the maximal flow from A to B ; i.e., $\max \sum_j x_{Aj}$ ($= \max \sum_j x_{jB}$) equals the “minimal cut,” the smallest sum of capacities of any collection of nodes and arcs which meets every path.

This result is due to Ford and Fulkerson [18]. A proof via the duality theorem of linear programming has been given by Dantzig and Fulkerson [19]. Note that the fact that “max flow \leq min cut” is trivial. The effort is to prove the equality. This is the analog of the situation in the previous section on systems of representatives where the necessity was trivial, and the only effort was required to prove the sufficiency.

Another result on flows, which does not specialize any of the nodes, is the “circulation theorem” [23]:

Let $a_{ij} \leq b_{ij}$ ($i \neq j$) be numbers associated with the arcs from i to j ,

$c_i \leq d_i$ be numbers associated with the i th node. A flow in the network is an assignment x_{ij} ($i \neq j$) satisfying

$$\begin{aligned} a_{ij} &\leq x_{ij} \leq b_{ij}, & i \neq j, \\ c_i &\leq \sum_j x_{ji} - \sum_j x_{ij} \leq d_i, & \text{all } i. \end{aligned}$$

Such a flow exists if and only if, for any subset S of the nodes (with S' its complement), we have

$$\sum_{i \in S; j \in S'} b_{ji} \geq \sum_{i \in S} c_i + \sum_{i \in S; j \in S'} a_{ij}$$

and

$$\sum_{i \in S; j \in S'} a_{ji} \leq \sum_{i \in S} d_i + \sum_{i \in S; j \in S'} b_{ij}.$$

In the event that $c_i = d_i = 0$ for all i (i.e., what enters a node must leave it), the two conditions collapse into the single condition: for any subset S of the nodes,

$$(3.1) \quad \sum_{i \in S; j \in S'} a_{ji} \leq \sum_{i \in S; j \in S'} b_{ij}.$$

This circulation theorem was originally proven through the theory of linear inequalities. It is closely related to Gale's "feasibility theorem" [24] for flows in undirected graphs. Gale has also shown [25] the equivalence of the circulation theorem and the min-cut max-flow theorem, and has further explored the relation (in theorems of this type) between the case of directed graphs and undirected graphs. Other remarks on this point are made in [19]. There has also been additional study of "dynamic flows" by Ford and Fulkerson [26], and by Gale [27].

We have thus seen that Hall's theorem can be generalized in various ways to results on systems of representatives and results on flows in networks. It is not at all difficult to show that any of the results already cited includes Hall's theorem as a special case. The tools for the generalization were the theory of linear inequalities and flow theorems—and although the latter can be discussed without invoking convex polyhedra and associated concepts, it is nevertheless quite natural to use them. Secondly [17], the demonstration that the flow theorems include the results on systems of representatives explicitly invokes the unimodular property in the manner of the discussion in the Introduction.

So it is in order for us to attribute combinatorial power to methods in the theory of linear inequalities, at least tentatively. But are these results actually stronger than Hall, or is it possible to deduce them as special cases of Hall's theorem? In short, while this trip has been fun, has it been entirely necessary?

At the present time, the answer appears to be yes and no. The next section will outline the way in which Hall's theorem may be twisted to yield

all the results so far described. §5 will, on the other hand, present a result that seems at this time to be genuinely stronger than Hall's theorem.

4. Deduction of previous results from Hall's theorem. As has been remarked earlier, it is sufficient to show that the circulation theorem is a consequence of Hall. Actually, it is sufficient merely to deduce the special case (3.1) of the circulation theorem. Ford and Fulkerson [17] have noted that there is an alternative path—from Hall to the min-cut max-flow theorem—and it is likely that their method is closely related to the one outlined below.

Although it would be possible to proceed directly from Hall to the circulation theorem, the notation would be very cumbersome, so we shall accomplish our aim in three steps:

(a) Let $(a_1, \dots, a_m), (b_1, \dots, b_n)$ be non-negative integers such that $\sum a_i = \sum b_j = S$. Let K be a subset of the set of all ordered pairs (i, j) $i = 1, \dots, m, j = 1, \dots, n$.

Then there exist integral x_{ij} , $i = 1, \dots, m, j = 1, \dots, n$ satisfying

$$(4.2) \quad \begin{aligned} x_{ij} \geq 0, (i, j) \in K &\text{ implies } x_{ij} = 0, \\ \sum_j x_{ij} = a_i & \quad i = 1, \dots, m, \\ \sum_i x_{ij} = b_j & \quad j = 1, \dots, n, \end{aligned}$$

if and only if, for every $I \subset \{1, \dots, m\} J \subset \{1, \dots, n\}$ such that $I \times J \subset K$, we have

$$(4.3) \quad 0 \geq \sum_{i \in I} a_i + \sum_{j \in J} b_j - S.$$

Proof. The necessity being trivial, we shall only discuss the sufficiency. Let R be a set with S elements $\{P_1, \dots, P_s\}$ $\mathcal{T} = \{T_1, \dots, T_s\}$ a family of S subsets of R , defined as follows: For each $j = 1, \dots, n$, we have b_j identical sets in \mathcal{T} . If $(1, j) \notin K$, then P_1, \dots, P_{a_1} are in each of these sets.

If $(1, j) \in K$, then P_1, \dots, P_{a_1} are not in any of these sets.

If $(2, j) \notin K$, then $P_{a_1+1}, \dots, P_{a_2}$ are in each of these sets.

If $(2, j) \in K$, then $P_{a_1+1}, \dots, P_{a_2}$ are not in any of these sets.

Continue in this fashion. Thus the sets in \mathcal{T} are in classes corresponding to the b_j , and the elements of R are in classes corresponding to the a_i .

Now we assert that (4.3) implies the existence of a system of distinct representatives for \mathcal{S} . To prove this, we verify (2.3). Let $L \subset \{1, \dots, S\}$. If we allow L to include all indices of sets belonging to any of the n classes of the sets in \mathcal{S} of which it contains already at least one index, then the left side of (2.3) is increased, although the right side stays the same, so (2.3) is possibly harder to satisfy. Accordingly, we may assume that L is composed of all T_i 's arising from some subset $J \subset \{1, \dots, n\}$. Then

$$\overline{\bigcup_{l \in L} S_l} = \sum_{\substack{(i, j) \notin K \\ \text{for some } j \in J}} a_i.$$

Hence, verifying (2.3) is equivalent to verifying that

$$\bar{L} = \sum_{j \in J} b_j \leq \sum_{\substack{(i,j) \notin K \\ \text{for some } j \in J}} a_i.$$

Alternatively, we must show that if $I \subset \{1, \dots, m\}$ is such that $I \times J \subset K$, then

$$\sum_{j \in J} b_j \leq S - \sum_{i \in I} a_i,$$

which is (4.3).

If we let x_{ij} be the number of elements of R in the system of distinct representatives which belong to the i th class of elements and represent the j th class of sets, then conditions (4.2) are obviously satisfied.

(b) Let $(a_1, \dots, a_m), (b_1, \dots, b_n)$ be given non-negative integers such that $\sum a_i = \sum b_j = S$. Let c_{ij} be non-negative integers, $i = 1, \dots, m, j = 1, \dots, n$. Then in order that there exist integers x_{ij} satisfying

$$(4.4) \quad \begin{aligned} 0 &\leq x_{ij} \leq c_{ij}, \\ \sum_j x_{ij} &= a_i & i = 1, \dots, m, \\ \sum_i x_{ij} &= b_j & j = 1, \dots, n, \end{aligned}$$

it is necessary and sufficient that, for all $I \subset \{1, \dots, m\}, J \subset \{1, \dots, n\}$, we have

$$(4.5) \quad \sum_{i \in I; j \in J} c_{ij} \geq \sum_{i \in I} a_i + \sum_{j \in J} b_j - S.$$

Proof. The necessity being easy, we treat only the sufficiency. This will be done by exploiting a device independently discovered by Kantorovich [28] and Dantzig [29]. An alternative device, which would have served equally well, has recently been discovered by Wagner [30].

Consider the vector $\alpha = (a_1, \dots, a_m; c_{11}, \dots, c_{mn})$ of $mn + m$ non-negative components, and the vector $\beta = (b_1, \dots, b_n; c_{11}, \dots, c_{mn})$ of $mn + n$ non-negative components. Observe that the sum of the co-ordinates of α is the same as the sum of the co-ordinates of β , namely $S + \sum_{i,j} c_{ij}$. We shall apply result (a) above, with the co-ordinates of α and β the respective row and column sums.

To specify the set K , first agree on the notation that co-ordinates of α are labeled either i or (i,j) and co-ordinates of β are labeled either j or (i,j) . Then K consists of the following combination:

Row	Column
i	j
i	(k,j) if and only if $k \neq i$
(i,j)	k if and only if $k \neq j$
(i,j)	all (k,l) except when $k = i, l = j$.

To prove that (4.5) implies (4.4), we shall show that (4.5) implies (4.3) in the present situation, and it is clear that this will prove (4.4), with $x_{ij,j} = x_{i,j} = x_{ij}$; $x_{ij,ij} = c_{ij} - x_{ij}$.

Let I be a subset of the row indices, J a subset of the column indices. Let $I = M \cup L$, where $M \subset \{1, \dots, m\}$, $L \subset \{(1,1), \dots, (m,n)\}$. Similarly, let $J = N \cup P$, where $N \subset \{1, \dots, n\}$, $P \subset \{(1,1), \dots, (m,n)\}$. Imagine M and N chosen. What are the possible choices for L and P so that $I \times J \subset K$?

Clearly,

$$L \subset \{(i,j) | j \notin N\},$$

$$P \subset \{(i,j) | i \notin M\},$$

$$L \cap P = \emptyset.$$

If we want to consider the choice of L and P to make (4.3) hardest to satisfy, we should (and may) choose them so that

$$L \cup P = \{(i,j), i \notin M \text{ or } j \notin N\}.$$

Making this choice, (4.3) becomes

$$\begin{aligned} 0 &\geq \sum_{i \in M} a_{it} + \sum_{j \in N} b_j + \sum_{(i,j) \in L \cup P} c_{ij} - S - \sum_{i,j} c_{ij} \\ &= \sum_{i \in M} a_{it} + \sum_{j \in N} b_j - S - \sum_{i \in M; j \in N} c_{ij}, \end{aligned}$$

which is (4.5).

(c) To prove the special case (3.1) of the circulation theorem, consider the following problem: find integral x_{ij} ($i,j = 1, \dots, n$) such that

$$\begin{aligned} 0 &\leq x_{ii} \leq \infty, \\ a_{it} &\leq x_{ij} \leq b_{it}, \\ \sum_i x_{ij} &= \sum_j x_{ij} = M, \end{aligned}$$

where M is an arbitrarily large integer.

Clearly this problem has a solution if and only if there is a flow. This problem can be transformed, by the substitution

$$y_{ij} = x_{ij} - a_{ij}, \quad y_{ii} = x_{ii}, \quad \text{into}$$

$$0 \leq y_{ii} \leq \infty,$$

$$0 \leq y_{ij} \leq b_{ij} - a_{ij},$$

$$\sum_j y_{ij} = M - \sum_j a_{ij}, \quad i = 1, \dots, n$$

$$\sum_i y_{ij} = M - \sum_i a_{ij}, \quad J = 1, \dots, n.$$

We can now apply (4.5). Obviously, since the upper bound on y_{ii} is ∞ , we need only consider the case $I \cap J = \emptyset$. Secondly, since $S = nM + \text{other}$

terms, (4.5) will be trivially satisfied unless $\bar{I} + \bar{J} = n$. In short, we need only consider the case where $I \subset \{1, \dots, n\}$ and J is the complement I' of I . In this case, (4.5) becomes

$$\sum_{i \in I; j \in I'} (b_{ij} - a_{ij}) \geq - \sum_{i \in I; \text{all } j} a_{ij} - \sum_{j \in I'; \text{all } i} a_{ij} + \sum_{i,j} a_{ij}.$$

With a little manipulation, this reduces to

$$\sum_{i \in I; j \in I'} b_{ij} \geq \sum_{i \in I; j \in I'} a_{ji}$$

which is (3.1).

5. The “most general” theorem of the Hall type. In the previous sections we have examined a number of combinatorial results on systems of representatives and network flow, and seen that they are all closely related—indeed, that the simplest implies all its more complicated generalizations.

This is a slight disappointment for anyone promulgating the thesis that linear inequalities are useful in discovering and proving theorems of this general character, though not a fatal disappointment, since it is a historical fact that some of these results were first reached that way. The case for linear inequalities can be made even stronger, however, by considering the following problem in systems of representatives, which appears to be quite general. Indeed, all problems considered thus far in this talk are subsumed in it.

Contemplation of the problem leads to a result [31] which appears in some sense to deserve the label of the “most general theorem” of this class.

Let R be a set with elements $\{P_1, \dots, P_n\}$ and let $\mathcal{S} = \{S_1, \dots, S_m\}$ be a family of subsets R . Let $a_i \leq b_i$ ($i = 1, \dots, m$) be given integers and $0 \leq w_j$ ($j = 1, \dots, n$) be given integers. Do there exist numbers x_j ($j = 1, \dots, n$) satisfying

$$(5.1) \quad \begin{aligned} 0 &\leq x_j \leq w_j && (j = 1, \dots, n), \\ &\text{and} \\ a_i &\leq \sum_{P_j \in S_i} x_j \leq b_i && (i = 1, \dots, m)? \end{aligned}$$

Define, for any $A \subset \{1, \dots, m\}$,

$$n_j(A) = \overline{\{i \mid P_j \in S_i, i \in A\}}, \quad j = 1, \dots, n.$$

Then it is easy to show that a necessary condition for the existence of a solution to (5.1) is that

$$(5.2) \quad \begin{aligned} A, B &\subset \{1, \dots, m\}, A \cap B = \emptyset, \text{ and} \\ |n_j(A) - n_j(B)| &\leq 1 \quad (j = 1, \dots, n) \quad \text{imply} \\ \sum_{i \in A} a_i &\leq \sum_{i \in B} b_i + \sum_{n_j(A) > n_j(B)} w_j. \end{aligned}$$

Condition (5.2) not only "sounds like" the condition (2.3) and all the other conditions we have met so far, but coincides with them when the combinatorial situations studied are put in such form that our "general problem" subsumes them. Then it is natural to inquire under what circumstances (5.2) is sufficient for the existence of a solution to (5.1). An answer is contained in the statement:

The following conditions are equivalent:

- (5.3) The m by n incidence matrix of sets S_i versus elements P_j has the unimodular property;
- (5.4) for every choice of integral $a_i \leq b_i$ ($i = 1, \dots, m$) and integral $w_j \geq 0$ ($j = 1, \dots, n$), (5.2) implies the existence of an integral solution to (5.1).

One can list other conditions, equivalent to the above two, which explore the relationship between real and integral solutions to (5.1), but from a combinatorial viewpoint, the principal interest is the equivalence of (5.3) and (5.4).

(5.3) implies (5.4) is the statement that if the incidence matrix has the unimodular property, then there is a theorem of the Hall type. (5.4) implies (5.3) is the statement that a theorem of the Hall type exists for all choices of boundary values *only* if the incidence matrix has the unimodular property. In summary, the unimodular property for the incidence matrix is a necessary and sufficient condition that (5.2) be a necessary and sufficient condition for the existence of an integral solution to (5.1). It is in this sense that the result may properly be regarded as the most general theorem of its kind.

While not difficult, the proof is long and will be published elsewhere. The key idea in the proof that (5.3) implies (5.4) is the consideration of dual sets of inequalities, where the unimodular property guarantees that if (5.1) is consistent, it has an integral solution, and further guarantees that examination of the extreme rays of the dual system is equivalent to checking (5.2). Ideas of the same general sort are involved in the proof that (5.4) implies (5.3).

In view of this theorem, it is of some interest to look for incidence matrices with the unimodular property. A special case of a result of Heller [32] shows that if all n columns of such a matrix are distinct and non-zero, and if there are n rows, then $n \leq m(m + 1)/2$. Further, this upper bound is attained if and only if the columns are all possible intervals of a simply ordered set of m points.

The most general sufficient condition known [5] for a matrix to have the unimodular property is that it be the incidence matrix of nodes versus directed paths in a directed graph all of whose loops are of even order with successive arcs alternating in direction. Until recently, every known incidence matrix with the unimodular property arose in this way, but the condition is

unaesthetically not symmetric with respect to rows and columns, although the unimodular property is. In fact, it is not difficult to give examples of incidence matrices where the rows represent nodes of an “alternating” graph, and the columns the directed paths, but no alternating graph can be found for which the columns represent nodes and the rows represent directed paths; for example, Heller [32],

$$\begin{matrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{matrix}$$

An example of an incidence matrix with the unimodular property that could not arise from an alternating graph whatever role be selected for rows and columns can be obtained by appending the column vector $(1, 1, 1, 1, 1)$ to the above matrix.

6. Remarks.

a. It is interesting to see instances of matrices with the unimodular property arising in various contexts. It has already been noted [5] that, in linear programming, the transportation problem, the warehouse problem in the form discussed by Cahn [33] and by Charnes and Cooper [34], the caterer problem of Jacobs [35 and 36] and certain production scheduling problems involving fulfilling cumulative requirements all involve incidence matrices arising from alternating graphs.

One can also see the possibility of direct application of the result in §5 in work of Mirsky [37] offering an alternative proof of Horn’s characterization [38] of the vector of diagonal elements of a hermitian matrix, and in Folner’s discussion [39] of Banach mean values in groups. In neither of these cases does the author use (5.2) to prove (5.1) but he could have.

b. Although our emphasis has been using inequalities to prove combinatorial theorems, there is some interest in the reverse process. For example, Birkhoff [40] used Hall’s theorem to prove that the vertices of the convex set of doubly stochastic matrices are the permutation matrices, a result for which many other proofs [41; 42; 43] have subsequently been offered.

c. It is also worth pointing out that, parallel to the theoretical interplay between linear inequalities and combinatorics, there has been a computational interplay. It was pointed out some time ago, see e.g. [44], that linear programming furnishes algorithms for certain combinatorial problems. So far, however, it has appeared from the Hungarian method of Kuhn [45] and its generalizations, modifications and extensions [46; 47; 48; 49] that the more fruitful relationship is the other way around: effective methods for choosing systems of distinct representatives yield algorithms for solving the transportation problem.

d. Several research questions are suggested by the material covered in this talk :

(1) The discovery of new classes of matrices with the unimodular property.

(2) The discovery of new ways of twisting problems so that matrices with the unimodular property appear. The theorem of Dilworth that the smallest number of chains whose union comprises all elements of a partially ordered set is the largest number of incomparable elements in the set (see Dilworth [49] and Fulkerson [50]) does not appear at first blush to be accessible by these methods since the matrix of elements versus chains does *not* have the unimodular property. But a method can be found to reformulate the problem [51] so that the unimodular property can be exploited. Perhaps the results of Ryser on term rank [52], and the graph theoretic theorems of Rabin and Norman [53], Berge [54], Tutte [55], [56], etc., which have at least a verbal similarity to the situations described in this talk, can be shown to be accessible by these methods. Thus far all attempts to derive them as corollaries of the main theorem of §5 have failed.

(3) The discovery of new applications of these results (which deal with finite sets) to infinite situations (through suitable finite approximation), and the discovery of non-trivial generalization—not accessible by finite approximation—to infinite combinatorial problems.

(4) There exist (see, e.g., Fan [57] and Duffin [58]) infinite-dimensional generalizations of the duality theorems in the theory of finite systems of linear inequalities; what use (if any) is the unimodular property in such circumstances, or what is the appropriate generalization of this property?

REFERENCES

1. H. B. Mann and H. J. Ryser, *Systems of distinct representatives*, Amer. Math. Monthly vol. 60 (1953) pp. 397–401.
2. E. Egervary, *Matrixok kombinatorius tulajdonsagairol*, Mat. es Fiz. Lapok vol. 38 (1931) pp. 16–28 (translated as *On combinatorial properties of matrices*, by H. W. Kuhn, Office of Naval Research Logistics Project Report, Department of Mathematics, Princeton University, 1953).
3. P. Hall, *On representatives of subsets*, J. London Math. Soc. vol. 10 (1935) pp. 26–30.
4. D. König, *Theorie der endlichen und unendlichen Graphen*, New York, Chelsea Publishing Co., 1950.
5. A. J. Hoffman and J. B. Kruskal, *Integral boundary points of convex polyhedra*, in [7, pp. 223–246].
6. I. Heller and C. B. Tompkins, *An extension of a theorem of Dantzig's*, in [7, pp. 247–254].
7. H. W. Kuhn and A. W. Tucker, eds., *Linear inequalities and related systems*, Annals of Mathematics Studies, no. 38, Princeton, 1956.
8. A. J. Hoffman, *Generalization of a theorem of König*, J. Washington Acad. Sci. vol. 46 (1956) p. 211.

9. R. Rado, *Theorems on linear combinatorial topology and general measure*, Ann. of Math. vol. 44 (1943) pp. 228–270.
10. T. S. Motzkin, *The assignment problem*, Proceedings of the Sixth Symposium in Applied Mathematics, New York, McGraw-Hill, 1956.
11. H. W. Kuhn, *A combinatorial algorithm for the assignment problem*, Issue 11 of Logistics Papers, George Washington University Logistics Research Project, 1954.
12. A. J. Hoffman, *Linear programming*, Applied Mechanics Reviews (9), 1956.
13. P. R. Halmos and Herbert E. Vaughan, *The marriage problem*, Amer. J. Math. vol. 72 (1950) pp. 214–215.
14. H. J. Ryser, *Geometrics and incidence matrices*, Slaught Memorial Paper Contributions to Geometry, Amer. Math. Monthly vol. 62 (1955) pp. 25–31.
15. A. J. Hoffman and H. W. Kuhn, *Systems of distinct representatives and linear programming*, Amer. Math. Monthly vol. 63 (1956) pp. 455–460.
16. ———, *On systems of distinct representatives*, in [7, pp. 199–206].
17. L. R. Ford, Jr. and D. R. Fulkerson, *Network flow and systems of representatives*, Canad. J. Math. vol. 10 (1958) pp. 78–84.
18. ———, *Maximal flow through a network*, Canad. J. Math. vol. 8 (1956) pp. 399–404.
19. G. B. Dantzig and D. R. Fulkerson, *On the min-cut max-flow theorem of networks*, in [7, pp. 215–221].
20. A. J. Hoffman, 1955, Unpublished.
21. K. Menger, *Zur allgemeinen Kurventheorie*, Fund. Math. vol. 10 (1927) pp. 96–115.
22. G. Hajos, *Zum Mengerschen Graphensatz*, Acta Litterarum ac Scientiarum, Szeged vol. 7 (1934) pp. 44–47.
23. A. H. Hoffman, 1956, Unpublished.
24. D. Gale, *A theorem on flows in networks*, Pacific J. Math. vol. 7 (1957) pp. 1073–1082.
25. ———, 1956, Unpublished.
26. L. R. Ford, Jr. and D. R. Fulkerson, *Dynamic network flow*, 1957, Unpublished.
27. D. Gale, *Transient flows in networks*, 1958, Unpublished.
28. L. V. Kantorovich and M. K. Gavurin, *Problems of increasing the effectiveness of transport works*, AN USSR, 1949, pp. 110–138. I am indebted to G. B. Dantzig for this reference.
29. G. B. Dantzig, *Upper bounds, secondary constraints and block triangularity in linear programming*, Econometrica vol. 23 (1955) pp. 174–183.
30. H. M. Wagner, *On the capacitated Hitchcock problem*, 1958, Unpublished.
31. A. J. Hoffman, 1956, Unpublished.
32. I. Heller, *On linear systems with integral valued solutions*, Pacific J. Math. vol. 7 (1957) pp. 1351–1364.
33. A. S. Cahn, *The warehouse problem*, Bull. Amer. Math. Soc. vol. 54 (1948) p. 1073 (abstract).
34. A. Charnes and W. W. Cooper, *Generalizations of the warehousing model*, Operations Res. Q. vol. 6 (1955) pp. 131–172.
35. W. W. Jacobs, *The caterer problem*, Naval Res. Logist. Quart. vol. 1 (1954) pp. 154–165.
36. J. W. Gaddum, A. J. Hoffman and D. Sokolowsky, *On the solution of the caterer problem*, Naval Res. Logist. Quart. vol. 1 (1954) pp. 223–229.
37. L. Mirsky, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc. vol. 33 (1958) pp. 14–21.
38. A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix*, Amer. J. Math. vol. 76 (1954) pp. 620–630.

39. E. Folner, *On groups with full Banach mean value*, Math. Scand. vol. 3 (1955) pp. 243–254.
40. G. Birkhoff, *Three observations on linear algebra*, Universidad Nacional de Tucuman, Revista Series A vol. 5 (1946) pp. 147–151.
41. A. J. Hoffman and H. W. Wielandt, *The variation of the spectrum of a normal matrix*, Duke Math. J. vol. 20 (1953) pp. 37–39.
42. J. von Neumann, *A certain zero-sum two-person game equivalent to the operational assignment problem*, in Contributions to the Theory of Games, vol. II, pp. 5–12 (edited by H. W. Kuhn and A. W. Tucker), Annals of Mathematics Studies, no. 28, Princeton, 1953.
43. J. Hammersley and W. Mauldon, *General principles of antithetic variates*, Proc. Cambridge Philos. Soc. vol. 52 (1956) pp. 476–481.
44. G. B. Dantzig, *Application of the simplex method to a transportation problem*, T. C. Koopmans, ed., *Activity Analysis of Production and Allocation*, Cowles Commission Monograph No. 13, New York, Wiley, 1951.
45. H. W. Kuhn, *The Hungarian method for solving the assignment problem*, Naval Res. Logist. Quart. vol. 2 (1955) pp. 83–97.
46. L. R. Ford, Jr. and D. R. Fulkerson, *A simple algorithm for finding maximal network flows and an application to the Hitchcock problem*, Canad. J. Math. vol. 9 (1957) pp. 210–218.
47. M. M. Flood, *The traveling salesman problem*, J. Operations Res. Soc. Amer. vol. 4 (1956) pp. 61–75.
48. J. R. Munkres, *Algorithms for the assignment and transportation problem*, J. Soc. Indust. Appl. Math. vol. 5 (1957) pp. 32–38.
49. R. P. Dilworth, *A decomposition theorem for partially ordered sets*, Ann. of Math. vol. 51 (1950) pp. 161–166.
50. D. R. Fulkerson, *Note on Dilworth's decomposition theorem for partially ordered sets*, Proc. Amer. Math. Soc. vol. 7 (1956) pp. 701–702.
51. G. B. Dantzig and A. J. Hoffman, *Dilworth's theorem on partially ordered sets*, in [7, pp. 207–214].
52. H. J. Ryser, *The term rank of a matrix*, Canad. J. Math. vol. 60 (1957) pp. 57–65.
53. M. O. Rabin and R. Z. Norman, *An algorithm for the minimum cover of a graph*, 1957, Unpublished.
54. C. Berge, *Two theorems in graph theory*, Proc. Nat. Acad. Sci. vol. 43 (1957) pp. 842–844.
55. W. T. Tutte, *The factorization of linear graphs*, J. London Math. Soc. vol. 22 (1947) pp. 107–111.
56. ———, *The factors of graphs*, Canad. J. Math. vol. 4 (1952) pp. 314–328.
57. K. Fan, *On systems of linear inequalities*, in [7, pp. 99–156].
58. R. J. Duffin, *Infinite programs*, in [7, pp. 157–170].

GENERAL ELECTRIC COMPANY,
NEW YORK, NEW YORK

This page intentionally left blank

A COMBINATORIAL EQUIVALENCE OF MATRICES

BY

A. W. TUCKER

This paper deals with an elementary equivalence relation on matrices, called "combinatorial equivalence" because each equivalence class contains just a *finite* number of matrices. It stems from an attempt to study for a general field, rather than an ordered field, the linear algebraic structure underlying the "simplex method" of G. B. Dantzig, so remarkably effective in Linear Programming. However, this structure is outlined here by itself as a topic that seems to have wide applicability and interest.

Given two m by n matrices A and B , with entries from an arbitrary field, we form the two systems of m homogeneous linear equations in $m + n$ variables

$$x + Ay = [I, A] \begin{bmatrix} x \\ y \end{bmatrix} = 0 \quad \text{and} \quad u + Bv = [I, B] \begin{bmatrix} u \\ v \end{bmatrix} = 0,$$

where x, u and y, v are column-vectors with m and n components (variables) and I is an identity matrix of order m . We say that A and B are *combinatorially equivalent*, and write $A :: B$, if the systems $x + Ay = 0$ and $u + Bv = 0$ have the same solutions except for a permutation of the component variables—that is, if there is a one-one correspondence between the solution sets

$$\{x, y | x + Ay = 0\} \quad \text{and} \quad \{u, v | u + Bv = 0\}$$

given by

$$\begin{bmatrix} x \\ y \end{bmatrix} = P \begin{bmatrix} u \\ v \end{bmatrix},$$

where P is a permutation matrix of order $m + n$. For brevity, we describe this relation between $x + Ay = 0$ and $u + Bv = 0$ by saying that the two systems (as an ordered pair) are "equivalent within P ."

The relation of combinatorial equivalence is (i) reflexive, (ii) symmetric, and (iii) transitive. For, we see: (i) that $A :: A$, since $x + Ay = 0$ is equivalent (within identity) to itself; (ii) that $A :: B$ implies $B :: A$, since $u + Bv = 0$ and $x + Ay = 0$ are equivalent within $P^{-1} = P^T$; and (iii) that $A :: B$ and $B :: C$ imply $A :: C$, since $x + Ay = 0$ and $s + Ct = 0$ are equivalent within PQ when $x + Ay = 0$ and $u + Bv = 0$ are equivalent within P and $u + Bv = 0$ and $s + Ct = 0$ are equivalent within Q .

THEOREM 1. *A :: B if, and only if, a matrix G can be formed from some m columns (in any order) of the m by m + n matrix [I, A] and a matrix H from the remaining n columns (in any order) such that*

$$(1) \quad GB = H.$$

Note that G must be nonsingular, since its columns form a basis for the columns of $[I, A]$, which has rank m .

Proof. The equation

$$(2) \quad [I, A]P = [G, H]$$

serves to determine a permutation matrix P when G and H are given, and vice versa. The systems $x + Ay = 0$ and $Gu + Hv = 0$ are equivalent within P . The systems $Gu + Hv = 0$ and $u + Bv = 0$ are equivalent (within identity) if, and only if,

$$(3) \quad [G, H] = G[I, B],$$

i.e. $H = GB$, G being nonsingular. Hence, $x + Ay = 0$ and $u + Bv = 0$ are equivalent within P if, and only if, (2) and (3) hold. Note that equations (2) and (3) together imply that G is nonsingular; otherwise the rank m of $[I, A]$ would not be conserved.

For example,

$$A = \begin{bmatrix} 3 & 4 & 0 \\ 5 & 9 & 2 \end{bmatrix} :: \begin{bmatrix} 3/4 & 1/4 & 0 \\ -7/4 & -9/4 & 2 \end{bmatrix} = B$$

since

$$\begin{aligned} [I, A]P &= \left[\begin{array}{cc|ccc} 1 & 0 & 3 & 4 & 0 \\ 0 & 1 & 5 & 9 & 2 \end{array} \right] \left[\begin{array}{ccccc} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right] = \left[\begin{array}{cc|cc} 4 & 0 & 3 & 1 & 0 \\ 9 & 1 & 5 & 0 & 2 \end{array} \right] = [G, H] \\ &= \begin{bmatrix} 4 & 0 \\ 9 & 1 \end{bmatrix} \left[\begin{array}{cc|ccc} 1 & 0 & 3/4 & 1/4 & 0 \\ 0 & 1 & -7/4 & -9/4 & 2 \end{array} \right] = G[I, B]. \end{aligned}$$

More briefly, the verification is made by

$$GB = \begin{bmatrix} 4 & 0 \\ 9 & 1 \end{bmatrix} \left[\begin{array}{ccc} 3/4 & 1/4 & 0 \\ -7/4 & -9/4 & 2 \end{array} \right] = \begin{bmatrix} 3 & 1 & 0 \\ 5 & 0 & 2 \end{bmatrix} = H.$$

This is the case marked *4/4* in the accompanying table, which exhibits the equation $GB = H$ for certain matrices B combinatorially equivalent to the given matrix A . (The meaning of the italicized matter and asterisks will appear presently.)

THEOREM 2. *The class $\{A\}$ of matrices combinatorially equivalent to A contains at most $(m + n)!$ matrices.*

Proof. For a given A , there are at most $(m + n)!$ cases in (2), since the number of permutation matrices P of order $m + n$ is $(m + n)!$. However,

TABLE

	$A = \begin{bmatrix} 3 & 4 & 0 \\ 5 & 9 & 2 \end{bmatrix}$
1/1	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3/1 & 4/1 & 0/1 \\ 5/1 & 9/1 & 2/1 \\ 8/1 & 6/1 & 7/1 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 0 \\ 5 & 9 & 2 \end{bmatrix}$
2/2	$\begin{bmatrix} 1 & 0 \\ 0 & 2^* \end{bmatrix} \begin{bmatrix} 6/2 & 8/2 & 0/2 \\ 5/2 & 9/2 & 1/2 \\ 4/2 & 3/2 & 7/2 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 0 \\ 5 & 9 & 1 \end{bmatrix}$
3/3	$\begin{bmatrix} 3^* & 0 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 4/3 & 0/3 \\ -5/3 & 7/3 & 6/3 \\ 8/3 & 2/3 & 9/3 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 0 \\ 0 & 9 & 2 \end{bmatrix}$
4/4	$\begin{bmatrix} 4^* & 0 \\ 9 & 1 \end{bmatrix} \begin{bmatrix} 3/4 & 1/4 & 0/4 \\ -7/4 & -9/4 & 8/4 \\ 2/4 & 6/4 & -5/4 \end{bmatrix} = \begin{bmatrix} 3 & 1 & 0 \\ 5 & 0 & 2 \end{bmatrix}$
5/5	$\begin{bmatrix} 1 & 3 \\ 0 & 5^* \end{bmatrix} \begin{bmatrix} -3/5 & -7/5 & -6/5 \\ 1/5 & 9/5 & 2/5 \\ 8/5 & 0/5 & -4/5 \end{bmatrix} = \begin{bmatrix} 0 & 4 & 0 \\ 1 & 9 & 2 \end{bmatrix}$
6/6	$\begin{bmatrix} 3^* & 0^* \\ 5^* & 2^* \end{bmatrix} \begin{bmatrix} 2/6 & 8/6 & 0/6 \\ -5/6 & 7/6 & 3/6 \\ 4/6 & 1/6 & 9/6 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 0 \\ 0 & 9 & 1 \end{bmatrix}$
7/7	$\begin{bmatrix} 3^* & 4^* \\ 5^* & 9^* \end{bmatrix} \begin{bmatrix} 9/7 & -4/7 & -8/7 \\ -5/7 & 3/7 & 6/7 \\ 0/7 & 2/7 & 1/7 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \end{bmatrix}$
8/8	$\begin{bmatrix} 4^* & 0^* \\ 9^* & 2^* \end{bmatrix} \begin{bmatrix} 6/8 & 2/8 & 0/8 \\ -7/8 & -9/8 & 4/8 \\ 1/8 & 3/8 & -5/8 \end{bmatrix} = \begin{bmatrix} 3 & 1 & 0 \\ 5 & 0 & 1 \end{bmatrix}$
9/9	$\begin{bmatrix} 1 & 4 \\ 0 & 9^* \end{bmatrix} \begin{bmatrix} 7/9 & -4/9 & -8/9 \\ 5/9 & 1/9 & 2/9 \\ 0/9 & 6/9 & 3/9 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 5 & 1 & 2 \end{bmatrix}$

a particular P can occur in (2) if, and only if, the corresponding G is non-singular.

For the particular two by three matrix A in the accompanying table, the combinatorial equivalence class $\{A\}$ contains $9 \cdot 12 = 108$ matrices: each of the nine matrices (with fractional entries) exhibited there represents a subclass of $2!3! = 12$ matrices obtained from it by permuting its two rows

and/or three columns. The full number $(2 + 3)! = 120$ is not attained because the fifth and second columns of

$$[I, A] = \left[\begin{array}{cc|cc} 1 & 0 & 3 & 4 & 0 \\ 0 & 1 & 5 & 9 & 2 \end{array} \right]$$

are linearly dependent, and so cannot be used to form a basis matrix G in terms of which a matrix H of the remaining columns is expressible as $H = GB$.

THEOREM 3. *Let $\{\alpha\}$ denote the set of square submatrices of A of all orders (including an empty submatrix ϕ of order zero). Let $\{\beta\}$ denote the like set for B , where $A :: B$. Then there is a one-one correspondence*

$$\beta \leftrightarrow \alpha$$

between $\{\beta\}$ and $\{\alpha\}$ such that corresponding subdeterminants $|\beta|$ and $|\alpha|$ are proportional within sign. Specifically,

$$(4) \quad |\beta| = \pm |\alpha| / |\alpha^*|,$$

where $\alpha = \alpha^$ corresponds to $\beta = \phi$ (taking $|\phi| = 1$). The nonsingular square submatrix α^* is called the pivot of (the ordered pair) $A :: B$.*

Proof. Let α denote a square submatrix¹ of A of order r , $0 \leq r \leq \min(m, n)$; for $r = 0$, α is the empty submatrix ϕ . Let A_α denote the m by r submatrix of A whose columns contain the entries of α , and I_α the m by r submatrix of I whose columns have their unit entries in the rows of $[I, A]$ which contain the entries of α . Then

$$\alpha = I_\alpha^T A_\alpha.$$

Let $\bar{\alpha}$ denote the m th order square submatrix $[I'_\alpha, A_\alpha]$ of $[I, A]$, where I'_α is the m by $m - r$ submatrix of I obtained by deleting I_α . Then

$$\bar{\alpha} = [I'_\alpha, A_\alpha] \leftrightarrow I_\alpha^T A_\alpha = \alpha$$

determines a one-one correspondence between the set $\{\bar{\alpha}\}$ of square submatrices of $[I, A]$ of order m and the set $\{\alpha\}$ of square submatrices of A of all orders, such that

$$|\bar{\alpha}| = \pm |\alpha|.$$

Note that $\alpha = \phi$ when $\bar{\alpha} = I$; hence we agree that $|\phi| = 1$. Analogous to (the inverse of) the correspondence $\bar{\alpha} \leftrightarrow \alpha$, there is a one-one correspondence

$$\beta = I_\beta^T B_\beta \leftrightarrow [I'_\beta, B_\beta] = \bar{\beta}$$

¹ By definition, a *submatrix* C_0 of a matrix C is obtained by deleting certain rows and/or columns from C (including none or all). Note that the rows and columns of C_0 occur in definite orders inherited from C .

such that

$$|\beta| = \pm |\bar{\beta}|.$$

The equation (see (2), above)

$$G[I, B] = [I, A]P$$

determines a one-one correspondence

$$\bar{\beta} = [I'_\beta, B_\beta] \leftrightarrow [I'_\alpha, A_\alpha] = \bar{\alpha}$$

in which

$$G\bar{\beta} = G[I'_\beta, B_\beta] = [I'_\alpha, A_\alpha]P^{(m)} = \bar{\alpha}P^{(m)},$$

where $P^{(m)}$ is a submatrix of P which is a permutation matrix of order m . Under this correspondence

$$|G||\beta| = \pm |\bar{\alpha}|.$$

Combining the three correspondences $\beta \leftrightarrow \bar{\beta} \leftrightarrow \bar{\alpha} \leftrightarrow \alpha$, we have

$$|G||\beta| = \pm |G||\bar{\beta}| = \pm |\bar{\alpha}| = \pm |\alpha|.$$

Hence

$$|\beta| = \pm |\alpha|/|G|.$$

Let α^* denote the submatrix of A , called the *pivot* of $A :: B$, that corresponds to the empty submatrix ϕ of B under $\beta \leftrightarrow \alpha$. Then, since $\beta = I$ when $\beta = \phi$,

$$|G||\phi| = |G||I| = \pm |\alpha^*|,$$

i.e., $|\alpha^*| = \pm |G| \neq 0$. Therefore, under the one-one correspondence $\beta \leftrightarrow \alpha$,

$$|\beta| = \pm |\alpha|/|\alpha^*|.$$

Note that the choice of sign, + or -, is left quite indeterminate throughout this discussion: regard each case of \pm as an independent choice.

For example, the matrix A in the accompanying table has been chosen so that its $(2+3)!/2!3! = 10$ subdeterminants have the values $0, 1, \dots, 9$. Specifically, the subdeterminant of order zero is 1; those of order one are 3, 4, 0, 5, 9, 2 (the individual entries of A); and those of order two are 8, 6, 7 (corresponding to the submatrices of A obtained by deleting the first, second and third columns, respectively). In the body of the table, the fractional entries of each matrix B are its subdeterminants of order one; just below each column of B appears in italics the subdeterminant of order two going with the submatrix formed by the remaining columns of B ; and at the left margin appears in italics the subdeterminant of order zero (expressed as a fraction). In each case we observe that the numerators of these ten subdeterminants of B involve precisely the values $0, 1, \dots, 9$ of the ten subdeterminants of A ; the common denominator is the determinant of the pivot α^* whose entries are starred in the square matrix G of order two that

premultiplies B (in the first case, marked 1/1, no entries are starred because $\alpha^* = \phi$).

THEOREM 4. *Let $\alpha^* = A_{11}$ be the pivot of $A :: B$. Then*

$$(5) \quad M^TBN = \begin{bmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix},$$

where A_{12}, A_{21}, A_{22} are the remaining submatrices of A determined by A_{11} , and M and N are permutation matrices of orders m and n . Conversely, given any nonsingular square submatrix A_{11} of A and any permutation matrices M and N of orders m and n , there is determined by (5) a matrix B such that $A :: B$.

Proof. Let $A :: B$ with α^* as pivot. Then, in the notation of the proof of Theorem 3,

$$\beta = \phi \leftrightarrow \bar{\beta} = I \leftrightarrow \bar{\alpha} = [I'_{\alpha^*}, A_{\alpha^*}] \leftrightarrow \alpha = \alpha^* = I_{\alpha^*}^T A_{\alpha^*},$$

with $|\alpha^*| = \pm |G| \neq 0$. For simpler notation, we now let

$$I_1 = I_{\alpha^*}, \quad I_2 = I'_{\alpha^*}, \quad A_1 = A_{\alpha^*}, \quad A_2 = A'_{\alpha^*},$$

where A'_{α^*} denotes the submatrix of A obtained by deleting A_{α^*} . Let K, L, M, N denote permutation matrices determined by

$$IK = K = [I_1, I_2], \quad AL = [A_1, A_2], \quad GM = [A_1, I_2], \quad HN = [I_1, A_2],$$

where G and H are as in (2) and (3). Then

$$K^T IK = [I_1, I_2]^T [I_1, I_2] = \begin{bmatrix} I_{11} & 0 \\ 0 & I_{22} \end{bmatrix},$$

$$K^T AL = [I_1, I_2]^T [A_1, A_2] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

$$K^T GM = [I_1, I_2]^T [A_1, I_2] = \begin{bmatrix} A_{11} & 0 \\ A_{21} & I_{22} \end{bmatrix},$$

$$K^T HN = [I_1, I_2]^T [I_1, A_2] = \begin{bmatrix} I_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where

$$I_{\mu\nu} = I_\mu^T I_\nu \quad \text{and} \quad A_{\mu\nu} = I_\mu^T A_\nu \quad (\mu, \nu = 1, 2).$$

Since A_{11} is the pivot α^* , A_{11} is a nonsingular square submatrix; since $K^T = K^{-1}$ (as with any permutation matrix), we have $K^T IK = I$ and so the submatrices I_{11} and I_{22} are identity matrices, while $I_{12} = 0$ and $I_{21} = 0$. From (1), rewritten as $B = G^{-1}H$, we have

$$M^T BN = M^T G^{-1} K K^T H N = (K^T GM)^{-1} (K^T H N).$$

Hence

$$M^TBN = \begin{bmatrix} A_{11} & 0 \\ A_{21} & I_{22} \end{bmatrix}^{-1} \begin{bmatrix} I_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}^{-1} & 0 \\ -A_{21}A_{11}^{-1} & I_{22} \end{bmatrix} \begin{bmatrix} I_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}.$$

This establishes (5).

Conversely, let A_{11} be any nonsingular square submatrix of A , let I_1 and A_1 be the m -rowed submatrices of I and A such that $A_{11} = I_1^T A_1$, let I_2 and A_2 be the submatrices obtained by deleting I_1 and A_1 from I and A , and let M and N be any permutation matrices of orders m and n . Then the above steps can be reversed to construct

$$B = M \begin{bmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} N^T$$

so that $B = G^{-1}H$, where

$$G = [A_1, I_2]M^T \quad \text{and} \quad H = [I_1, A_2]N^T.$$

That is, $GB = H$ and consequently, by Theorem 1, the constructed B is such that $A :: B$. This completes the proof of Theorem 4.

Each nonsingular square submatrix α^* of A determines within the class $\{A\}$ of matrices combinatorially equivalent to A a subclass $\{A\}_{\alpha^*}$ formed by the set of matrices B for which α^* is the pivot of $A :: B$. In this way the class $\{A\}$ is partitioned into subclasses $\{A\}_{\alpha^*}$ in one-one correspondence with the nonsingular square submatrices α^* of A . The subclass $\{A\}_{\alpha^*}$ consists of the $m!n!$ matrices B obtained by allowing M and N in Theorem 4 to vary over all permutation matrices of orders m and n . That is, $\{A\}_{\alpha^*}$ is generated by taking one particular B and then permuting its rows and columns in all possible ways. As the preferred "generator" of the permutation subclass $\{A\}_{\alpha^*}$, we choose the particular matrix B obtained by taking $M = K$ and $N = L$, where K and L are permutation matrices uniquely determined by $\alpha^* = A_{11}$ (see the proof of Theorem 4). We call this particular B the *pivotal transform* of A by α^* . Then

$$(6) \quad K^TBL = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

and, reciprocally,

$$(7) \quad K^TAL = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} B_{11}^{-1} & B_{11}^{-1}B_{12} \\ -B_{21}B_{11}^{-1} & B_{22} - B_{21}B_{11}^{-1}B_{12} \end{bmatrix}$$

Thus, A is the pivotal transform of B by $\beta^* = B_{11} = A_{11}^{-1} = \alpha^{*-1}$ if B is the pivotal transform of A by α^* . We note that the pivotal transformation from A to B by α^* involves a permutation P in (2) which merely

interchanges the columns of $I_1 = I_{\alpha^*}$ with the columns of $A_1 = A_{\alpha^*}$, paired off in the order they stand (the first column of I_1 with the first of A_1 , the second with the second, and so on). This is apparent from

$$[IK, AL] = [I_1, I_2; A_1, A_2] \quad \text{and} \quad [GK, HL] = [A_1, I_2; I_1, A_2].$$

We say that the pivotal transformation from A to B by α^* has *order* r^* , if the nonsingular square submatrix α^* has order r^* . When $r^* = 0$, the pivot α^* is the empty submatrix ϕ and the pivotal transform of A by ϕ is just A itself. When $r^* = 1$, the pivot α^* is a nonzero entry of A , say a_{pq} . Then the pivotal transform of A by $a_{pq} (\neq 0)$ is the matrix B with entries as follows :

$$\begin{aligned} b_{pq} &= a_{pq}^{-1} = 1/a_{pq}, \\ b_{pj} &= a_{pq}^{-1} a_{pj}, & (j \neq q), \\ b_{iq} &= -a_{iq} a_{pq}^{-1}, & (i \neq p), \\ b_{ij} &= a_{ij} - a_{iq} a_{pq}^{-1} a_{pj}, & (i \neq p, j \neq q), \\ &= a_{ij} + b_{iq} a_{pj}. \end{aligned}$$

(Note that the arithmetic of this computation can be accomplished efficiently in $2mn - m - n + 1$ individual steps : one division, $mn - 1$ multiplications, and $(m - 1)(n - 1)$ additions.) Such pivotal transformations of order one will play a crucial rôle in Theorem 7.

For example, the nine matrices B exhibited in the table are all the possible pivotal transforms of the matrix A given there : one of order zero (case 1/1), five of order one ($2/2, 3/3, 4/4, 5/5, 9/9$), and three of order two ($6/6, 7/7, 8/8$). Note in each case that $[G, H]$ is formed from $[I, A]$ by direct interchange of columns (none, one, or two) of I with columns of A , paired off in the order in which they stand. By permuting rows and columns of each B in the table in all possible ways, one obtains a subclass $\{A\}_{\alpha^*}$ of $2!3! = 12$ members. The nine subclasses $\{A\}_{\alpha^*}$ unite to form the class $\{A\}$ of all $9 \cdot 12 = 108$ matrices combinatorially equivalent to the given A .

REMARK. Let B be the pivotal transform of A by $\alpha^* = A_{11}$ (as defined above) and let x_1 and x_2 , y_1 and y_2 , u_1 and u_2 , v_1 and v_2 be subvectors of x, y, u, v such that

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = K^T x, \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = L^T y, \quad \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = K^T u, \quad \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = L^T v.$$

Then, equivalence within P of the systems $x + Ay = 0$ and $u + Bv = 0$ corresponds exactly to equivalence within

$$x_1 = v_1, \quad x_2 = u_2, \quad y_1 = u_1, \quad y_2 = v_2$$

of the systems

$$x_1 + A_{11}y_1 + A_{12}y_2 = 0 \quad u_1 + B_{11}v_1 + B_{12}v_2 = 0$$

and

$$x_2 + A_{21}y_1 + A_{22}y_2 = 0 \quad u_2 + B_{21}v_1 + B_{22}v_2 = 0.$$

That is, we must be able to pass from one system to the other by solving the first equation for y_1 or v_1 and substituting in the second equation. Carrying this through, we get the relations between $A_{\mu\nu}$ and $B_{\mu\nu}$ set forth above in (6) and (7).

THEOREM 5. $(-A^T) :: (-B^T)$ if, and only if, $A :: B$.

Proof. Transposing rows and columns consistently throughout (5) and at the same time multiplying by minus one, we get

$$N^T(-B^T)M = \begin{bmatrix} (-A_{11}^T)^{-1} & (-A_{11}^T)^{-1}(-A_{21}^T) \\ -(-A_{12}^T)(-A_{11}^T)^{-1} & (-A_{22}^T) - (-A_{12}^T)(-A_{11}^T)^{-1}(-A_{21}^T) \end{bmatrix}.$$

Hence, by Theorem 4, $(-A^T) :: (-B^T)$ if, and only if, $A :: B$.

This negative-transpose “duality” of combinatorial equivalence corresponds in a fundamental way with the duality in matrix games and in pairs of dual linear programs. Note that it is not true in general that $A :: B$ implies either $A^T :: B^T$ or $(-A) :: (-B)$.

THEOREM 6. Let α be a square submatrix of A of order r that contains as submatrix the pivot α^* of $A :: B$, the order of α^* being r^* . Then the square submatrix β to which α corresponds under the one-one correspondence $\beta \leftrightarrow \alpha$ (of Theorem 3) has order $s = r - r^*$. Moreover, α is the pivot of $A :: C$ if, and only if, β is the pivot of $B :: C$.

Proof. For simplicity of notation, we assume that the nonsingular square matrix α^* stands in the first r^* rows and r^* columns of A and that α stands in the first r rows and r columns of A . (Otherwise, permutation matrices K , L , etc., come into play in a far more heavy-handed way than in the proof of Theorem 4.) We partition I and A to have

$$I = [I_1, I_2, I_3] \quad \text{and} \quad A = [A_1, A_2, A_3],$$

where I_μ and A_ν are m -rowed submatrices with m_μ and n_ν columns ($\mu, \nu = 1, 2, 3$) such that $m_1 = n_1 = r^*$, $m_2 = n_2 = s = r - r^*$, and $m_3 = m - r$ and $n_3 = n - r$. By further partitioning,

$$I = \begin{bmatrix} I_{11} & 0 & 0 \\ 0 & I_{22} & 0 \\ 0 & 0 & I_{33} \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

where

$$I_{\mu\nu} = I_\mu^T I_\nu \quad \text{and} \quad A_{\mu\nu} = I_\mu^T A_\nu \quad (\mu, \nu = 1, 2, 3)$$

are m_μ by m_ν and m_μ by n_ν submatrices. Then

$$\alpha^* = A_{11} \quad \text{and} \quad \alpha = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Let the permutation matrix P in (2) and the matrix B in (3), partitioned into $[B_1, B_2, B_3]$ in the same manner as A , be taken so that

$$[I_1, I_2, I_3; A_1, A_2, A_3]P = [A_1, I_2, I_3; I_1, A_2, A_3] = \\ [A_1, I_2, I_3][I_1, I_2, I_3; B_1, B_2, B_3].$$

Then, in particular, it follows from the second of these equations that

$$[A_1, A_2, I_3] = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ A_{31} & A_{32} & I_{33} \end{bmatrix} = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & I_{22} & 0 \\ A_{31} & 0 & I_{33} \end{bmatrix} \begin{bmatrix} I_{11} & B_{12} & 0 \\ 0 & B_{22} & 0 \\ 0 & B_{32} & I_{33} \end{bmatrix} \\ = [A_1, I_2, I_3][I_1, B_2, I_3],$$

where $B_{\mu\nu}$ denote m_μ by n_ν submatrices obtained by further partitioning (just the same as with A). Hence

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{11}| |B_{22}|.$$

In the notation used in the proof of Theorem 3,

$$\beta = B_{22} \leftrightarrow \bar{\beta} = [I_1, B_2, I_3] \leftrightarrow [A_1, A_2, I_3] = \bar{\alpha} \leftrightarrow \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \alpha$$

and

$$|\alpha| = |\alpha^*| |\beta|, \quad \text{i.e.} \quad |\beta| = |\alpha| / |\alpha^*|.$$

(There is no ambiguity of sign here due to the “main-diagonal” location of α^* , α and β .) The square submatrix $\beta = B_{22}$ has order $m_2 = n_2 = s = r^* - r$.

Turning to the second part of Theorem 6, we assume that α and β are nonsingular ($|\beta| \neq 0$ if, and only if, $|\alpha| \neq 0$). Let

$$[I_1, I_2, I_3; B_1, B_2, B_3]Q = [I_1, B_2, I_3; B_1, I_2, B_3] = \\ [I_1, B_2, I_3][I_1, I_2, I_3; C_1, C_2, C_3],$$

where $C = [C_1, C_2, C_3]$ is partitioned in the same way as A and B . Then $B :: C$ with $I_2^T B_2 = \beta$ as pivot. However, using the relations between A and B at the beginning of the preceding paragraph,

$$[I_1, I_2, I_3; A_1, A_2, A_3]PQ = [A_1, I_2, I_3; I_1, A_2, A_3]Q \\ = [A_1, A_2, I_3; I_1, I_2, A_3] \\ = [A_1, I_2, I_3][I_1, I_2, I_3; B_1, B_2, B_3]Q \\ = [A_1, I_2, I_3][I_1, B_2, I_3][I_1, I_2, I_3; C_1, C_2, C_3] \\ = [A_1, A_2, I_3][I_1, I_2, I_3; C_1, C_2, C_3].$$

Hence $A :: C$ with $[I_1, I_2]^T [A_1, A_2] = \alpha$ as pivot. Also, these steps can be reversed to show that β is the pivot of $B :: C$ if α is the pivot of $A :: C$. This completes the proof of Theorem 6, under the assumption that α^* , α and β

have “main-diagonal” location. However, Theorem 6 holds in general since it is clear that the stated results carry over (within sign) to any matrices A' , B' and C' drawn from the permutation subclasses $\{A\}_\phi$, $\{A\}_{\alpha^*}$ and $\{A\}_\alpha = \{B\}_\beta$.

As a corollary of the first part of Theorem 6, we observe that the one-one correspondence $\beta \leftrightarrow \alpha$ (of Theorem 3) preserves the orders of these square submatrices if, and only if, the pivot of $A :: B$ is the empty submatrix ϕ (which is a submatrix of every α of A). That is, orders are preserved if, and only if, B belongs to the subclass $\{A\}_\phi$ of matrices obtained from A by mere permutation of rows and/or columns.

THEOREM 7. *Two matrices are combinatorially equivalent if, and only if, it is possible to pass from the one to the other by a finite succession of elementary operations of the following three types :*

- (i) *interchange of any two rows,*
- (ii) *interchange of any two columns,*
- (iii) *a pivotal transformation of order one.*

Proof. The “if” part is obvious, since each elementary operation leads to a combinatorially equivalent matrix and combinatorial equivalence is transitive. To prove the “only if” part, we take any two combinatorially equivalent matrices A and C . Let the pivot of $A :: C$ be the nonsingular square submatrix α of order r . We then set up a succession of r pivotal transformations of order one as follows :

$$\begin{array}{ccccccc} A & \xrightarrow{a} & A^1 & \cdots & A^k & \xrightarrow{a^k} & A^{k+1} & \cdots & A^r \\ | & & | & & | & & | & & | \\ \alpha & & \alpha^1 & & \alpha^k & & \alpha^{k+1} & & \alpha^r = \phi. \\ C & & C & & C & & C & & C \end{array}$$

A^1 is the pivotal transform of A by any selected entry $a \neq 0$ of the pivot α of $A :: C$. Then, by Theorem 6 (with a as α^* , A^1 as B , and α^1 as β), the pivot α^1 of $A^1 :: C$ has order $r - 1$. Likewise, for $k = 1, \dots, r - 1$, the matrix A^{k+1} is the pivotal transform of A^k by any selected entry $a^k \neq 0$ of the pivot α^k of $A^k :: C$. Then, by Theorem 6 (with A^k as A , a^k as α^* , A^{k+1} as B , and α^{k+1} as β), the pivot α^{k+1} of $A^{k+1} :: C$ has order $r - k - 1$. Finally, the pivot α^r of $A^r :: C$ has order zero, i.e. $\alpha^r = \phi$. Hence C belongs to the permutation subclass $\{A^r\}_\phi$ and we can pass from A^r to C by a finite succession of elementary operations of types (i) and (ii). This completes the proof of Theorem 7.

THEOREM 8. *If A is a nonsingular square matrix, then $A :: A^{-1}$. In particular, A^{-1} is the pivotal transform of A by $\alpha^* = A$.*

Proof. Take P in (2) so that $[I, A]P = [A, I]$. Then (3) becomes $[A, I] = A[I, A^{-1}]$. So $A :: A^{-1}$. Moreover, (6) and (7) reduce to

$$B = B_{11} = A_{11}^{-1} \quad \text{and} \quad A = A_{11} = B_{11}^{-1},$$

since the remaining $A_{\mu\nu}$ and $B_{\mu\nu}$ are vacuous. Hence A^{-1} is the pivotal transform of A by $\alpha^* = A$.

We note that the inversion of a nonsingular square matrix of order n can be performed efficiently by the elementary operations of Theorem 7, using n operations of type (iii) at most. (These pivotal operations involve a total of n^3 individual steps of division and multiplication at most.)

Remarks (added in proof). M. M. Flood has observed that $A :: B$ if, and only if,

$$[I, A]P \begin{bmatrix} -B \\ J \end{bmatrix} = 0,$$

where I and J are identity matrices of orders m and n , and P is a permutation matrix of order $m + n$. In particular, this yields a simple proof of Theorem 5.

The relation $s = r - r^*$ in Theorem 6 generalizes to $s = r + r^* - t$ for arbitrary α and $\beta \leftrightarrow \alpha$, where t denotes the total number of rows and columns of A in each of which there are entries of α and entries of the pivot α^* . (If α contains α^* , then $t = 2r^*$ and s reduces to $r - r^*$.)

PRINCETON UNIVERSITY,
PRINCETON, NEW JERSEY

LINEAR INEQUALITIES AND THE PAULI PRINCIPLE¹

BY

HAROLD W. KUHN

Introduction. The basic physical problem that underlies this paper is the energy eigenvalue problem of an n -fermion assembly under r -body interaction. In a series of papers, S. Watanabe [1] has shown that this can be reduced to an extremum problem involving an r -body density-matrix provided that the auxiliary condition, corresponding to the Pauli Exclusion Principle, can be formulated in terms of the r -body density-matrix. For the case of a one-body density-matrix, he also gave a necessary condition which seemed, on physical grounds, to be sufficient. Recently Dr. Watanabe suggested that sufficiency of the condition was equivalent to the solvability of a certain system of linear inequalities. The purpose of this paper is to present four proofs of the sufficiency of the condition for a one-body density-matrix. In addition, the problem of a two-body density-matrix is formulated and certain partial results described. The discussion is purely mathematical and starts from the formulation given to the physical problem by Watanabe.

The one-body problem. The one-body problem can be given three, essentially equivalent, statements; each is suggested by and is suited to a particular approach to the proof. These will be presented separately, and then their interrelations investigated.

I. Urn Problem. Suppose an urn contains balls of g colors which have been placed in the urn in batches of n balls at a time, where $n \leq g$, and no batch contains two balls of the same color. Let n_i be the number of balls of color i , for $i = 1, \dots, g$, and let $N = n_1 + \dots + n_g$. What urn compositions (n_1, \dots, n_g) are possible?

SOLUTION. The composition (n_1, \dots, n_g) is possible if, and only if, n divides N and $nn_i \leq N$ for $i = 1, \dots, g$.

Proof. A proof of the sufficiency of the conditions will be given by induction on $b = N/n$, the number of batches of balls used to fill the urn. (The necessity is clear, since $N = nb$ and no color can appear more often than b times. That is, $n_i \leq b = N/n$ or $nn_i \leq N$ for all i .)

If $b = 1$, the conditions require $nn_i \leq N = nb = n$. That is, $0 \leq n_i \leq 1$ and $n_1 + \dots + n_g = n$. Thus the conditions imply that the urn contains n balls of distinct colors, which make up the single batch. Suppose that sufficiency has been proved for $b - 1$ and consider an urn composed of nb

¹ The preparation of this paper was supported, in part, by the Princeton-IBM Mathematics Research Project, Princeton University.

balls satisfying the conditions. Let the colors be indexed so that $n_1 \geq n_2 \geq \dots \geq n_g$. There are surely n distinct colors in the urn since otherwise $n_1 + \dots + n_g = n_1 + \dots + n_{n-1} \leq (n-1)n_1 < nn_1 \leq N$. Remove a batch consisting of balls of colors $1, \dots, n$. Since at most n colors can appear as often as b times in the urn, $n_i \leq b-1$ for $i = n+1, \dots, g$. Hence

$$(1) \quad nn_i \leq n(b-1) = n\left(\frac{N}{n} - 1\right) = N - n$$

for $i = n+1, \dots, g$. On the other hand, the new frequency for the color of a ball which has been removed is $n_i - 1$. Hence

$$(2) \quad n(n_i - 1) = nn_j - n \leq N - n$$

for $i = 1, \dots, n$. Therefore the conditions are satisfied for all colors in the new urn composed of $N - n = n(b-1)$ balls and the Urn Problem is solved.

II. Polyhedron Problem. Let ν denote an arbitrary subset of n distinct indices $\{i_1, \dots, i_n\}$ chosen from $\{1, \dots, g\}$. Let $X^\nu = (x_i^\nu)$ denote the point in g -dimensional affine space defined by $x_i^\nu = 1/n$ if $i \in \nu$ and $x_i^\nu = 0$ otherwise, for $i = 1, \dots, g$. Let C denote the convex hull of the points X^ν . Characterize C as the intersection of halfspaces.

SOLUTION. $C = \{X = (x_i) | 0 \leq x_i \leq 1/n \text{ and } x_1 + \dots + x_g = 1\}$.

Proof. Let $D = \{X = (x_i) | 0 \leq x_i \leq 1/n \text{ and } x_1 + \dots + x_g = 1\}$.

Since D is a compact convex set, it is the convex hull of its extreme points. It is clear that X^ν is an extreme point of D for every ν . To show that every extreme point of D is an X^ν , we shall show that every extreme point of D has no more than n positive components. Hence, to sum to one, there are exactly n positive components, all equal to $1/n$ and the extreme point is an X^ν .

Suppose $X = (x_1, x_2, \dots, x_g)$ is extreme in D and has $n+1$ (or more) positive components. Hence, at least two positive components are less than $1/n$, say,

$$(3) \quad 0 < x_1 < 1/n \quad \text{and} \quad 0 < x_2 < 1/n.$$

Then, for small enough ϵ ,

$$(4) \quad X = 1/2(x_1 + \epsilon, x_2 - \epsilon, x_3, \dots, x_g) + 1/2(x_1 - \epsilon, x_2 + \epsilon, x_3, \dots, x_g)$$

exhibits X as the midpoint of two distinct points of D , and proves that X is not extreme. This completes the solution of the Polyhedron Problem.

This solution can be sharpened somewhat by following a parallel to the solution of the Urn Problem. For $X = (x_i) \in D$, let $d(X)$ be the number of components x_i such that $0 < x_i < 1/n$, if there are any such. It is clear that $d(X) \neq 1$ under this definition. We shall assign $d(X) = 1$ if all of the components of X are equal to 0 or $1/n$, i.e., if X is an X^ν .

SHARPENED SOLUTION. Let $d = d(X)$ for $X \in D$. Then there exist $\lambda_1 \geq 0, \dots, \lambda_d \geq 0$ with $\lambda_1 + \dots + \lambda_d = 1$ such that

$$(5) \quad X = \lambda_1 X^{\nu_1} + \dots + \lambda_d X^{\nu_d}$$

for a suitable set of extreme points $X^{\nu_1}, \dots, X^{\nu_d}$.

It will be convenient to state and prove a corollary dealing with rational X at the same time.

COROLLARY. Let $R = (r_1, \dots, r_g) \in D$ with $d = d(R)$ and all components r_i rational. If b is an integer such that bnr_i is integral for all i , then there exists a decomposition

$$(6) \quad R = \lambda_1 X^{\nu_1} + \dots + \lambda_d X^{\nu_d}$$

with all $\lambda_i \geq 0$, $\lambda_1 + \dots + \lambda_d = 1$ and $\lambda_i b$ integral for all i .

Proof. The proof of the Sharpened Solution is made by induction on $d(X)$. It is obvious for $d(X) = 1$; suppose that it has been proved for $d(Z) < d = d(X)$ and let the components of X be indexed so that $x_1 \geq x_2 \geq \dots \geq x_g$. Since $d(X) \geq 2$, $1/n > x_n \geq x_{n+1} > 0$ and

$$\lambda = \min(nx_n, 1 - nx_{n+1})$$

satisfies $0 < \lambda < 1$. The decomposition $X = \lambda Y + (1 - \lambda)Z$ defined by

$$(7) \quad Y = (1/n, \dots, 1/n, 0, \dots, 0),$$

and

$$(8) \quad Z = \left(\frac{nx_1 - \lambda}{n(1 - \lambda)}, \dots, \frac{nx_n - \lambda}{n(1 - \lambda)}, \frac{x_{n+1}}{1 - \lambda}, \dots, \frac{x_g}{1 - \lambda} \right)$$

then holds identically.

By the choice of λ , for $i = 1, \dots, n$,

$$(9) \quad z_i = \frac{nx_i - \lambda}{n(1 - \lambda)} \geq \frac{nx_n - \lambda}{n(1 - \lambda)} \geq 0,$$

while $x_i \leq 1/n$ implies

$$(10) \quad z_i = \frac{nx_i - \lambda}{n(1 - \lambda)} \leq \frac{1 - \lambda}{n(1 - \lambda)} = \frac{1}{n}.$$

By the choice of λ , for $i = n + 1, \dots, g$,

$$(11) \quad z_i = \frac{x_i}{1 - \lambda} \leq \frac{x_{n+1}}{1 - \lambda} \leq \frac{x_{n+1}}{nx_{n+1}} = \frac{1}{n},$$

while $x_i \geq 0$ implies

$$(12) \quad z_i = \frac{x_i}{1 - \lambda} \geq 0.$$

Hence $Z \in D$. On the other hand, it is easily verified that, if $x_i = 0$ (or $1/n$) then $z_i = 0$ (or $1/n$), while $z_n = 0$ if $\lambda = nx_n$ and $z_{n+1} = 1/n$ if $\lambda = 1 - nx_{n+1}$. Hence $d(Z) < d(X)$ and the Sharpened Solution is complete.

The corollary is also proved by an induction on $d(R)$. If $d(R) = 1$ then R is an X^ν , any integer may play the rôle of b , and $R = 1 \cdot X^\nu$ validates the conclusion. Suppose the corollary proved for all Z with $d(Z) < d = d(R)$, and decompose R as in the Sharpened Solution,

$$(13) \quad R = \lambda Y + (1 - \lambda)Z,$$

where $\lambda = \min(nr_n, 1 - nr_{n+1})$. If bnr_i is integral for all i , then $b\lambda$ is also. Hence, by the definition of Z , $(1 - \lambda)bnz_i$ is integral for all i and the induction hypothesis may be applied to Z with $(1 - \lambda)b$ playing the rôle of b . Hence,

$$(14) \quad Z = \lambda'_2 X^{\nu_2} + \cdots + \lambda'_d X^{\nu_d}$$

with $\lambda'_i(1 - \lambda)b$ integral for $i = 2, \dots, d$. Therefore,

$$(15) \quad R = \lambda_1 X^{\nu_1} + \cdots + \lambda_d X^{\nu_d},$$

with $\lambda_1 = \lambda$, $\lambda_i = (1 - \lambda)\lambda'_i$ for $i = 2, \dots, d$, and $X^{\nu_1} = Y$. Furthermore, $\lambda_i b$ is integral for all i and so the corollary is proved.

III. Solvability Problem. Let ν be as before, and let g constants $x_i \geq 0$, $x_1 + \cdots + x_g = 1$ be given. Under what conditions do the g equations

$$(16) \quad \sum_{\nu, i \in \nu} \lambda_\nu = nx_i \quad (i = 1, \dots, g)$$

have a non-negative solution for the unknowns λ_ν ?

SOLUTION. Necessary and sufficient conditions are $x_i \leq 1/n$ for $i = 1, \dots, g$.

Proof. To prove the necessity, note that each λ_ν appears exactly n times on the left side of the system. Hence,

$$(17) \quad n \sum_\nu \lambda_\nu = n \sum_i x_i = n$$

and

$$(18) \quad nx_i = \sum_{\nu, i \in \nu} \lambda_\nu \leq \sum_\nu \lambda_\nu = 1 \quad (i = 1, \dots, g).$$

To prove the sufficiency, we appeal to a criterion for the solvability of a system of linear inequalities [2]. Either the system

$$(19) \quad \lambda_\nu \geqq 0 \quad (\text{all } \nu),$$

$$(20) \quad \sum_{\nu, i \in \nu} \lambda_\nu = nx_i \quad (i = 1, \dots, g)$$

is solvable, or there exist w_i , $i = 1, \dots, g$, such that

$$(21) \quad \sum_{i \in \nu} w_i \leq 0 \quad (\text{all } \nu)$$

and

$$(22) \quad \sum_i w_i x_i > 0,$$

but not both systems are solvable. We shall show that, under the conditions $x_i \leq 1/n$, the second system is not solvable.

LEMMA. *An extreme (ray) solution of*

$$(23) \quad \sum_{i \in \nu} w_i \leq 0 \quad (\text{all } \nu)$$

is either the negative of a unit vector or has at least $g - n + 1$ components $w_i = \min_i w_i$.

Proof. Let $W \neq 0$ be an extreme (ray) solution and let $w_1 = \min_i w_i$, reindexing if necessary. Then $w_1 < 0$ since, by adding all of the inequalities of the system,

$$(24) \quad w_1 + \cdots + w_g \leq 0.$$

Consider the sets ν such that $1 \in \nu$ and $w_i \neq 0$ for some $i \in \nu$ such that $i \neq 1$. (If the given solution is not the negative of a unit vector, there must exist at least one such.) For at least one of these,

$$(25) \quad \sum_{i \in \nu} w_i = 0,$$

since otherwise the given solution is the sum of the vectors

$$(26) \quad 1/2(w_1 \pm \epsilon, w_2, \dots, w_g)$$

which are distinct (as rays) and are solutions for small enough $\epsilon > 0$. Let $\nu' = (\nu - \{1\}) \cup \{j\}$ for $j \notin \nu$. Then

$$(27) \quad 0 \geq \sum_{i \in \nu'} w_i = \sum_{i \in \nu} w_i - w_1 + w_j \geq \sum_{i \in \nu} w_i = 0,$$

and

$$(28) \quad w_j = w_1 = \min_i w_i \quad (\text{all } j \notin \nu).$$

This completes the proof of the lemma.

To complete the solution of the Solvability Problem, let $W = (w_i)$ be any extreme ray solution and let $X = (x_i)$ be any point in D . If W is the negative of a unit vector, then we have

$$(29) \quad w_1 x_1 + \cdots + w_g x_g \leq 0$$

trivially, since $x_i \geq 0$ for all i . Otherwise, assume that $w_{n+1} = \cdots = w_g = \min_i w_i = m$ (note that this is somewhat weaker than the lemma). Then

$$\begin{aligned} (30) \quad w_1 x_1 + \cdots + w_g x_g &= (w_1 - m)x_1 + \cdots + (w_n - m)x_n \\ &\quad + mx_1 + \cdots + mx_g \\ &\leq [(w_1 - m) + \cdots + (w_n - m)] 1/n + m \\ &= (w_1 + \cdots + w_n) 1/n - m + m \\ &\leq 0. \end{aligned}$$

This proves that the system of inequalities (21) and (22) is not solvable and completes the proof of III.

The connections between the three formulations of the one-body problem are close and almost obvious. Since, by (17), in any non-negative solution to (16) necessarily $\sum \lambda_v = 1$, the Solvability Problem asks: Which X can be expressed as convex combinations of the points X^v ? This is exactly the same as the Polyhedron Problem and receives the same answer, namely, exactly the points of D .

The proofs given for the Polyhedron Problem and the Solvability Problem also have a close geometrical connection. The polyhedron $C = D$ has a representation (certainly redundant in some cases, such as $g = 3, n = 2$) as the intersection of the hyperplane $H = \{X | x_1 + \dots + x_g = 1\}$ with the $2g$ half spaces $x_j \geq 0$ and $nx_j \leq 1$ for $j = 1, \dots, g$. Define

$$(31) \quad \hat{C} = \{Y | Y = tX \text{ for } X \in C \text{ and } t \geq 0\}.$$

Then \hat{C} is a polyhedral cone with $C = \hat{C} \cap H$. Since $C = D$, the extreme faces of \hat{C} are clearly contained among the half spaces

$$(32) \quad -y_j \leq 0 \quad (j = 1, \dots, g)$$

and

$$(33) \quad -y_1 - y_2 - \dots - (1 - n)y_j - \dots - y_g \leq 0 \quad (j = 1, \dots, g).$$

Define \hat{C}^* , the polar cone of \hat{C} , by

$$(34) \quad \hat{C}^* = \{W | w_1 y_1 + \dots + w_g y_g \leq 0 \text{ for all } Y \in \hat{C}\}.$$

Then the extreme rays of \hat{C}^* are contained among the vectors

$$(35) \quad U^j = (u_i) \text{ where } u_i = \begin{cases} -1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } j = 1, \dots, g,$$

and

$$(36) \quad V^j = (v_i) \text{ where } v_i = \begin{cases} n - 1 & \text{if } i = j, \\ -1 & \text{otherwise,} \end{cases} \quad \text{for } j = 1, \dots, g.$$

The extreme (ray) solutions of system (23) must appear in this set; this provides a sharper version of the lemma used in the solution of the Solvability Problem.

The Urn Problem may be regarded as a discrete version of either of these problems. The passage from I to II or III is made by an obvious limiting process, first approximating (x_1, \dots, x_g) by rational (r_1, \dots, r_g) and then constructing an integral urn composition $(n_1, \dots, n_g) = (bnr_1, \dots, bnr_g)$ with a suitably chosen b . The reverse implication is obtained from the corollary to the sharpened solution to the Polyhedron Problem. Let (n_1, \dots, n_g) be an urn composition such that n divides $N = n_1 + \dots + n_g$ with integral quotient b and $nn_i \leq N$ for all i . Set $r_i = n_i/nb$ for $i = 1, \dots, g$. Then $R = (r_1, \dots, r_g)$ satisfies the hypotheses of the corollary. By (6),

$$(37) \quad R = \lambda_1 X^{r_1} + \dots + \lambda^g X^{r_g},$$

with $\lambda_i b$ integral for $i = 1, \dots, d$. Hence the urn can be made up by inserting $\lambda_i b$ batches of composition ν_i for $i = 1, \dots, d$.

The two-body problem. The two-body problem can also be given three statements. Although attacks have been made on all three fronts, the problem seems to be much harder and only very meager results are available. To place the results known at present on record, we shall formulate the problem only in its polyhedron form.

Polyhedron Problem. Let ν denote an arbitrary subset of n distinct indices $\{i_1, \dots, i_n\}$ chosen from $\{1, \dots, g\}$. Let $X^\nu = (x_{ij}^\nu)$, where $1 \leq i < j \leq g$, denote the point in $g(g - 1)/2$ -dimensional affine space defined by $x_{ij}^\nu = 2/n(n - 1)$ if $i \in \nu$ and $j \in \nu$, and $x_{ij}^\nu = 0$ otherwise. Let C denote the convex hull of the points X^ν . Characterize C as the intersection of halfspaces.

In the following table, distinct indices are to take on all possible sets of distinct values. Except in the case $g = 6, n = 3$, the sets of constraints are known to be complete.² In all cases, $\sum_{1 \leq i < j \leq g} x_{ij} = 1$ is to hold.

g	n	Faces of C
4	3	$x_{ij} + x_{ik} + x_{il} \leq 2/3$ $x_{ij} + x_{kl} = x_{ik} + x_{jl}$
5	3	$x_{ik} + x_{il} + x_{im} + x_{jk} + x_{jl} + x_{jm} \leq 2/3$
5	4	$x_{ij} + x_{ik} + x_{il} + x_{im} \leq 1/2$ $x_{ij} + x_{kl} = x_{ik} + x_{jl}$
6	3	$x_{ij} \geq 0$ $x_{ik} + x_{il} + x_{im} + x_{in} + x_{jk} + x_{jl} + x_{jm} + x_{jn} \leq 2/3$ $x_{il} + x_{im} + x_{in} + x_{jl} + x_{jm} + x_{jn} + x_{kl} + x_{km} + x_{kn} \leq 2/3$ $x_{ij} \leq x_{ik} + x_{il} + x_{im} + x_{in}$

REFERENCES

1. S. Watanabe, Zeitschrift für Physik vol. 113 (1939) p. 482; Kagaku (in Japanese) vol. 14 (1944) pp. 82, 122 and 169.

2. H. W. Kuhn, Amer. Math. Monthly vol. 63 (1956) pp. 217–232.

BRYN MAWR COLLEGE,
BRYN MAWR, PENNSYLVANIA

² Added in proof, October 23, 1959. The constraints given for $g = 6, n = 3$ are now known to be complete. Details and additional results will be published elsewhere.

This page intentionally left blank

COMPOUND AND INDUCED MATRICES IN COMBINATORIAL ANALYSIS

BY

H. J. RYSER

1. Introduction. The present work is an outgrowth of a recent paper by the author entitled *Inequalities of compound and induced matrices with applications to combinatorial analysis* [21]. We study inequalities involving the elementary symmetric functions and the homogeneous product sums of the characteristic roots of a nonnegative hermitian matrix and the applications of these inequalities to problems in combinatorial analysis.

Let a be a complex number and let \bar{a} be the complex conjugate of a . Let $A = [a_{rs}]$ be a matrix with elements in the complex field. Throughout the discussion \bar{A} denotes the matrix $[\bar{a}_{rs}]$, A^T the transpose of A , $\det(A)$ the determinant of A , $\text{per}(A)$ the permanent of A , A^{-1} the inverse of A for $\det(A) \neq 0$, $\text{tr}(A)$ the trace of A , $C_r(A)$ the r th compound or adjugate of A , $P_r(A)$ the r th induced or power matrix of A . I denotes the identity matrix, and S denotes the matrix all of whose entries are 1's.

In §2 we summarize the pertinent literature on compound and induced matrices. §3 contains the main inequalities. The theorems in this section are generalizations of those appearing in [21]. Theorem 3.5 illustrates the type of result studied [21]. Let H be a nonnegative hermitian matrix of order v . Let

$$(1.1) \quad \text{tr}(H) = k^*v,$$

$$(1.2) \quad SHS = \mu S,$$

$$(1.3) \quad \mu = (k^* + (v - 1)\lambda^*)v,$$

where S is the v by v matrix of 1's. Now define the matrix B^* of order v by

$$(1.4) \quad B^* = (k^* - \lambda^*)I + \lambda^*S.$$

Then Theorem 3.5 asserts that

$$(1.5) \quad \text{tr}(C_r(H)) \leq \text{tr}(C_r(B^*)) \quad (1 \leq r \leq v).$$

Equality holds for $r = 1$. If $k^* + (v - 1)\lambda^* \neq 0$ and equality holds for an r ,

$$1 < r \leq v,$$

or if $k^* + (v - 1)\lambda^* = 0$ and equality holds for an r ,

$$1 < r < v,$$

then

$$H = B^*.$$

An analogous result holds for the r th induced or power matrix $P_r(H)$ of H , where for this case $\text{tr}(P_r(H)) \geq \text{tr}(P_r(H^*))$. Theorem 3.5 is applied in §4 to various algebraic topics. These include Vandermonde matrices, incidence matrices of oriented graphs, and incidence matrices of v, k, λ configurations.

2. Compound and induced matrices. Compound and induced matrices have been studied very extensively over a number of years. We confine ourselves here to the definitions of these matrices and a listing of some of their essential properties. The following references may be consulted for more detailed information [2; 11; 13; 14; 22; 26].

Let A be an n by n matrix with elements in the complex field. Let r be an integer such that $1 \leq r \leq n$, and let $\{n_r\}$ be the collection of all subsets of r elements chosen from the set $1, \dots, n$. Let the elements of $\{n_r\}$ be denoted by $\sigma_1, \dots, \sigma_N$, where the σ 's are ordered lexicographically and

$$N = \binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!}.$$

If in the matrix A all rows are deleted whose indices do not belong to σ_i and all columns are deleted whose indices do not belong to σ_j , then there remains an r by r submatrix of A , which we denote by $A_{\sigma_i \sigma_j}$. We define the r th compound or the r th adjugate of A by

$$C_r(A) = [\det(A_{\sigma_i \sigma_j})] \quad (i, j = 1, \dots, N).$$

The matrix $C_r(A)$ of order N satisfies:

$$(2.1) \quad C_r(A)C_r(B) = C_r(AB) \quad (A \text{ and } B \text{ of order } n),$$

$$(2.2) \quad C_r(A^T) = (C_r(A))^T,$$

$$(2.3) \quad C_r(A^{-1}) = (C_r(A))^{-1} \quad (\det(A) \neq 0),$$

$$(2.4) \quad \det(C_r(A)) = (\det(A))^M \quad \left(M = \binom{n-1}{r-1} \right).$$

Moreover, if $\alpha_1, \dots, \alpha_n$ are the n characteristic roots of A , then the characteristic roots of $C_r(A)$ are the N products of the α_i 's taken r at a time. If

$$a_r(A) = \sum_{\sigma} \det(A_{\sigma\sigma}),$$

where σ runs through $\{n_r\}$, then the $a_r(A)$'s are the coefficients, apart from the signs, of the characteristic polynomial of A . Indeed,

$$(2.5) \quad \prod_{i=1}^n (x + \alpha_i) = x^n + a_1(A)x^{n-1} + \cdots + a_n(A),$$

whence it follows that

$$(2.6) \quad a_r(A) = \text{tr}(C_r(A)) \quad (r = 1, \dots, n).$$

In particular, note that

$$(2.7) \quad a_1(A) = \text{tr}(A),$$

$$(2.8) \quad a_n(A) = \det(A).$$

Now let r be an arbitrary positive integer, and let

$$(2.9) \quad y_i = a_{i_1}x_1 + \cdots + a_{i_n}x_n \quad (i = 1, \dots, n),$$

where x_i and y_i are indeterminates. We form the

$$N^* = \binom{n+r-1}{r}$$

products of the y_i 's

$$(2.10) \quad y_1^{\gamma_1}y_2^{\gamma_2}\cdots y_n^{\gamma_n},$$

where $\sum \gamma_i = r$. We agree to order the products (2.10) lexicographically in the sense that the product $y_1^{\gamma_1}y_2^{\gamma_2}\cdots y_n^{\gamma_n}$ stands before the product $y_1^{\delta_1}y_2^{\delta_2}\cdots y_n^{\delta_n}$ provided that the first nonvanishing difference $\gamma_1 - \delta_1, \gamma_2 - \delta_2, \dots, \gamma_n - \delta_n$ is positive. Denote the products (2.10) written in this order by

$$Y_1, Y_2, \dots, Y_{N^*},$$

and denote the corresponding products of the x_i 's written in the same order by

$$X_1, X_2, \dots, X_{N^*}.$$

Let X be the column vector with components X_1, \dots, X_{N^*} , and let Y be the column vector with components Y_1, \dots, Y_{N^*} . Then by (2.9) we may write

$$Y = P_r(A)X,$$

where $P_r(A)$ is a matrix of order N^* . This matrix is called the r th *induced matrix* or *power matrix* of A . Note that $P_r(A)$ is defined for every positive integer r . Induced matrices satisfy:

$$(2.11) \quad P_r(A)P_r(B) = P_r(AB) \quad (A \text{ and } B \text{ of order } n),$$

$$(2.12) \quad P_r(A^{-1}) = (P_r(A))^{-1} \quad (\det(A) \neq 0),$$

$$(2.13) \quad \det(P_r(A)) = (\det(A))^{M^*} \quad \left(M^* = \binom{n+r-1}{n} \right).$$

Also if $\alpha_1, \dots, \alpha_n$ are the characteristic roots of A , then the characteristic roots of $P_r(A)$ are the N^* products of the α_i 's of degree r . Thus if

$$\frac{1}{\sum_{i=1}^n (1 - \alpha_i x)} = 1 + h_1 x + \cdots + h_r x^r + \cdots,$$

then for every positive integer r ,

$$(2.14) \quad h_r = \text{tr}(P_r(A)).$$

3. The main inequalities. In this section we generalize the inequalities in [21] that involve the trace functions $\text{tr}(C_r(H))$ and $\text{tr}(P_r(H))$ of a non-negative hermitian matrix H . We begin with the concept of majorization for polynomials and formal power series. If $f(x) = \sum a_i x^i$ and $g(x) = \sum b_i x^i$ are polynomials of degree n , where the coefficients a_i and b_i are nonnegative reals, and if

$$a_i \leq b_i \quad (i = 0, 1, \dots, n),$$

then $f(x)$ is *majorized* by $g(x)$, written

$$(3.1) \quad f \prec g \quad \text{or} \quad g \succ f.$$

Similarly if $f(x) = \sum a_i x^i$ and $g(x) = \sum b_i x^i$ are formal power series, we write $f \prec g$ or $g \succ f$ provided

$$0 \leq a_i \leq b_i \quad (i = 0, 1, 2, \dots).$$

Note that $f \prec g$ and $f_1 \prec g_1$ imply $ff_1 \prec gg_1$.

We state four lemmas derived in [21] that are basic to the inequalities on trace functions that follow. Lemma 3.1 is trivial and Lemma 3.2 is a well-known inequality that is an easy consequence of Lemma 3.1.

LEMMA 3.1. *If $\alpha \geq \beta \geq 0$ and $\epsilon \geq 0$, then*

$$(x + \alpha + \epsilon)(x + \beta) \prec (x + \alpha)(x + \beta + \epsilon).$$

Equality holds for the coefficients of x^2 and x . Equality holds for the coefficient of x^0 if and only if $\alpha = \beta$ or $\epsilon = 0$.

LEMMA 3.2. *If $e = (\alpha_1 + \dots + \alpha_n)/n$ and $\alpha_i \geq 0$, then*

$$\prod_{i=1}^n (x + \alpha_i) \prec (x + e)^n.$$

Equality holds for the coefficients of x^n and x^{n-1} . If equality holds for one of the other coefficients, then each $\alpha_i = e$, and equality holds throughout.

LEMMA 3.3. *If $\alpha \geq \beta \geq 0$ and $\epsilon \geq 0$, then*

$$\frac{1}{(1 - (\alpha + \epsilon)x)(1 - \beta x)} \succ \frac{1}{(1 - \alpha x)(1 - (\beta + \epsilon)x)}.$$

Equality holds for the coefficients of x^0 and x . If equality holds for one of the other coefficients, then $\alpha = \beta$ or $\epsilon = 0$, and equality holds throughout.

LEMMA 3.4. *If $e = (\alpha_1 + \dots + \alpha_n)/n$ and $\alpha_i \geq 0$, then*

$$\frac{1}{\prod_{i=1}^n (1 - \alpha_i x)} \succ \frac{1}{(1 - ex)^n}.$$

Equality holds for the coefficients of x^0 and x . If equality holds for one of the other coefficients, then each $\alpha_i = e$, and equality holds throughout.

Now let H be a nonnegative hermitian matrix of order v and rank e . Let the characteristic roots of H be $\lambda_1, \dots, \lambda_v$, where

$$(3.2) \quad \lambda_1 \geq \dots \geq \lambda_e > \lambda_{e+1} = \dots = \lambda_v = 0.$$

Let h be an integer ≥ 1 such that

$$(3.3) \quad e \leq h \leq v,$$

and define k and λ by

$$(3.4) \quad \text{tr}(H) = kh,$$

$$(3.5) \quad \lambda_h \leq k + (h - 1)\lambda \leq \lambda_1.$$

Define the matrix B of order h by

$$B = (k - \lambda)I + \lambda S,$$

where I is the identity matrix and S is the matrix all of whose entries are 1's. Finally let

$$(3.6) \quad B_0 = B + 0,$$

where B_0 of order v is the direct sum of the matrix B of order h and the zero matrix of order $v - h$.

The characteristic polynomial of B is easily computed by subtracting column one of $\det(xI - B)$ from each of the other columns, and then adding to row one each of the remaining rows. Thus

$$\det(xI - B_0) = x^{v-h}(x - (k + (h - 1)\lambda))(x - (k - \lambda))^{h-1},$$

and hence the v characteristic roots of B_0 are $k + (h - 1)\lambda$ taken once, $k - \lambda$ taken $h - 1$ times, and 0 taken $v - h$ times. This allows us to evaluate $\text{tr}(C_r(B_0))$ and $\text{tr}(P_r(B_0))$ explicitly. Evidently

$$(3.7) \quad \text{tr}(C_r(B_0)) = \binom{h}{r} (k + (r - 1)\lambda)(k - \lambda)^{r-1}$$

and

$$(3.8) \quad \text{tr}(P_r(B_0)) = \sum_{i=0}^r \binom{h+i-2}{i} (k + (h - 1)\lambda)^{r-i} (k - \lambda)^i.$$

Note that $\text{tr}(C_r(B_0)) = 0$ for $r > h$.

THEOREM 3.1. *The matrices H and B_0 satisfy*

$$\text{tr}(C_r(H)) \leq \text{tr}(C_r(B_0)) \quad (1 \leq r \leq v).$$

Equality holds for $r = 1, h + 1, \dots, v$. If $k + (h - 1)\lambda \neq 0$ and equality holds for an r ,

$$1 < r \leq h,$$

or if $k + (h - 1)\lambda = 0$ and equality holds for an r ,

$$1 < r < h,$$

then there exists a unitary U such that

$$H = U^{-1}B_0U.$$

Let $h > 1$ and in Lemma 3.1 set $\epsilon = \lambda_1 - (k + (h - 1)\lambda)$, $\alpha = \lambda_1 - \epsilon$, and $\beta = \lambda_h$. Then

$$(x + \lambda_1)(x + \lambda_h) \prec (x + k + (h - 1)\lambda)(x + \lambda_h + \epsilon),$$

and

$$\begin{aligned} (\lambda_h + \epsilon + \lambda_2 + \cdots + \lambda_{h-1})/(h - 1) &= (kh - \lambda_1 + \epsilon)/(h - 1) \\ &= k - \lambda. \end{aligned}$$

Hence by Lemma 3.2,

$$\begin{aligned} (3.9) \quad (x + \lambda_1)(x + \lambda_h) \prod_{i=2}^{h-1} (x + \lambda_i) \\ &\prec (x + k + (h - 1)\lambda)(x + \lambda_h + \epsilon) \prod_{i=2}^{h-1} (x + \lambda_i) \\ &\prec (x + k + (h - 1)\lambda)(x + (k - \lambda))^{h-1}, \end{aligned}$$

whence the first conclusion follows.

Suppose that $k + (h - 1)\lambda \neq 0$ and that equality holds in the theorem for an r , $1 < r \leq h$. Then equality holds throughout (3.9) for some coefficient of x^r , where $r \neq h$, $h - 1$. But then equality holds for some coefficient of x^r in

$$(x + \lambda_h + \epsilon) \prod_{i=2}^{h-1} (x + \lambda_i) \prec (x + (k - \lambda))^{h-1},$$

where $r \neq h - 1$, $h - 2$. By Lemma 3.2,

$$\lambda_2 = \cdots = \lambda_{h-1} = \lambda_h + \epsilon = k - \lambda.$$

Then for $k - \lambda > 0$ and also for $k - \lambda = 0$,

$$\lambda_1\lambda_h = (k + (h - 1)\lambda)(k - \lambda)$$

and

$$\lambda_1 + \lambda_h = (k - \lambda) + (k + (h - 1)\lambda).$$

Hence $\lambda_1 = k + (h - 1)\lambda$ and $\lambda_h = k - \lambda$, or $\lambda_1 = k - \lambda$ and $\lambda_h = k + (h - 1)\lambda$. Thus the characteristic roots of H are $k + (h - 1)\lambda$ taken once, $k - \lambda$ taken $h - 1$ times, and 0 taken $v - h$ times. This means H and B_0 have the same characteristic roots, and hence there exists a unitary U such that

$$H = U^{-1}B_0U.$$

Suppose that $k + (h - 1)\lambda = 0$ and that equality holds in the theorem for an r , $1 < r < h$. Then $\lambda_h = 0$ and equality holds for some coefficient of x^r in

$$x \prod_{i=1}^{h-1} (x + \lambda_i) \prec x(x + (k - \lambda))^{h-1},$$

where $r \neq h$, $h = 1$, or 0 . But then equality holds for some coefficient of x^r in

$$\prod_{i=1}^{h-1} (x + \lambda_i) \prec (x + (k - \lambda))^{h-1},$$

where $r \neq h - 1, h - 2$. But then by Lemma 3.2, $\lambda_1 = \dots = \lambda_{h-1} = k - \lambda$. Hence the characteristic roots of H are $k - \lambda$ taken $h - 1$ times, and 0 taken $v - h + 1$ times. Thus H and B_0 have the same characteristic roots, and hence there exists a unitary U such that

$$H = U^{-1}B_0U.$$

Suppose now that in Theorem 3.1 we set $h = v$. Let

$$(3.10) \quad \text{tr}(H) = k'v,$$

$$(3.11) \quad \lambda_v \leq k' + (v - 1)\lambda' \leq \lambda_1,$$

and define the matrix B' of order v by

$$(3.12) \quad B' = (k' - \lambda')I + \lambda'S.$$

Then we obtain [21]

THEOREM 3.2. *The matrices H and B' satisfy*

$$\text{tr}(C_r(H)) \leq \text{tr}(C_r(B')) \quad (1 \leq r \leq v).$$

Equality holds for $r = 1$. If $k' + (v - 1)\lambda' \neq 0$ and equality holds for an r ,

$$1 < r \leq v,$$

or if $k' + (v - 1)\lambda' = 0$ and equality holds for an r ,

$$1 < r < v,$$

then there exists a unitary U such that

$$H = U^{-1}B'U.$$

Let A be a matrix of order v and rank $e > 0$. Then

$$(3.13) \quad A\bar{A}^T = H$$

is nonnegative hermitian of rank e and

$$(3.14) \quad \text{tr}(H) = \|A\|^2,$$

where $\|A\|$ denotes the norm of A . In Theorem 3.1 it is permissible to set $h = e$ and $\lambda = 0$. Then (3.4) and (3.6) imply

$$(3.15) \quad k = \frac{\text{tr}(H)}{e} = \frac{\|A\|^2}{e},$$

$$(3.16) \quad B_0 = \frac{\|A\|^2}{e} I + 0.$$

Here I is the identity matrix of order e , and 0 is the zero matrix of order $v - e$. This yields an inequality of Marcus [15].

THEOREM 3.3. *Let A be a matrix of order v and rank $e > 0$. Then*

$$\text{tr}(C_r(A\bar{A}^T)) \leq \binom{e}{r} e^{-r} \|A\|^{2r} \quad (1 \leq r \leq v).$$

If equality holds for an r , $1 < r \leq e$, then there exists a unitary U such that

$$A\bar{A}^T = U^{-1} \left[\frac{\|A\|^2}{e} I + 0 \right] U.$$

We consider next useful variations of Theorem 3.2. Let \tilde{k} be defined by
(3.17) $\text{tr}(H) = \tilde{k}v$.

Let

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_v \end{pmatrix}$$

be a nonzero vector of v components and set

$$(3.18) \quad \bar{X}^T H X = \eta,$$

where

$$(3.19) \quad \eta = (\tilde{k} + (v - 1)\tilde{\lambda}) \sum_{i=1}^v x_i \bar{x}_i.$$

Define the matrix \tilde{B} of order v by

$$(3.20) \quad \tilde{B} = (\tilde{k} - \tilde{\lambda})I + \tilde{\lambda}S.$$

THEOREM 3.4. *The matrices H and \tilde{B} satisfy*

$$\text{tr}(C_r(H)) \leq \text{tr}(C_r(\tilde{B})) \quad (1 \leq r \leq v).$$

Equality holds for $r = 1$. If $\tilde{k} + (v - 1)\tilde{\lambda} \neq 0$ and equality holds for an r ,

$$1 < r \leq v,$$

or if $\tilde{k} + (v - 1)\tilde{\lambda} = 0$ and equality holds for an r ,

$$1 < r < v,$$

then

$$H = (\tilde{k} - \tilde{\lambda})I + \frac{\tilde{\lambda}v}{\sum x_i \bar{x}_i} [x_r \bar{x}_s].$$

For by the well-known minimax property of hermitian matrices (see, for example, [2] or [5]),

$$(3.21) \quad \lambda_v(\sum x_i \bar{x}_i) \leq \eta \leq \lambda_1(\sum x_i \bar{x}_i),$$

whence it follows that

$$(3.22) \quad \lambda_v \leq \tilde{k} + (v - 1)\tilde{\lambda} \leq \lambda_1.$$

Thus by Theorem 3.2,

$$\text{tr}(C_r(H)) \leq \text{tr}(C_r(\tilde{B})) \quad (1 \leq r \leq v).$$

Suppose now that $\tilde{k} + (v - 1)\tilde{\lambda} \neq 0$ and equality holds for an r , $1 < r \leq v$, or that $\tilde{k} + (v - 1)\tilde{\lambda} = 0$ and equality holds for an r , $1 < r < v$. Then there exists a unitary U such that

$$(3.23) \quad H = \bar{U}^T \tilde{B} U = (\tilde{k} - \tilde{\lambda})I + \tilde{\lambda} \bar{U}^T S U.$$

Let u_i denote the sum of column i of U . Then

$$\begin{aligned} \bar{X}^T H X &= (\tilde{k} - \tilde{\lambda}) \sum x_i \bar{x}_i + \tilde{\lambda} (\sum x_i u_i) (\sum \bar{x}_i \bar{u}_i) \\ &= (\tilde{k} - \tilde{\lambda} + \tilde{\lambda} v) \sum x_i \bar{x}_i, \end{aligned}$$

whence

$$\tilde{\lambda} (\sum x_i u_i) (\sum \bar{x}_i \bar{u}_i) = \tilde{\lambda} v \sum x_i \bar{x}_i.$$

If $\tilde{\lambda} = 0$, then $H = \tilde{k}I$, and if $\tilde{\lambda} \neq 0$, then

$$(3.24) \quad (\sum x_i u_i) (\sum \bar{x}_i \bar{u}_i) = v \sum x_i \bar{x}_i.$$

Now $U \bar{U}^T = I$ implies

$$\sum u_i \bar{u}_i = v,$$

and Cauchy's inequality implies

$$\begin{aligned} v \sum x_i \bar{x}_i &= (\sum x_i u_i) (\sum \bar{x}_i \bar{u}_i) \\ &\leq \sum x_i \bar{x}_i \sum u_i \bar{u}_i = v \sum x_i \bar{x}_i. \end{aligned}$$

Since equality holds, we must have

$$u_i = c \bar{x}_i \quad (i = 1, \dots, v),$$

where

$$c\bar{c} = \frac{v}{\sum x_i \bar{x}_i}.$$

Hence

$$H = (\tilde{k} - \tilde{\lambda})I + \tilde{\lambda} [\bar{u}_r u_s] = (\tilde{k} - \tilde{\lambda})I + \frac{\tilde{\lambda} v}{\sum x_i \bar{x}_i} [x_r \bar{x}_s].$$

Now let $\tilde{k} = k^*$ and set X equal to the column vector of v 1's. Let λ^* denote the value of $\tilde{\lambda}$ for this choice of X . Define the matrix B^* of order v by

$$(3.25) \quad B^* = (k^* - \lambda^*)I + \lambda^* S.$$

Then we obtain [21]

THEOREM 3.5. *The matrices H and B^* satisfy*

$$\text{tr}(C_r(H)) \leq \text{tr}(C_r(B^*)) \quad (1 \leq r \leq v).$$

Equality holds for $r = 1$. If $k^ + (v - 1)\lambda^* \neq 0$ and equality holds for an r ,*

$$1 < r \leq v,$$

or if $k^ + (v - 1)\lambda^* = 0$ and equality holds for an r ,*

$$1 < r < v,$$

then

$$H = B^*.$$

We discuss briefly analogues of the preceding theorems for induced matrices. We begin with the analogue of Theorem 3.1.

THEOREM 3.6. *The matrices H and B_0 satisfy*

$$\text{tr}(P_r(H)) \geq \text{tr}(P_r(B_0)) \quad (r = 1, 2, \dots).$$

Equality holds for $r = 1$. If equality holds for an $r > 1$, then there exists a unitary U such that

$$H = U^{-1}B_0U.$$

In Lemma 3.3, let $\epsilon = \lambda_1 - (k + (h - 1)\lambda)$, $\alpha = \lambda_1 - \epsilon$, and $\beta = \lambda_h$. Then

$$(3.26) \quad \frac{1}{(1 - \lambda_1x)(1 - \lambda_hx)} > \frac{1}{(1 - (k + (h - 1)\lambda)x)(1 - (\lambda_h + \epsilon)x)}.$$

By (3.26) and Lemma 3.4,

$$(3.27) \quad \begin{aligned} & \frac{1}{(1 - \lambda_1x)(1 - \lambda_hx) \prod_{i=2}^{h-1} (1 - \lambda_i x)} \\ & > \frac{1}{(1 - (k + (h - 1)\lambda)x)(1 - (\lambda_h + \epsilon)x) \prod_{i=2}^{h-1} (1 - \lambda_i x)} \\ & > \frac{1}{(1 - (k + (h - 1)\lambda)x)(1 - (k - \lambda)x)^{h-1}}, \end{aligned}$$

and the first conclusion follows.

Suppose that equality holds throughout (3.27) for some coefficient of x^r , where $r \neq 0, 1$. Then equality holds for some coefficient of x^r in

$$\frac{1}{(1 - (\lambda_h + \epsilon)x) \prod_{i=2}^{h-1} (1 - \lambda_i x)} > \frac{1}{(1 - (k - \lambda)x)^{h-1}},$$

where $r \neq 0, 1$. Thus

$$\lambda_2 = \dots = \lambda_{h-1} = \lambda_h + \epsilon = k - \lambda.$$

Also equality holds for some coefficient of x^r in

$$\frac{1}{(1 - \lambda_1x)(1 - \lambda_hx)} > \frac{1}{(1 - (k + (h - 1)\lambda)x)(1 - (\lambda_h + \epsilon)x)},$$

where $r \neq 0, 1$. Thus we must have $\epsilon = 0$, $\lambda_1 = k + (h - 1)\lambda$, and $\lambda_h = k - \lambda$, or $\alpha = \beta$, $\lambda_1 = k - \lambda$, and $\lambda_h = k + (h - 1)\lambda$. Hence the characteristic roots of H are $k + (h - 1)\lambda$ taken once, $k - \lambda$ taken $h - 1$ times, and 0 taken $v - h$ times. Thus there exists a unitary U such that

$$H = U^{-1}B_0U.$$

It is clear that analogues of the other theorems in this section follow without difficulty.

THEOREM 3.7. *The nonnegative hermitian matrix H satisfies*

$$\text{tr}(P_r(H)) \geq \text{tr}(P_r(B')), \text{tr}(P_r(\tilde{B})), \text{tr}(P_r(B^*)) \\ (r = 1, 2, \dots).$$

Equality holds in each of these inequalities for $r = 1$. If equality holds in one of these inequalities for an $r > 1$, then the matrix H corresponding to the case of equality equals

$$U^{-1}B'U, (\tilde{k} - \tilde{\lambda})I + \sum \frac{\tilde{\lambda}_v}{x_i\bar{x}_i} [x_r\bar{x}_s], B^*,$$

respectively.

Note further that if A is a matrix of order v and rank $e > 0$, then

$$\text{tr}(P_r(A\bar{A}^T)) \geq \binom{e+r-1}{r} e^{-r} \|A\|^{2r} \quad (r = 1, 2, \dots).$$

If equality holds for an $r > 1$, then there exists a unitary U such that

$$A\bar{A}^T = U^{-1} \left[\frac{\|A\|^2}{e} I + 0 \right] U.$$

4. Applications. In this section we apply Theorem 3.5 to a variety of algebraic situations. We make no attempt to study these applications exhaustively. The same letters v , k , and λ arise in each of the problems studied, but the varied meanings of these symbols should not cause confusion.

In the applications we usually encounter a certain matrix that contains the essential information of the combinatorial or algebraic problem under investigation. Let us for the moment designate such a matrix by the letter E . The structure of the matrix E may vary considerably from problem to problem. E need not be hermitian or even square. However, the matrix

$$E\bar{E}^T = F$$

is nonnegative hermitian. Suppose that we perform arbitrary permutations to the rows and to the columns of E . This is equivalent to multiplying E on the left by a permutation matrix P_1 and on the right by a permutation matrix P_2 . Now if $E^* = P_1EP_2$ and $F^* = E^*\bar{E}^{*T}$, then $F^* = P_1FP_1^{-1}$

and $\text{tr}(C_r(F^*)) = \text{tr}(C_r(F))$. Similar remarks hold for the induced matrices $P_r(F)$. Thus the trace functions $\text{tr}(C_r(E\bar{E}^T))$ and $\text{tr}(P_r(E\bar{E}^T))$ are invariant under arbitrary permutations of the rows and of the columns of E . Such functions are frequently of combinatorial interest because they describe properties that are independent of the particular labeling of objects. The inequalities of §3 may be used to deduce upper bounds for $\text{tr}(C_r(E\bar{E}^T))$ and lower bounds for $\text{tr}(P_r(E\bar{E}^T))$. In the applications it is usually very revealing to investigate carefully those cases for which equality is attained.

Let

$$(4.1) \quad f(x) = x^v + a_1x^{v-1} + \cdots + a_v$$

be a polynomial of degree v with integer coefficients and with leading coefficient equal to one. The first application involves the algebraic integers $\alpha_1, \dots, \alpha_v$ that are the zeros of this polynomial. We form the Vandermonde matrix

$$(4.2) \quad V = \begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_1^{v-1} \\ \vdots & & & \\ 1 & \alpha_v & \cdots & \alpha_v^{v-1} \end{bmatrix}$$

of order v and define

$$(4.3) \quad \bar{V}^T V = H.$$

Now let

$$(4.4) \quad \text{tr}(H) = kv,$$

$$(4.5) \quad SHS = \mu S,$$

$$(4.6) \quad \mu = (k + (v - 1)\lambda)v,$$

where S is the v by v matrix of 1's. Evidently

$$(4.7) \quad k = \left(v + \sum_{i=1}^v \alpha_i \bar{\alpha}_i + \cdots + \sum_{i=1}^v \alpha_i^{v-1} \bar{\alpha}_i^{v-1} \right) / v,$$

$$(4.8) \quad \mu = \left(1 + \sum_{i=1}^{v-1} \alpha_1^i \right) \left(1 + \sum_{i=1}^{v-1} \bar{\alpha}_1^i \right) + \cdots + \left(1 + \sum_{i=1}^{v-1} \alpha_v^i \right) \left(1 + \sum_{i=1}^{v-1} \bar{\alpha}_v^i \right),$$

$$(4.9) \quad \det(H) = \prod_{i,j=1; i < j}^v (\alpha_i - \alpha_j)(\bar{\alpha}_i - \bar{\alpha}_j).$$

THEOREM 4.1. Let $f(x) = x^v + a_1x^{v-1} + \cdots + a_v$ be a polynomial of degree $v \geq 3$ with integer coefficients and with leading coefficient equal to one. Let the zeros $\alpha_1, \dots, \alpha_v$ of this polynomial be distinct and define k and λ by (4.4) and (4.6), respectively. Then

$$(4.10) \quad \prod_{i,j=1; i < j}^v (\alpha_i - \alpha_j)(\bar{\alpha}_i - \bar{\alpha}_j) \leq (k + (v - 1)\lambda)(k - \lambda)^{v-1}.$$

Equality holds if and only if

$$f(x) = x^v - 1, \quad x^v + 1, \quad \text{or} \quad x^v + x^{v-1} + \cdots + x + 1.$$

The inequality (4.10) is an immediate consequence of Theorem 3.5 with $r = v$. Suppose that equality holds in (4.10). We are assuming distinct zeros so that $k + (v - 1)\lambda \neq 0$, and hence by Theorem 3.5,

$$(4.11) \quad H = (k - \lambda)I + \lambda S.$$

By (4.11),

$$k = \alpha_1\bar{\alpha}_1 + \cdots + \alpha_v\bar{\alpha}_v = \alpha_1^2\bar{\alpha}_1^2 + \cdots + \alpha_v^2\bar{\alpha}_v^2 = v.$$

But

$$v^2 = (\alpha_1\bar{\alpha}_1 + \cdots + \alpha_v\bar{\alpha}_v)^2 \leq (\alpha_1^2\bar{\alpha}_1^2 + \cdots + \alpha_v^2\bar{\alpha}_v^2)v = v^2,$$

and hence

$$(4.12) \quad \alpha_1\bar{\alpha}_1 = \cdots = \alpha_v\bar{\alpha}_v = 1.$$

Now

$$\lambda = \alpha_1 + \cdots + \alpha_v < v,$$

and since $\mu > 0$,

$$\lambda > \frac{-v}{v-1}.$$

Thus λ is an integer such that

$$(4.13) \quad -1 \leq \lambda \leq v - 1.$$

It follows from (4.11) that

$$(4.14) \quad \alpha_1^i + \cdots + \alpha_v^i = \lambda \quad (i = 1, \dots, v - 1).$$

The Newton identities imply

$$(4.15) \quad \lambda + a_1\lambda + a_2\lambda + \cdots + a_i\lambda + (i+1)a_{i+1} = 0 \quad (i = 1, \dots, v-2).$$

Hence

$$a_i\lambda - ia_i + (i+1)a_{i+1} = 0 \quad (i = 1, \dots, v-2),$$

and

$$(4.16) \quad a_i = \frac{-\lambda(1-\lambda)\cdots((i-1)-\lambda)}{i!} \quad (i = 1, \dots, v-1).$$

Consider the case $\lambda = -1$. By (4.12) and (4.16),

$$(4.17) \quad f(x) = x^v + x^{v-1} + \cdots + x + 1$$

or

$$(4.18) \quad f(x) = x^v + x^{v-1} + \cdots + x - 1.$$

The polynomial (4.18) does not have a real zero equal to ± 1 . But then by (4.12), this polynomial has no real zero. This is impossible for both odd and even v , and thus (4.18) is excluded.

There remain the integral λ that satisfy $0 \leq \lambda \leq v - 1$. If $\lambda = 0$, then

$$(4.19) \quad f(x) = x^v \pm 1.$$

Suppose that $1 \leq \lambda \leq v - 1$. Then (4.16) implies

$$(4.20) \quad f(x) = x^{v-\lambda}(x - 1)^\lambda \pm 1.$$

By (4.12) these polynomials have no real zeros. But then v is even and

$$(4.21) \quad f(x) = x^{v-\lambda}(x - 1)^\lambda + 1.$$

Let α be a zero of (4.21). Then

$$\alpha^{v-\lambda}(\alpha - 1)^\lambda = \bar{\alpha}^{v-\lambda}(\bar{\alpha} - 1)^\lambda = -1,$$

whence

$$[(\alpha - 1)(\bar{\alpha} - 1)]^\lambda = 1$$

and

$$(\alpha - 1)(\bar{\alpha} - 1) = 1.$$

This implies $\alpha + \bar{\alpha} = 1$ and $\alpha^2 + \bar{\alpha}^2 = -1$. Thus

$$(4.22) \quad \lambda = \alpha_1^2 + \cdots + \alpha_v^2 = -v/2,$$

and this excludes (4.21).

One may study related problems that concern algebraic number fields. For instance, let K be an algebraic number field over the field R of rational numbers. Let K be of degree v and normal over R . Let $\theta_1, \dots, \theta_v$ denote the v automorphisms of K relative to R , and let $\alpha_1, \dots, \alpha_v$ be arbitrary in K . Then

$$D = [\theta_i(\alpha_j)]$$

defines a matrix of order v and

$$\bar{D}^T D = J$$

is nonnegative hermitian. An application of Theorem 3.5 to the matrix J yields a determinantal inequality as in Theorem 4.1. There remains the problem of determining the sets $\alpha_1, \dots, \alpha_v$ for which equality holds. We shall not pursue this topic here. Some material pertinent to this subject may be found in [7].

Our next application deals with oriented graphs. A *graph* G consists of a non-null set V of objects called *points* and a set W of objects called *lines*, the two sets having no elements in common. With each line there are associated just two distinct points, called its *endpoints*. The line is said to *join* its endpoints. Isolated points are permitted and two or more lines may join the same endpoints. If the number of lines joining distinct point pairs is the same for every such pair, then G is *complete*. G is *finite* if both V and W are finite. G is *oriented* if each line is assigned a direction in one of the two possible ways. We consider finite oriented graphs. We exclude from consideration the trivial graphs for which the set W is empty.

Let P_1, \dots, P_v be the points and let L_1, \dots, L_w be the lines of the oriented graph G . Let $p_{ij} = 1$ if P_i is the initial point of the directed line L_j , let

$p_{ij} = -1$ if P_i is the terminal point of the directed line L_j , and let $p_{ij} = 0$ if P_i is not an endpoint of L_j . Then

$$(4.23) \quad P = [p_{ij}]$$

defines a matrix of size v by w . This matrix is called the *incidence matrix* of G and contains all of the essential information given by G . Each column of P has a single entry equal to 1, a single entry equal to -1 , and all of the remaining entries of the column equal to 0. Conversely, every matrix of this type defines a finite oriented graph.

The matrix

$$(4.24) \quad PP^T = H$$

is nonnegative symmetric. As before let

$$\begin{aligned} \text{tr}(H) &= kv, \\ SHS &= \mu S, \\ \mu &= (k + (v - 1)\lambda)v, \end{aligned}$$

where S is the v by v matrix of 1's. Evidently,

$$(4.25) \quad k = \frac{2w}{v}$$

and

$$(4.26) \quad \lambda = \frac{-2w}{v(v - 1)}.$$

Moreover, if

$$(4.27) \quad B = (k - \lambda)I + \lambda S$$

is of order v , then

$$(4.28) \quad \text{tr}(C_r(B)) = \binom{v}{r} \left(\frac{v - r}{v} \right) \left(\frac{2w}{v - 1} \right)^r \quad (r = 1, \dots, v).$$

Note that G is complete if and only if $PP^T = B$.

Incidence matrices of oriented graphs have been treated by Poincaré, Veblen, and a number of other writers [4; 12; 18; 24]. The rank and the subdeterminants of P are especially significant from the graph-theoretic viewpoint. Every subdeterminant of P of order r is equal to $+1$, -1 , or 0 . A detailed discussion of the graphical interpretation of this result may be found in [4].

Now let \tilde{P} be the square matrix obtained by bordering P with a suitable number of zero rows or zero columns. Then by (2.1) and (2.2),

$$C_r(\tilde{P}\tilde{P}^T) = C_r(\tilde{P})(C_r(\tilde{P}))^T.$$

But

$$\text{tr}(C_r(PP^T)) = \text{tr}(C_r(\tilde{P}\tilde{P}^T)) = \text{tr}(C_r(\tilde{P})(C_r(\tilde{P}))^T),$$

and the last of these expressions equals the sum of the squares of the minors

of order r of P . Hence $\text{tr}(C_r(PPT))$ equals the number of nonzero minors of order r of the incidence matrix P .

THEOREM 4.2. *The incidence matrix P of an oriented graph G of v points and w lines satisfies*

$$(4.29) \quad \text{tr}(C_r(PPT)) \leq \binom{v}{r} \left(\frac{v-r}{v}\right) \left(\frac{2w}{v-1}\right)^r \quad (r = 1, \dots, v).$$

Equality holds for $r = 1$ and $r = v$. Equality holds for an r , $1 < r < v$, if and only if G is complete.

Inequality (4.29) follows from Theorem 3.5 and (4.28). We now have $k + (v-1)\lambda = 0$, and if equality holds for an r , $1 < r < v$, then Theorem 3.5 implies $PPT = B$, whence G is complete.

Our concluding application concerns the combinatorial designs known as v, k, λ configurations. This topic has been studied previously in [21], and we confine ourselves to a summary of the main results. Let $Q = [q_{ij}]$ be a matrix of order v , all of whose entries are 0's and 1's. Let $v > 1$, and let τ denote the total number of 1's in Q . The matrix Q may be viewed as an incidence matrix for an arrangement of v elements x_1, \dots, x_v into v sets S_1, \dots, S_v , where $q_{ij} = 1$ if x_j is in S_i , and $q_{ij} = 0$ if x_j is not in S_i . With Q we associate the nonnegative symmetric matrix

$$(4.30) \quad QQ^T = W,$$

and we investigate the arrangement of the τ 1's in Q in order that $\text{tr}(C_r(W))$ attain the maximum value.

Let

$$(4.31) \quad \text{tr}(W) = kv = \tau,$$

$$(4.32) \quad SWS = \mu S,$$

$$(4.33) \quad \mu = (k + (v-1)(\lambda(Q)))v,$$

where S is the v by v matrix of 1's and $\lambda(Q)$ is a rational number determined by the particular 0, 1 arrangement within Q . Now define

$$(4.34) \quad \lambda = \frac{k(k-1)}{v-1}.$$

The 0, 1 matrix Q of order v containing $\tau = kv$ 1's satisfies

$$(4.35) \quad \lambda \leq \lambda(Q) \leq k.$$

The inequalities (4.35) are easy to establish [21].

We describe now some 0, 1 matrices A of order v called incidence matrices of v, k, λ configurations. Let v elements x_1, \dots, x_v be arranged into v sets S_1, \dots, S_v such that every set contains exactly k distinct elements and such that every pair of sets has exactly λ elements in common, $0 < \lambda < k < v$.

Such an arrangement is a v, k, λ configuration. Every v, k, λ configuration satisfies (4.34). For such a configuration, let $a_{ij} = 1$ if x_j is an element of S_i , and let $a_{ij} = 0$ if x_j is not an element of S_i . The v by v matrix

$$(4.36) \quad A = [a_{ij}]$$

of 0's and 1's is the *incidence matrix of the v, k, λ configuration*. Define the matrix B of order v by

$$(4.37) \quad B = (k - \lambda)I + \lambda S.$$

One verifies readily that if $0 < \lambda < k < v$, then a v, k, λ configuration exists if and only if there exists a 0, 1 matrix A of order v such that

$$(4.38) \quad AAT = B$$

These v, k, λ configurations and their related incidence matrices have an extensive literature. The main problem deals with the determination of the precise range of values of v, k, λ for which configurations exist. Certain nonexistence theorems are derived in [1] and [3], but the general problem is still unsolved. A survey of the literature concerning these configurations is available in [8; 17; 19].

THEOREM 4.3. *Let Q be a 0, 1 matrix of order v , containing exactly $\tau = kv$ 1's. Let $\lambda = k(k - 1)/(v - 1)$ and $B = (k - \lambda)I + \lambda S$, where $0 < \lambda < k < v$. Then*

$$\text{tr}(C_r(QQ^T)) \leq \text{tr}(C_r(B)) \quad (r = 1, \dots, v).$$

Equality holds for $r = 1$. Equality holds for an $r > 1$ if and only if Q is the incidence matrix of a v, k, λ configuration.

This theorem follows from Theorem 3.5 and (4.35) [21]. Note that Theorem 4.3 implies

$$(4.39) \quad (\det(Q))^2 \leq k^2(k - \lambda)^{v-1},$$

where equality holds if and only if Q is the incidence matrix of a v, k, λ configuration [20].

Suppose that the parameters v, k, λ are specified in such a way that a v, k, λ configuration does not exist. Let Q be the 0, 1 matrix of order v containing $\tau = kv$ 1's. Then there remains the problem of devising an arrangement of the τ 1's in Q in such a way that $\text{tr}(C_r(QQ^T))$ attains the maximum value. This problem is not solved by Theorem 4.3, and a general solution appears to be very difficult.

In the preceding applications we have not mentioned the analogous theorems for induced matrices, but it is clear that these results follow without difficulty. The traces of compound and induced matrices are by no means the only functions of importance in combinatorial extremal problems of the

type described here. Another is the *permanent* of a matrix $A = [a_{ij}]$ of order n . This is defined by

$$(4.40) \quad \text{per}(A) = \sum a_{1i_1}a_{2i_2}\cdots a_{ni_n},$$

where the summation extends over the $n!$ permutations of the integers i_1, i_2, \dots, i_n . Thus $\text{per}(A)$ is like $\det(A)$ without sign changes. We mention in passing that $\text{per}(A)$ is one of the terms appearing on the main diagonal of the induced matrix $P_n(A)$, and thus $\text{per}(A)$ contributes a term to $\text{tr}(P_n(A))$. Furthermore, $\text{per}(A)$ is *invariant under arbitrary permutations of the rows and of the columns of A* .

Extremal problems involving permanents are usually very difficult. Suppose that X is a doubly stochastic matrix of order n . This means that the entries of X are nonnegative reals and that the row sums and the column sums of X are each equal to 1. In 1926 van der Waerden suggested the problem of determining the minimum of $\text{per}(X)$, where X is doubly stochastic [12, p. 238; 25]. This problem is still unsolved. Recently Marcus and Newman [16] have made advances toward a solution. The current conjecture is that $\text{per}(X) \geq n!/n^n$, with equality if and only if X is the doubly stochastic matrix all of whose entries are $1/n$.

A number of interesting extremal problems involve the permanents of 0, 1 matrices. Let Q be a 0, 1 matrix of order v . Then Q may be viewed as an incidence matrix for an arrangement of v elements into v sets, and $\text{per}(Q)$ equals the total number of systems of distinct representatives for the arrangement of the v elements into the v sets. See [9] for the definition and basic properties of systems of distinct representatives. Let \mathfrak{A} denote the class of all 0, 1 matrices of order v , with exactly k 1's in each row and in each column. Let E be in \mathfrak{A} . It is well-known that $\text{per}(E) > 0$ [12]. But very little is known about the minimal value of $\text{per}(E)$ for E in \mathfrak{A} . If \mathfrak{A} contains an incidence matrix A of a v, k, λ configuration, then very limited empirical information suggests that $\text{per}(A)$ is small or even minimal in \mathfrak{A} . This is in some ways unexpected, because (4.39) asserts that incidence matrices of v, k, λ configurations have determinants that are in absolute value maximal in \mathfrak{A} .

In conclusion we mention a remarkable property of the incidence matrix A of a configuration with parameters $v = 7, k = 3, \lambda = 1$. Such a matrix satisfies

$$(4.41) \quad \text{per}(A) = \text{abs. val. } (\det(A)) = 24.$$

These matrices of order 7 play a unique role in the forthcoming paper by Tinsley on permanents of cyclic 0, 1 matrices [23].

REFERENCES

1. R. H. Bruck and H. J. Ryser, *The nonexistence of certain finite projective planes*, Canad. J. Math. vol. 1 (1949) pp. 88–93.

2. N. G. de Bruijn, *Inequalities concerning minors and eigenvalues*, Nieuw Arch. Wisk. (3) vol. 4 (1956) pp. 18–35.
3. S. Chowla and H. J. Ryser, *Combinatorial problems*, Canad. J. Math. vol. 2 (1950) pp. 93–99.
4. Jules Chuard, *Questions d'analysis situs*, Rend. Circ. Mat. Palermo, vol. 46 (1922) pp. 185–224.
5. R. Courant and D. Hilbert, *Methoden der mathematischen Physik*, vol. 1, Berlin, 1931.
6. C. J. Everett and H. J. Ryser, *Rational vector spaces I*, Duke Math. J. vol. 16 (1949) pp. 553–570.
7. ———, *Rational vector spaces II*, Duke Math. J. vol. 17 (1950) pp. 135–145.
8. Marshall Hall, Jr., *Projective planes and related topics*, California Institute of Technology, 1954.
9. P. Hall, *On representatives of subsets*, J. London Math. Soc. vol. 10 (1935) pp. 26–30.
10. G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge, 1952.
11. A. Hurwitz, *Zur Invariantentheorie*, Math. Ann. vol. 45 (1894) pp. 381–404.
12. Dénes König, *Theorie der endlichen und unendlichen Graphen*, New York, Chelsea, 1950.
13. Dudley E. Littlewood, *The theory of group characters and matrix representations of groups*, Oxford, 1950.
14. C. C. MacDuffee, *The theory of matrices*, Berlin, 1933.
15. Marvin Marcus, *On subdeterminants of doubly stochastic matrices*, Illinois J. Math. vol. 1 (1957) pp. 583–590.
16. Marvin Marcus and Morris Newman, *On the minimum of the permanent of a doubly stochastic matrix*, Duke Math. J. vol. 26 (1959) pp. 61–72.
17. Günter Pickert, *Projektive Ebenen*, Berlin, 1955.
18. H. Poincaré, *Second complément à l'analysis situs*, Proc. London Math. Soc. vol. 32 (1901) pp. 277–308.
19. H. J. Ryser, *Geometries and incidence matrices*, Slaught Papers no. 4, Mathematical Association of America, 1955.
20. ———, *Maximal determinants in combinatorial investigations*, Canad. J. Math. vol. 8 (1956) pp. 245–249.
21. ———, *Inequalities of compound and induced matrices with applications to combinatorial analysis*, Illinois J. Math. vol. 2 (1958) pp. 240–253.
22. Issai Schur, *Ueber eine Klasse von Matrizen die sich einer gegebenen Matrix zuordnen lassen*, Dissertation, Berlin, 1901.
23. Marion F. Tinsley, *Permanents of cyclic matrices*, to appear in Pacific J. Math.
24. Oswald Veblen, *Analysis situs*, Amer. Math. Soc. Colloquium Publications, vol. 5, 1931.
25. B. L. van der Waerden, *Aufgabe 45*, Jahresbericht der deutschen Mathematiker-Vereinigung, vol. 35 (1926) p. 117.
26. J. H. M. Wedderburn, *Lectures on matrices*, Amer. Math. Soc. Colloquium Publications, vol. 17, 1934.

THE OHIO STATE UNIVERSITY,
COLUMBUS, OHIO

This page intentionally left blank

PERMANENTS OF DOUBLY STOCHASTIC MATRICES

BY

MARVIN MARCUS¹ AND MORRIS NEWMAN

I. Introduction. This note contains a brief account of the extreme values of the permanent of a doubly stochastic matrix. Full details and further results are given in [4]. The problem of determining the minimum seems first to have been proposed by B. van der Waerden [5] and later by D. König [2], and was brought to our attention by H. Ryser. As far as we are aware, no work on this problem has previously appeared.

A doubly stochastic matrix is one whose entries are non-negative and whose row and column sums are all 1. We let K_n be the (convex) set of all $n \times n$ doubly stochastic matrices, K_n^0 the relative interior of K_n in the Euclidean topology. The permanent of X is the function

$$\text{per}(X) = \sum x_{1i_1}x_{2i_2}\cdots x_{ni_n}$$

summed over all permutations (i_1, i_2, \dots, i_n) of the integers $(1, 2, \dots, n)$.

A first elementary result is contained in

LEMMA 1. *For S in K_n , $\text{per}(S) \leq 1$ with equality if and only if S is a permutation matrix. Also,*

$$(1) \quad \text{per}(S) \geq (n^2 - n + 1)^{1-n}.$$

This lemma is easily implied by G. Birkhoff's result [1] that a doubly stochastic matrix is in the convex hull of at most $n^2 - n + 1$ permutation matrices, and that

$$\text{per}(A + B) \geq \text{per}(A) + \text{per}(B)$$

for A and B any pair of matrices with non-negative elements.

We introduce some notation. An $n \times n$ matrix C will sometimes be written in terms of its row vectors as (C_1, C_2, \dots, C_n) . We denote the vector all of whose entries are 1 by ϵ . J_n is the $n \times n$ matrix in K_n whose entries are all $1/n$. $C_{ij}(A)$ denotes the permanent of the $(n - 1) \times (n - 1)$ matrix obtained by deleting row i and column j of A and will be called a p -minor. An $n \times n$ matrix S is called *decomposable* if there is a permutation matrix P such that

$$P'SP = \begin{pmatrix} A & X \\ 0 & B \end{pmatrix}$$

where A and B are square matrices, and is called *completely decomposable* if there is a permutation matrix P such that $P'SP = A + B$ (the direct

¹ National Research Council—National Bureau of Standards Postdoctoral Research Fellow, 1956–1957.

sum of A and B). If S is not decomposable it is called *indecomposable*. A decomposable doubly stochastic matrix is completely decomposable.

The following facts are of importance:

$$(2) \quad \text{per}(PAQ) = \text{per}(A), \quad P, Q \text{ permutation matrices.}$$

$$(3) \quad AJ_n = J_n A = J_n, \quad \text{if } A \text{ is in } K_n.$$

$$(4) \quad \text{per} \begin{pmatrix} A & X \\ 0 & B \end{pmatrix} = \text{per}(A) \text{ per}(B), \text{ where } A, B \text{ are square matrices.}$$

For an $n \times n$ matrix A

$$(5) \quad \text{per}(A) = \sum_{j=1}^n a_{ij} C_{ij}(A), \quad 1 \leq i \leq n,$$

and so

$$(6) \quad \frac{\partial \text{per}(A)}{\partial a_{ij}} = C_{ij}(A).$$

(7) The permanent of a matrix is a multilinear function of its rows (columns).

We also need the following results of Perron, Frobenius and Wielandt [3] on matrices with non-negative elements:

I. If A is an indecomposable $n \times n$ matrix with non-negative elements then A has a simple eigenvalue $\lambda > 0$ such that if α is an eigenvalue of A then $|\alpha| \leq \lambda$.

II. If S is a symmetric doubly stochastic matrix then the null space corresponding to the maximal eigenvalue 1 is one-dimensional.

We obtain II from I immediately since $\lambda = 1$ is clearly the eigenvalue of S with maximum modulus and the null space of $S - I$ has the same dimension as the multiplicity of $\lambda = 1$.

LEMMA 2. If $\text{per}(A) = \min_{S \in K_n} \text{per}(S)$ then A is indecomposable.

Proof. Since a doubly stochastic decomposable matrix A is completely decomposable we need only show that A is not completely decomposable.

We actually prove the stronger statement that PAQ is never a direct sum for permutation matrices P and Q . Suppose

$$G = PAQ = L + M$$

where $L \in K_u$ and $M \in K_v$. We may assume that $l_{uu}C_{uu}(L) > 0$ and $m_{11}C_{11}(M) > 0$ since by (1) some term both in the expansion of L by its last row and M by its first row must be positive. Define $G(\theta)$ to be the matrix obtained from G by subtracting θ from l_{uu} and m_{11} and adding θ to the zeros in the $u, u+1$ and $u+1, u$ positions of G . If $0 < \theta$ is sufficiently small then $G(\theta) \in K_n$. We find by differentiation that

$$\text{per}(G(\theta)) = \text{per}(G) - k\theta + O(\theta^2)$$

where k is a sum of four terms as follows :

$$C_{\mathbf{u}} \per(M) + C_{11}(M) \per(L) - C_{\mathbf{u}+1}(G) - C_{\mathbf{u}+1 \cdot \mathbf{u}}(G).$$

Now from (4), the last two terms are 0 since $C_{\mathbf{u}+1}(G)$ (resp. $C_{\mathbf{u}+1 \cdot \mathbf{u}}(G)$) is the product of the permanents of two matrices one of which has a column (resp. row) of zeros. Thus for $0 < \theta$ sufficiently small

$$\per(G(\theta)) < \per(G),$$

a contradiction. Hence the lemma is proved.

II. The extreme values.

THEOREM 1. *If $\per(A) = \min_{S \in K_n} \per(S)$ then there exists a set R of pairs of integers (i,j) such that*

- (a) $a_{ij} = 0$ if and only if $(i,j) \in R$,
- (b) $(i,j) \notin R$ implies $C_{ij}(A) = \per(A)$,
- (c) $(i,j) \in R$ implies $C_{ij}(A) = \per(A) + \beta$ for $\beta \geq 0$ and independent of (i,j) .

The proof of this theorem, which is long, is given in [4]. The method of Lagrange multipliers is used to determine

$$\min_{S \in F} \per(S)$$

where F is the face of least dimension of K_n containing an absolute minimizing matrix in its interior. Lemma 2 and result II are of fundamental importance here.

We now introduce an averaging process, and set

$$M = M_n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} \\ & & \ddots & & \\ \frac{1}{2^{n-1}} & \frac{1}{2^{n-1}} & \frac{1}{2^{n-2}} & \cdots & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2^{n-1}} & \frac{1}{2^{n-1}} & \frac{1}{2^{n-2}} & \cdots & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Clearly M is doubly stochastic and indecomposable. By I and the Gershgorin circle theorem we see that all the eigenvalues of M are strictly less

than 1 in modulus with the exception of the simple root 1. Thus M is similar to a direct sum,

$$M \sim (1) + B_1 + \cdots + B_k, \quad B_t = \begin{bmatrix} \lambda_t & c_t \\ & \ddots \\ & & \ddots \\ & & & \lambda_t & c_t \end{bmatrix}, \quad |\lambda_t| < 1.$$

Thus $B_t^t \rightarrow 0$ as $t \rightarrow \infty$, and so

$$L = \lim_{t \rightarrow \infty} M^t \sim (1) + 0.$$

Thus L is a doubly stochastic matrix of rank 1, which implies that $L = J_n$. We have proved therefore the following lemma :

LEMMA 3.

$$\lim_{t \rightarrow \infty} M^t = J_n.$$

It is easily verified that

$$M = (I_{n-2} + J_2)(I_{n-3} + J_2 + I_1) \cdots (J_2 + I_{n-2}).$$

Thus if

$$E = J_2 + I_{n-2},$$

we have for suitable permutation matrices P_t ,

$$(8) \quad M = P_1 E P_2 E \cdots P_{n-1} E.$$

LEMMA 4. Suppose the p -minors of A are all equal. Then $\text{per}(EA) = \text{per}(A)$.

Proof. From (7) we have

$$\begin{aligned} \text{per}(EA) &= \frac{1}{2} \text{per}(A) + \frac{1}{4} \text{per}(A_1, A_1, A_3, \dots, A_n) + \\ &\quad \frac{1}{4} \text{per}(A_2, A_2, A_3, \dots, A_n). \end{aligned}$$

Since the p -minors of A are all equal,

$$\text{per}(A_1, A_1, A_3, \dots, A_n) = \text{per}(A_2, A_2, A_3, \dots, A_n) = \text{per}(A),$$

and so the lemma is proved.

THEOREM 2. Suppose that $\text{per}(A) = \min_{S \in K_n} \text{per}(S)$, and $A \in K_n^0$. Then

$$\text{per}(A) = \text{per}(J_n) = \frac{n!}{n^n}.$$

Proof. Since A is an interior absolute minimum, all the p -minors of A are equal by Theorem 1. Therefore by (7) $\text{per}(EA) = \text{per}(A)$. It is clear

that EA is an interior point of K_n whenever A is and so EA is also an interior absolute minimum. Theorem 1 now implies that the p -minors of EA are all equal and so we find that $\text{per}(MA) = \text{per}(A)$, from (2) and (8). Reasoning in the same manner with M we deduce that for all positive integral t ,

$$\text{per}(M^t A) = \text{per}(A).$$

By Lemma 3, we find

$$\begin{aligned} \text{per}(A) &= \lim_{t \rightarrow \infty} \text{per}(M^t A) = \text{per}(J_n A) \\ &= \text{per}(J_n) = \frac{n!}{n^n}. \end{aligned}$$

The proof of the theorem is thus complete.

We can show that an interior absolute minimum is necessarily unique and equal to J_n . For this purpose we require the following lemma:

LEMMA 5. *If $A \neq J_n$ and A is an interior point of K_n in a sufficiently small neighborhood of J_n then*

$$\text{per}(A) > \text{per}(J_n).$$

The details of this proof, which is easy, are to be found in [4].

THEOREM 3. *If the absolute minimum is attained in the interior of K_n then it is assumed uniquely for the matrix J_n .*

Proof. Let A be an interior absolute minimum and assume that $A \neq J_n$. Consider the sequence of matrices

$$(9) \quad A, MA, M^2A, \dots$$

Suppose that $M^r A = M^s A$, $r > s$; put $r = s + t$. Then

$$M^s A = M^t(M^s A) = M^{2t}(M^s A) = \dots,$$

so that $M^s A = J_n M^s A = J_n$, by Lemma 3 and (3). Put $M^{s-1} A = B$. Then $MB = J_n$, and so

$$\begin{aligned} \frac{1}{2} B_1 + \frac{1}{2} B_2 &= \frac{1}{4} B_1 + \frac{1}{4} B_2 + \frac{1}{2} B_3 = \dots \\ &= \frac{1}{2^{n-1}} B_1 + \frac{1}{2^{n-1}} B_2 + \dots + \frac{1}{2} B_n = \frac{1}{n} \epsilon, \end{aligned}$$

which implies that

$$\frac{1}{2} (B_1 + B_2) = B_3 = B_4 = \dots = B_n = \frac{1}{n}.$$

Since B is an interior absolute minimum, the p -minors of B are all equal. This implies after a small calculation that $B = J_n$.

Repeating this argument we deduce that $A = J_n$, a contradiction. The sequence (9) therefore is an infinite sequence of interior absolute minima containing no repetitions.

Since $M^t \rightarrow J_n$, this implies that in arbitrarily small neighborhoods of J_n there are matrices distinct from J_n with the same permanent as J_n , contradicting Lemma 5. Theorem 3 is thus proved.

REFERENCES

1. G. Birkhoff, *Three observations on linear algebra*, Univ. Nac. Tucumán Rev. ser. A vol. 5 (1946) pp. 147–151.
2. D. König, *Theorie der Graphen*, New York, Chelsea, 1950, p. 238.
3. H. Wielandt, *Unzerlegbare, nicht negative Matrizen*, Math. Z. vol. 52 (1950) pp. 642–648.
4. M. Marcus and M. Newman, *On the minimum of the permanent of a doubly stochastic matrix*, Duke Math. J. vol. 26 (1959) pp. 61–72.
5. B. van der Waerden, *Aufgabe 45*, Jber. Deutsch. Math. Verein. vol. 35 (1926) p. 117.

UNIVERSITY OF BRITISH COLUMBIA,
VANCOUVER, CANADA

NATIONAL BUREAU OF STANDARDS,
WASHINGTON, D.C.

A SEARCH PROBLEM IN THE *N*-CUBE

BY

ANDREW M. GLEASON¹

It is frequently important to find the maximum value of a function defined on a finite set. Any finite combinatorial problem can be recast into this form, but in general the new formulation will be of no value. It is entirely clear that there is no certain method of finding the maximum short of computing the function at each point of the set in question and no statistical method can offer a probability higher than the proportion of the space on which the computation is made, unless the function has special properties known in advance.

In many cases the argument set has, in effect, a topology; that is a notion of distance which makes some pairs of points closer than others. Correspondingly the function to be maximized has a sort of continuity which may be expressed by saying that variations in the function due to a small change in the argument are small compared to the total variation of the function. Under these circumstances a systematic search for a maximum can be made by the method of ascent. One starts anywhere in the argument space, computes the value there and at each of the points within some distance (of which there will presumably be only a few), moves to that point with the largest function value and repeats the process until no move is indicated. The end result is a local maximum of the function, that is a point such that nowhere within the prescribed searching distance is the function larger. In case of ties, one might move to a tied point in the hope of finding a new way up. If this is continued exhaustively one finds either a strict local maximum point or an entire plateau of tied points.

There is, however, no guarantee that one will find the global maximum point by this method. The situation can be compared roughly to the following experiment: If one is put down at random in the United States and tries the method of ascent, what is the probability that he will wind up atop Mt. Whitney? This will depend on the size of the regions scanned but will remain quite small unless they are very large. In combinatorial problems of interest the peak sought is usually a great deal higher than the other local maxima and presumably its base covers an area considerably larger than those of the lesser peaks. The vital question is whether this base has area large relative to the whole space, for this relative area is the probability

¹ This study was supported jointly by the Army, Navy, and Air Force under contract with the Massachusetts Institute of Technology. Particular thanks are due to Mr. Roland Silver of Lincoln Laboratory who ably programmed the experiments described herein.

of success of the method. By analogy with the geographical model, one might expect to find high subsidiary peaks near the main peak; hence if the method should lead to a function value which is very high we might be justified in thinking the true maximum is attained nearby. It seems quite hopeless to compute how these matters will work out in any non-trivial case, so an experimental approach with the aid of high speed equipment is indicated.

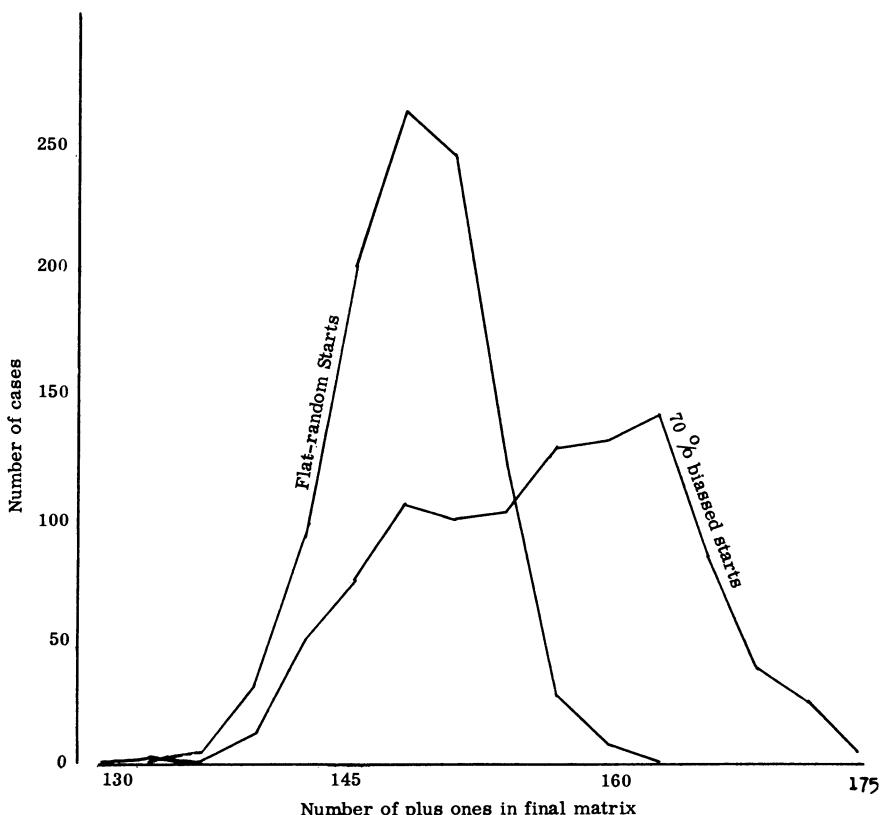
The problem was as follows: Given a matrix we are allowed to multiply any row or set of rows and any column or set of columns through by -1 and the object is to maximize the sum of the entries in the matrix. The matrices used were 15×15 and consisted of ± 1 's. Because changing signs in all rows produces the same effect as changing all columns there are 2^{29} different ways to modify each matrix, so the problem may be regarded as finding the best matrix among 2^{29} matrices. One may conveniently think of the 2^{30} corners of a 30-dimensional cube and a function which takes the same value at any pair of opposite corners. Each dimension of the cube corresponds to one of the possible row or column changes. Two corners are neighboring if they are ends of an edge, that is, if the matrices differ by changing the sign in a single row or column. Each matrix has 30 neighbors to be examined in the search procedure. Since the matrices were of odd size, changing the signs in a single row or column had to change the score so that no plateaus could exist.

Two variations of the method of ascent were tried, steepest ascent in which one always moved to the largest neighboring score and least ascent in which one moved to the neighboring score which gave the least true gain. When ties appeared concerning which move to make they were resolved by a rule of precedence. A priori it was argued that the method of least ascent, since it took longer to find a peak and wandered farther in the space might have a better chance of finding the highest peak. Actually it developed that the methods were about equally good at finding high peaks, but slowest ascent seems to find a slightly higher random peak!

Four experiments were run each involving one thousand matrices. The matrices were prepared in different ways in the various experiments. Each starting matrix was run with each of the ascent methods and the height of the resulting peaks compared and recorded.

In the first experiment the starting matrices were generated by a 50-50 pseudo-random stream of plus and minus ones. In the thousand trials, the slowest ascent method obtained a higher peak than the steepest ascent in 434 cases, a lower peak in 384 with 182 ties. An average random peak turned out to have 149 (or about 66%) plus ones. The distribution was quite tight having a standard deviation of 4.2. (A binomial distribution with this mean has a standard deviation of about 7.1.)

In the subsequent experiments the matrices were generated by a biased pseudo-random stream and then certain rows and columns were changed at



random. This produced a starting matrix which superficially resembles a flat random start but which is known to have a fairly high peak somewhere.

When the initial bias was 60% the final results are almost indistinguishable from the 50% starts, as might be expected since we are not introducing a peak larger than would be there anyway.

When the bias was 70%, however, there was a marked increase in the height of peaks found (see chart). This indicates that even when the main peak is just a little higher than random peaks there is a good probability of finding it by either method or at least of finding a large subsidiary peak. Unfortunately the experiment was not performed so as to be able to decide whether the various high peaks found were near one another. Both methods found peaks higher than the originally generated matrix in over 70% of the cases, but again the least ascent method outscored the steepest 300-277 with 423 ties.

With an initial bias of 80% both methods of search found a peak as high as the originally generated matrix in 974 cases; in short, one is virtually

certain to find the main range if not the actual summit. Analysis of the twenty-six failures indicates that the completely false peaks average lower than random peaks from the flat-start cases. When the initial bias was 90%, both methods always found a peak of the same height and this peak was at least as high as the original matrix.

Further experimentation is planned.

HARVARD UNIVERSITY,
CAMBRIDGE, MASSACHUSETTS

TEACHING COMBINATORIAL TRICKS TO A COMPUTER

BY

D. H. LEHMER

The widespread renewed interest in Combinatorial Analysis is largely due to the physical existence of the automatic digital computer which makes possible the actual carrying out of processes hitherto only talked about. Moreover, some of the characteristics of these computers have rendered feasible processes that are entirely beyond the ability of a human being to follow in detail. However, the modern "all purpose" computer is in reality not adept at combinatorial problems, being designed to do rational operations, in its own peculiar arithmetic, operations intended to mimic the corresponding multiplications, additions, divisions, and subtractions in the idealized real number system of the applied mathematician. The fact that a machine has been designed to do a 10×10 multiplication in N microseconds may not be very interesting to a combinatorial analyst with a problem involving no multiplication at all. He may be much more interested in the fact that the machine is capable of some rapid simple logical operation, such as recognition of overflow, which he can apply to his problem. To get the computer to do combinatorial problems efficiently requires a good deal of thoughtful teaching, some of which is done by the computer.

Many steps in the solution of combinatorial problems seem relatively small but nevertheless are not quite straightforward when programmed in the language of the computer's limited repertory of instructions. We give in what follows some suggestions for teaching the machine to handle some of the simpler types of combinatorial operations. It is hoped that some of the ideas set forth may be of use in dealing with the much more elaborate problems of this Symposium.

As a matter of fact, because of the flexibility of the computer control, a surprisingly large percentage of the instructions in an essentially non-combinatorial problem, for example matrix inversion, are actually of a combinatorial nature, although of a quite rudimentary sort such as tallying, comparing, and the other operations usually described as "address arithmetic."

To begin with the simplest problem of the next higher category we consider the problem of set inclusion. Suppose that a set S of n numbers

$$(1) \quad a_1, a_2, \dots, a_n$$

is stored in the machine in n addresses. The machine produces a number x and is asked to discover whether x belongs to the set S or not. To make things easy we may suppose that the n addresses are consecutive or even that the address of a_k is k or, what is the same, the word $w(k)$ at address k

is a_k . Without further information on the a 's there can be no other method than a methodical "house to house" search. A simple block diagram for this routine is given in Figure 1.

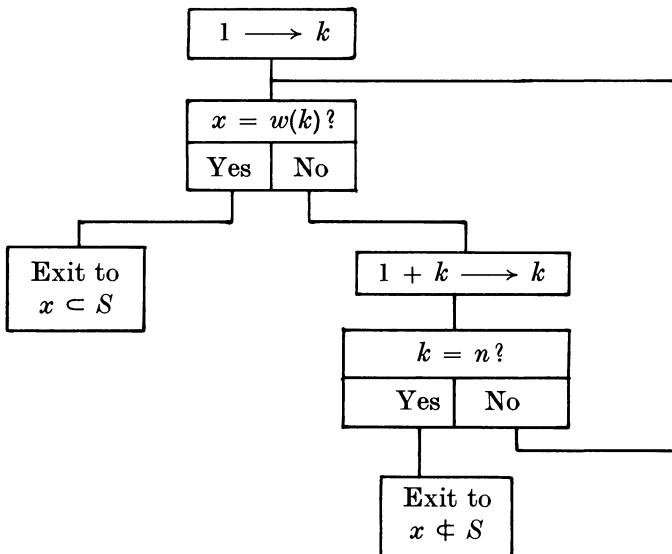


FIGURE 1

This routine may be expected to loop around $n/2$ times on the average. In case the a 's in (1) are monotonely increasing, another search method is available in which the effort is proportional to $\log n$ instead of n . It proceeds by successive dichotomies as indicated in Figure 2. In the third box $[(a + b)/2]$ denotes the greatest integer not exceeding the average of a and b and is found, of course, by shifting the sum $a + b$ one right.

The routine of Figure 2 is, to be sure, very much faster than the general routine of Figure 1. A still faster process may be used when the members a_k of S (or some function of them) are small positive integers. In this case the machine may ascertain whether x belongs to S by a method in which the time is practically independent of n . In this scheme the set S is represented by its "characteristic binary number" whose r th digit is 1 or 0 according as r is a member of S or not. (In some cases it may be more convenient to interchange the roles of 1 and 0.) For example, the odd prime numbers may be represented by the characteristic binary number

11101101101001100101101001...

in which the r th digit is 1 or 0 according as $2r + 1$ is prime or composite. Thousands of these digits may be stored in the high speed memory of the machine while millions of others may be stored on magnetic tapes if necessary. Whenever the machine produces a new number x it is instructed to

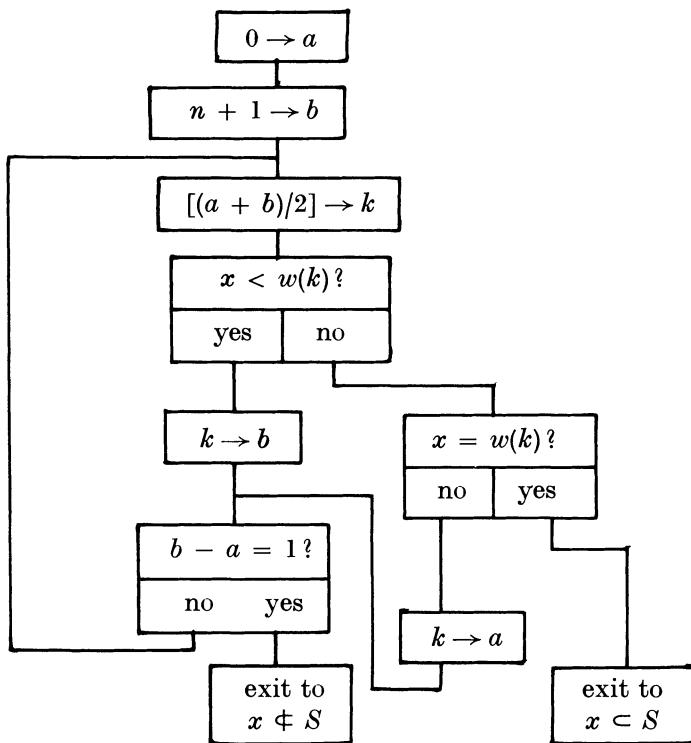


FIGURE 2

“extract” the x th binary digit of the characteristic member and ascertain whether or not it is zero. This will involve a minor amount of address arithmetic in order to choose the appropriate word containing the x th bit and the appropriate extractor to mask out all but the bit in question. This will generally call for a division of x by the number of bits in a word with scrupulous retention of the remainder. However if x runs over consecutive integer values, a simpler trick is available, namely the use of intentional overflow. If we denote the characteristic word by B and if x starts from 1, the simple diagram of Figure 3 illustrates how rapidly the question of set inclusion can be handled in this case.

Actually this routine is oversimplified. Eventually A , which acquires an extra terminal zero at each step, will consist wholly of zeros so that when x finally exceeds the number of bits in the word B it is time either to stop or to rejuvenate the word A and continue as before. One way to keep A alive indefinitely is to insert the command $A + 2^{-m} \rightarrow A$ just after the “yes.” This produces a strictly periodic pattern of yes’s and no’s of period m ; a sort of high speed roulette wheel which has a large number of combinatorial uses. For example such a subroutine can be placed after another one so as to steer the control of the program into one of two channels according to a

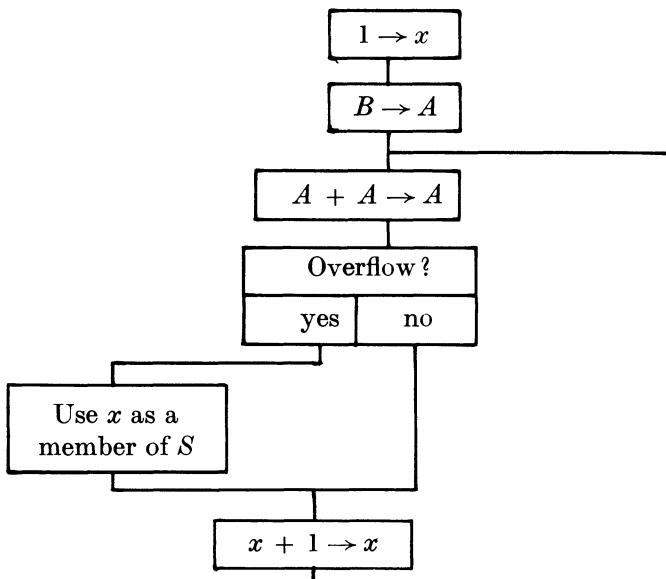


FIGURE 3

prearranged pattern, however complicated. A number of such subroutines can be compounded to produce pseudo-random digits. By using sets of such signals as extractors one produces a high speed parallel "sieve" that can be used to identify the answers to a wide class of diophantine problems [1]. The above brief remarks will serve to indicate what can come out of the simple question: "Does x belong to S ? ", a question almost never asked by the mathematician who "works" in set theory.

Changing the subject abruptly, there are simple combinatorial problems associated with multiple sums and other operators on functions of many variables. The problem of evaluating $f(x)$ for equally spaced values of x is easily generalized to the case of several variables having equal ranges. With a little address arithmetic one may reduce the problem to the consideration of the m^n lattice points in the n dimensional cube of side m in the "first orthant of E_n ", that is, the vectors

$$(k_1, k_2, \dots, k_n) \quad \begin{cases} k_i = 0(1)m - 1, \\ i = 1(1)n. \end{cases}$$

In dealing with this "unrestricted case" one has only to pretend that the k 's are the separate digits of an n -digit integer N written to base m . All such vectors are generated by the simple Peano process $N + 1 \rightarrow N$, starting with $N = 0$, observing the ordinary carry rule of base m arithmetic, and halting as soon as $N = m^n$.

Such straightforward programming can be modified to cover the case in which each k has its own upper bound. This involves only a minor change

in the carry rule. A special case of this rectangular, rather than cubical, array of lattice points arises in connection with permutation problems discussed later.

Another variant of the carry rule takes care of the problem of generating vectors whose components are monotone, say,

$$0 \leq k_1 \leq k_2 \leq \cdots \leq k_n \leq m - 1,$$

by resetting the overgrown component k_i not by zero but by the newly increased value of k_{i-1} , with only slight complications in case of carry propagation. The case of strict monotoneity is handled by using $1 + k_{i-1}$ as a resetting value.

Suppose that a problem calls for vectors of n non-negative integers

$$(k_1, k_2, \dots, k_n)$$

subject only to the conditions that their sum

$$(2) \quad k_1 + k_2 + \cdots + k_n = C$$

has a prescribed constant value. Since for each component

$$(3) \quad 0 \leq k_i \leq C,$$

one could set $m = C + 1$ and use the unrestricted case program to generate each of the n^{C+1} vectors satisfying (3). Then the condition (2) could be imposed to eliminate nearly every candidate as soon as it is generated. This forthright procedure would be very wasteful of machine time producing large quantities of chaff for only a handful of wheat. A more efficient method is the following. We replace n by $n - 1$ and generate vectors

$$(\delta_1, \delta_2, \dots, \delta_{n-1})$$

for which the δ 's are monotone:

$$(4) \quad 0 \leq \delta_1 \leq \delta_2 \leq \cdots \leq \delta_{n-1} \leq C$$

as described above. Then we set

$$\begin{aligned} k_1 &= \delta_1 - 0 \geq 0, \\ k_2 &= \delta_2 - \delta_1 \geq 0, \\ &\vdots && \vdots \\ k_{n-1} &= \delta_{n-1} - \delta_{n-2} \geq 0, \\ k_n &= C - \delta_{n-1} \geq 0. \end{aligned}$$

It is clear that (2) and (3) are satisfied and to every instance of a vector of k 's satisfying these conditions there corresponds uniquely a vector of δ 's as described by (4).

If a problem requires the k 's to be positive, (4) may be replaced by the condition of strict monotoneity. Alternatively one may replace k_i by $h_i + 1$ and C by $C - n$ and solve the original problem for the h 's.

In some problems the parameters n , m , or C are so large that the number of vectors becomes unreasonably great. In such cases one may wish to make use of sampling methods. This may be done by replacing the methodical generation procedures by random variable generation in obvious ways. Thus in the preceding problem the δ 's are to be selected at random, not the first $n - 1$ k 's.

Another rather basic combinatorial problem is that of selecting, in turn, all possible combinations of n objects k at a time. For the machine, these n objects are words stored in the memory in addresses which without loss of generality may be taken to be $1, 2, \dots, n$. The problem is then to get the machine methodically to select k of these addresses and deliver the corresponding words to k other addresses which we may take to be $n + 1, n + 2, \dots, n + k$. It is worth pointing out that we do not care about the order in which the k selected words are arranged when delivered. To leave this matter to the discretion of the machine would be very unwise since it cannot deliver words in an unspecified manner like a human being drawing a handful of balls from an urn. The simplest way to specify this ordering is to insist on preservation of precedence; that is, if w_1 and w_2 are two selected words and if the address of w_1 in storage is less than that of w_2 the same shall be true on delivery. Thus if the words in storage are numbers in monotone sequence the same will be true of all the selected sets of k . With this additional requirement the subroutine can now be written. A diagram of such a routine is given in Figure 4. It exhibits a typical feature of many combinatorial routines. The sole purpose of the routine is to fetch k words from addresses $1(1)n$ and to deposit them into addresses $n + 1(1)n + k$. Yet there is only one command that does this. All the other commands are needed to process this single fetch and deposit instruction. We note that the arithmetic operations involved consist only of adding and subtracting unity. The number t is a tally that rises and falls between 1 and k and finally becomes zero after the last selection has been made. The numbers A_t are the selected addresses. The routine selects first the k words stored in addresses 1 through k and finally selects the k words in addresses $n - k + 1$ through n . If the n objects are the first n letters of the alphabet, the $C_{n,k}$ k -letter words not only will have the letters of each word in alphabetical order but the words themselves will be in lexicographical order.

Problems involving permutations occur frequently in combinatorial problems. There are, to the writer's knowledge, ten different methods of producing permutations automatically.

A method may be judged on its ability to satisfy one or more of the following requirements:

- (a) To generate all permutations,
- (b) To generate all favorable permutations,
- (c) To generate sample or random permutations,
- (d) To generate custom-made permutations.

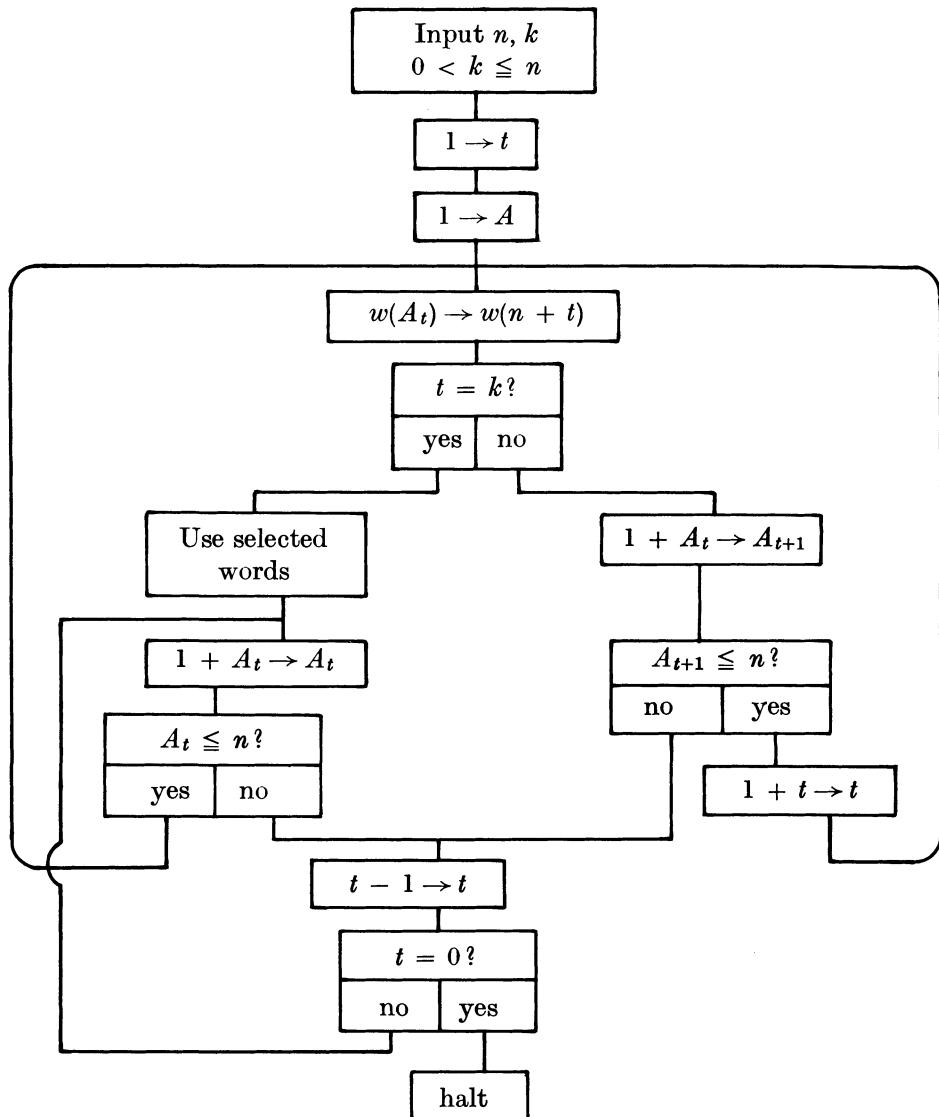


FIGURE 4

The following comments on these requirements may help to clarify the problem.

As far as (a) is concerned, it is well to note that $n!$ is approximately $1.55 \cdot 10^{25}$ for $n = 25$. Hence unless we can generate and utilize more than 10000 permutations per second our problem for $n = 25$ will last more than 10000 times the present age of the earth. For $n = 12$ the time will be about 13 hours. For $n = 6$ the time is negligible.

Because of the preceding, the ability of a method to meet requirement (b) is very important when n is more than 10 or 12. In this case we attempt to skip over millions of unwanted permutations in convenient blocks. For example in the travelling salesman problem with cities in two clusters as shown in Figure 5 the minimum circuit visiting all 10 cities and returning

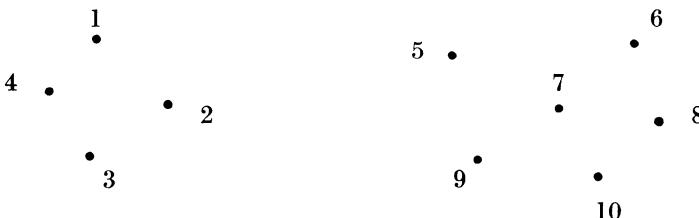


FIGURE 5

home would not be given by any permutation of the form 1 5 2 6 $x x x x x x$ no matter what digits the x 's represent. Hence we should skip over all 720 permutations that begin in this way. To get the machine to recognize and take advantage of this opportunity as quickly as a human being can, requires a human programmer with a suitable method.

As for (c), it may be possible in some problems when n is large to use random sampling of the huge population of $n!$ permutations to make shrewd guesses about some function of permutations. Acceptability tests for randomness of permutations have not been generally agreed upon as yet. S. Ulam, in a recent letter, suggests two such tests, one for frequency and one for gaps. In generating random permutations we are generating isolated ones. Thus our program differs greatly from those required by (b) and especially (a) where a recursive formula or algorithm is indicated.

The requirement (d) is often encountered with quite large values of n but with a relatively small number of admissible permutations. It has been possible to deal with $n = 20$ in one example, for instance. All requisite permutations were found in a few minutes by a systematic filling in of addresses by marks according to the very restrictive conditions imposed on the permutations.

A word or two about representing permutations in a computer may be in order. It is clear that the different objects being permuted can be made computer words. A permutation is then merely an assignment of these words or copies of these words to n memory cells. To store all $n!$ permutations in the machine at one time will require a small n and a big memory. Hence in most interesting problems the permutations are made to pass in review timewise. A subroutine inspects each permutation, rejects or gathers information about it, and then consigns it to extinction. However, in some cases one or more good-looking permutations are set aside for output.

Instead of permuting whole words full of information we can make a permanent file of these words in some fixed addresses and then proceed to permute these addresses instead. These addresses may be taken as $0, 1, 2, \dots, n - 1$. Now we are handling such small numbers that most of the arithmetic unit is processing zeros. This suggests that there should be some attempt to parallel the arithmetic by using "fractional precision" methods, that is by storing many marks in one word. In some programs this serves to speed up the generation considerably. To offset this advantage is the fact that the separate marks are less accessible individually without recourse to often fussy extract or shift operations.

In generating or counting permutations a special "factorial" representation of integers is often convenient. Every non-negative integer N less than $n!$ can be uniquely written

$$N = S_1! + S_2! + S_3! + \cdots + S_{n-1}(n-1)!$$

where the "factorial digits", S_k , satisfy

$$0 \leq S_k \leq k.$$

Thus, S_1 is a binary digit, S_2 a ternary digit, etc. Following the well-established backwards Arabic way of writing decimal digits we can also write

$$(5) \quad N = S_{n-1}, S_{n-2}, \dots, S_2, S_1.$$

Whatever the complexities of the arithmetic of this number representation, it cannot be accused of favoring any one particular base. There are two ways of computing the factorial digits of a given number N . To obtain the higher ordered digits first one simply divides N by $(n-1)!$. The quotient is S_{n-1} and the remainder is divided by $(n-2)!$ to obtain S_{n-2} , etc. To get the digits in reverse order, one divides N by 2; the remainder is S_1 and the quotient is divided by 3 to obtain S_2 , etc.

To have a method for generating permutations one has only to establish a one-to-one correspondence between the set of all permutations on n marks and the set of all $(n-1)$ -digit numbers of the form (5) or what is really the same, the set of all vectors of $n-1$ non-negative integer components, the k th component not exceeding k , or, again the same, the set of lattice points in and on an $(n-1)$ -dimensional rectangular parallelepiped of sides 1, 2, $\dots, n-1$.

Since we have already discussed the methodical recursive generation of such vectors by addition, using a simple variant of the usual carry rule, the corresponding permutations are duly generated also. If isolated or random permutations are needed it is a simple matter to generate isolated or random vectors of the right type.

It remains to establish such a one-to-one correspondence. This may be done in more than one way. Three ways are discussed in what follows. Four other entirely different methods will be described later.

The Tompkins-Paige cyclic method makes use of the fact that every permutation on n marks is the product of $n - 1$ cyclic permutations in the following sense. The simple permutation which replaces

$$\begin{array}{cccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array}$$

by

$$\begin{array}{cccccccccc} 0 & 1 & 2 & 6 & 7 & 8 & 9 & 3 & 4 & 5 \end{array}$$

may be said to be of order 7 and degree 3, by which we mean that the last 7 marks of the original permutation have been mounted on a wheel and the wheel rolled forward three spokes. If this result is subjected to permutation of order 6 and degree 0 it remains unchanged; but if the degree is 4 we thus get

$$\begin{array}{cccccccccc} 0 & 1 & 2 & 6 & 4 & 5 & 7 & 8 & 9 & 3. \end{array}$$

Clearly every permutation is the result of $n - 1$ superposed transformations of order $k + 1$ and degree S_k ($k = 1, 2, \dots, n - 1, 0 \leq S_k \leq k$), which are uniquely determined either by the permutation of the set of factorial digits S_k . Thus the correspondence is established. For $n = 3$ the correspondences are :

S_2	S_1	permutation
0	0	0 1 2
0	1	0 2 1
1	0	1 2 0
1	1	1 0 2
2	0	2 0 1
2	1	2 1 0

As set up for the machine, the recursive generation of the next permutation from a given one is done as follows. The last two marks of the given permutation are interchanged and 1 is added to the number

$$S_{n-1}, S_{n-2}, \dots, S_2, S_1.$$

If this produces no carry ($S_1 = 0$) we have our new permutation. If there is a carry to S_2 only, we have an old permutation which is now subjected to a cyclic permutation of order 3 and degree 1, which gives our new permutation. If, however, the carry propagates to S_3 and stops there, a cyclic permutation of order 4 and degree 1 is in order, etc. In this routine all cyclic permutations are of degree 1. This simple method has been coded for the SWAC and the IBM 701 and carefully polished for speed. For the SWAC the time required just to generate and count permutations is close to 2456 microseconds per permutation. The 701 is somewhat slower requiring nearly 3600 microseconds. For this timing, each mark is stored as an individual word. If several marks are stored in a single word, cyclic permutation can be paralleled by an obvious use of the shift command. This should speed the program considerably, though it would also delay

the actual use of the permutation in a practical problem. Of all the methods discussed, the Tompkins-Paige cyclic method is the fastest.

In some problems it is necessary to know whether a permutation is odd or even. The parity of the permutation corresponding to the Tompkins-Paige method is the same as that of the number

$$S_1 + 2S_2 + 3S_3 + \cdots + (n - 1)S_{n-1}.$$

We note that this routine does not generate permutations in lexicographical order since in the above table for $n = 3$ the permutation 120 precedes 102. For some purposes this may be a drawback.

If the first few marks of a permutation are somehow undesirable, the routine can easily be altered to skip over all the permutations that begin in this way simply by adding a unit to the appropriate S_k and performing the corresponding cyclic permutation. Thus the method meets requirement (b). It fails to satisfy requirements (c) and (d).

In an unpublished paper, Tompkins has suggested a different way of realizing the above correspondence which allows it to meet requirement (c). In this method the machine computes the mark located at a given address to produce the permutation directly from its corresponding vector of factorial digits.

We turn now to a second way of making a permutation correspond to its factorial digits. This method was suggested by Marshall Hall and may be called the Method of Derangements. In the previous method the objects being permuted can be any computer words. In the Hall method the objects must be the numbers $0(1)n - 1$. In any such permutation we may, for each mark $k > 0$, ask how many of the k marks less than k actually follow k . Denoting this number by S_k we see at once that

$$S_{n-1}, S_{n-2}, \dots, S_2, S_1$$

is a set of factorial digits of a number which corresponds to the given permutation and which, conversely, characterizes this permutation. We have for example the following correspondencies when $n = 7$.

S_6	S_5	S_4	S_3	S_2	S_1	permutation
0	0	0	0	0	0	0 1 2 3 4 5 6
3	1	4	1	2	1	4 2 1 6 3 5 0
1	2	2	3	1	1	3 1 4 5 2 6 0
6	5	4	3	2	1	6 5 4 3 2 1 0

The coding of this method is fairly straightforward. The resulting routine is a good deal slower than the Tompkins-Paige method. The parities of successive permutations strictly alternate. The method is well suited to requirement (c).

A third way of setting up a correspondence was, in effect, suggested by

D. N. Lehmer as long ago as 1906. It may be called the lexicographic method since it generates permutations in this order.

If in any permutation

$$a_1, a_2, \dots, a_n$$

of the numbers $0(1)n - 1$ we strike out a_1 and reduce by unity all the marks which exceed a_1 , we get a new permutation

$$\alpha_1, \alpha_2, \dots, \alpha_{n-1}$$

of the numbers $0(1)n - 2$, which we may denote by

$$M(a_1, a_2, \dots, a_n).$$

If we now define a rank function $R(a_1, a_2, \dots, a_n)$ recursively by

$$R(0) = 0,$$

$$R(a_1, a_2, \dots, a_n) = a_1(n - 1)! + R(M(a_1, a_2, \dots, a_n))$$

it is seen that R is nothing but the rank or serial number of the permutation (a_1, a_2, \dots, a_n) in the lexicographical list of all permutations. In fact, in this list the first $(n - 1)!$ permutations have 0 as their first mark, the next $(n - 1)!$ permutations have 1 as their first mark, and so on. Since our permutation has a_1 as its first mark it is preceded by $a_1(n - 1)!$ permutations whose first mark is less than a_1 . Among those permutations which begin with a_1 , ours has rank $R(M(a_1, a_2, \dots, a_n))$. If one successively applies the operation M we get a sequence of $n - 1$ permutations whose first elements are the factorial digits of the rank of the original permutation.

Thus for example, for $n = 7$, the permutation 1 4 2 0 5 6 3 gives rise to

$$\begin{array}{r} 1\ 4\ 2\ 0\ 5\ 6\ 3 \\ 3\ 1\ 0\ 4\ 5\ 2 \\ 1\ 0\ 3\ 4\ 2 \\ 0\ 2\ 3\ 1 \\ 1\ 2\ 0 \\ 1\ 0 \\ 0 \end{array}$$

Hence the rank of 1 4 2 0 5 6 3 is

$$1, 3, 1, 0, 1, 1 = 6! + 3 \cdot 5! + 4! + 2! + 1! = 1107.$$

Conversely given the rank and its factorial digits

$$S_{n-1}, S_{n-2}, \dots, S_2, S_1$$

we may reconstruct the permutation having this rank. Beginning with 0 we affix S_1 and in case S_1 is 0, we increase the original 0 to 1. Thus we get

$$\begin{aligned} 0\ 1 &\quad \text{if } S_1 = 0, \\ 1\ 0 &\quad \text{if } S_1 = 1. \end{aligned}$$

Inductively having reached a permutation of $0(1)k - 1$ we affix S_k and adjust upward by a unit those elements which exceed S_k . Finally S_{n-1} is attached and a final adjustment, if necessary, completes the permutation. Thus the millionth permutation on 10 marks is found as follows: The factorial digits of a million are

$$10^6 = 2, 6, 6, 2, 5, 1, 2, 2, 0.$$

Hence we write the succession of permutations

	0								
0	1								
2	0	1							
2	3	0	1						
1	3	4	0	2					
5	1	3	4	0	2				
2	6	1	4	5	0	3			
6	2	7	1	4	5	0	3		
6	7	2	8	1	4	5	0	3	
2	7	8	3	9	1	5	6	0	4

of which the last is the millionth. Thus the correspondence is established. The parity of the permutation of rank R is that of $[(R + 1)/2]$. The method satisfies requirements (a), (b) and (c).

We mention briefly four quite different methods based on various aspects of permutations.

The Walker Backtrack Method, given elsewhere in this volume in more general form, is described by him as "completely unsophisticated." One regards a permutation of the marks $0, 1, 2, \dots, (n - 1)$ as simply a vector whose components are taken from the non-negative integers $< n$ but are all distinct. One proceeds to construct such vectors starting from $0, 0, 0, \dots, 0, 0$ filling in at each opportunity the least available mark. When a permutation is completed the last two marks are removed and the penultimate address is filled by the next largest mark available. If there is no next largest element available one more mark is removed and replaced by the next larger available mark, etc. The result is a complete set of permutations in lexicographical order. This process was coded by Walker, for a general n , for the SWAC using only twenty-two commands. The program is slower than the Tompkins routine by a factor of two. I believe it could be altered to give random permutation and to skip over blocks of unwanted permutations. A similar program was devised by the writer to meet requirement (d) in 1955. Such programs are difficult to describe except in very general terms. In brief the machine keeps a sort of registry which shows at a glance which marks have been assigned to the permutation under construction and thereby avoids placing two marks in the same place and provides an automatic waiting list of marks as yet unassigned.

We pass on to what may be called the Constant Difference Method. Given a permutation like

$$2 \ 3 \ 1 \ 5 \ 4 \ 0 \ 7 \ 9 \ 6 \ 8$$

one can obtain immediately another one by increasing every mark by unity, replacing 9 by 0 rather than 10; thus

$$3 \ 4 \ 2 \ 6 \ 5 \ 1 \ 8 \ 0 \ 7 \ 9.$$

In fact, we get in this way 10 permutations all with the same set of differences modulo 10 between consecutive marks, namely

$$1 \ 8 \ 4 \ 9 \ 6 \ 7 \ 2 \ 5 \ 2.$$

One may take as representative of these 10 permutation whose first element is zero, namely

$$0 \ 1 \ 9 \ 3 \ 2 \ 8 \ 5 \ 7 \ 4 \ 6.$$

Similarly the permutation

$$1 \ 0 \ 3 \ 2 \ 8 \ 5 \ 7 \ 4 \ 6$$

in which we have taken the marks modulo 9, is one of 9 represented by

$$0 \ 8 \ 2 \ 1 \ 7 \ 4 \ 6 \ 3 \ 5.$$

This continues on down to the case of only two marks 0 1. This suggests the following method exemplified by the case of $n = 5$. We begin with the permutation 0 1 2 3 4. Adding 1 1 1 1 1 modulo 5 five times to return to 0 1 2 3 4. We now subtract 1 1 1 1 and then add it back again, this time modulo 4, obtaining 0 1 2 3 0. Once more we add 1 1 1 1, this time modulo 5, obtaining 0 2 3 4 1. This is our next permutation and there are four others it represents. Continuing we come to 0 4 1 2 3 which, after giving 1 0 2 3 4, 2 1 3 4 0, 3 2 4 0 1, 4 3 0 1 2, 0 4 1 2 3 gives rise in turn to

$$0 \ 3 \ 0 \ 1 \ 2, \ 0 \ 0 \ 1 \ 2 \ 3, \ 0 \ 0 \ 0 \ 1 \ 2, \ 0 \ 0 \ 1 \ 2 \ 0, \ 0 \ 0 \ 2 \ 3 \ 1$$

and finally 0 1 3 4 2, our next permutation. The process finally returns to 0 1 2 3 4.

This process has been coded for the SWAC and for the 701. It is about as fast as the Walker method. If permutations with specified properties of the differences between consecutive marks are required the process is very much faster than any previous one. An example of such a property is the requirement of the differences themselves forming a permutation as in cable splicing and other management problems. The method lends itself to fractional precision representation. For $n = 8$, for example, one permutation can be made from its predecessor in 128 microseconds on the SWAC.

Another method, called the Addition Method, may be explained briefly. Starting with $x = 0$ and programming " $x + 1$ replaces x " we can generate n^n n -digit numbers to the base n . Those numbers whose digits are distinct are the desired $n!$ permutations of $0, 1, \dots, n - 1$.

Of course for $n = 10$ this method would be very wasteful since only 1 number in $10^{10}/10! = 2756$ has distinct digits. To make this more efficient one should add (when possible) more than 1 to each successive number, in fact, as much as possible. To this effect we can formulate the following rule based on the notion of an “offending digit” of a number to the base n . If the digits of the number are not a permutation, the offending digit is the first digit which is equal to a preceding digit. If the digits are a permutation then the penultimate digit is the offending one. The rule of procedure now becomes: Add 1 to the offending digit (carrying to base n if necessary) and replace all succeeding digits by 0, 1, 2, 3, Starting with the number 0 1 2 . . . (n - 1) we thus obtain all permutations in lexicographical order. For example for $n = 3$ we have

$$\begin{array}{r} \underline{0 \ 1 \ 2} \\ 1 \ 1 \ 0 \end{array} \quad \begin{array}{r} \underline{0 \ 2 \ 0} \\ 1 \ 2 \ 0 \end{array} \quad \begin{array}{r} \underline{0 \ 2 \ 1} \\ 2 \ 0 \ 0 \end{array} \quad \begin{array}{r} 1 \ 0 \ 0 \\ \underline{2 \ 0 \ 1} \end{array} \quad \begin{array}{r} 1 \ 0 \ 1 \\ \underline{2 \ 1 \ 0} \end{array} \quad \begin{array}{r} \underline{1 \ 0 \ 2} \\ 2 \ 2 \ 0 \end{array}$$

where the permutations are underlined. The efficiency of this method depends upon the ease with which the machine can locate the offending digit. This problem can be solved by a registry method mentioned before in connection with the Walker method. The total number of numbers generated is $n - 1$ times the number of permutations produced. The method is capable of improvement.

An unpublished method of Ulam which might be called the Random Product Method is designed only to meet requirement (c). It makes use of the fundamental fact that a permutation of a permutation is another permutation, and is used to generate random permutations for n as large as 100 or more. Two permutations P_1 and P_2 are put into the machine. With probability $1/2$, P_1 or P_2 is chosen to be applied to P_1 to produce either

$$P_3 = P_1^2 \text{ or } P_3 = P_1 P_2,$$

a new permutation. P_3 in turn is multiplied by either P_1 or P_2 depending again on a random event of probability $1/2$. The process continues indefinitely. Tests of randomness are applied and new “starters” P_1 , P_2 are chosen if the tests show unsatisfactory characteristics.

In conclusion it is worth pointing out that special purpose electronic equipment attached to the arithmetic unit of a fast computer can add an order of magnitude to the speed of permutation problems. For example, a simple set of n ring counters of periods 2, 3, 4, . . . , n running in parallel would be very useful in the fractional precision handling of permutations. Various micro-programming techniques introduced into the design of a computer would make for more efficiency with permutation problems. Whatever improvements are made, however, one has only to increase n a little and obtain a hopeless problem about permutations.

This page intentionally left blank

ISOMORPH REJECTION IN EXHAUSTIVE SEARCH TECHNIQUES¹

BY

J. D. SWIFT

1. Introduction. In many problems involving exhaustive searches the limiting factor with respect to speed or completion may not be the efficiency with which the search as such is conducted but rather the number of times the same basic problem is investigated. That is, the search routine may be effective in rejecting impossible cases in large blocks and still fail to accomplish its purpose in that the cases which must be investigated are looked at too frequently. It is the purpose of this paper to investigate the problem in some detail particularly with respect to certain examples of searches completed, abandoned and projected.

2. Scope of the problem. We shall consider the following class of search problems: Initially k marks are designated. A sequence of length n whose elements are the given marks is to be produced subject to recursively expressible rules which determine an allowable subset of the marks at the j th place in terms of j and the previous $j - 1$ marks. These rules must not be absolute; specifically, the sequence resulting when an allowable sequence is subjected to a permutation of the marks must be allowable. Such a sequence will be called a (complete) solution. An allowable sequence of length $m < n$ is termed a partial solution. Partial or complete solutions are called isomorphic if they may be transformed into one another by a permutation of the marks. The problem is to determine all non-isomorphic complete solutions.

Certain problems may be reduced to a set of restricted problems, that is, problems in which an (allowable) initial sequence of length $r < n$ is given and it is required to find all non-isomorphic solutions with this initial segment.

Some of the standard problems which may be dealt with within this classification are problems of existence or listing of latin squares, greco-latin squares, semi-groups of fixed order, finite projective planes, Steiner triple systems and other finite algebraic structures. It is clearly of no importance to the definition that in actual search procedures elements of the sequence are frequently entered or altered in blocks or that, as in orthogonal latin square problems, two or more sequences are stored separately to facilitate computation as long as there is some fixed order of selection. On the other

¹ The preparation of this paper was sponsored in part by the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

hand many of the problems referred to are not normally stated in sequence language; the use of this terminology predicates a particular type of method of solution. This method is of so general use, particularly in connection with the employment of high speed digital machinery, that we may essentially subsume the original problem in it.

The general problem now clearly reduces to two parts, the employment of the rules for generating the sequences and the avoidance or rejection of isomorphic solutions. It is obviously desirable to recognize isomorphic partial solutions as early as possible for if $A = \langle a_1, a_2, \dots, a_n \rangle$ is such that the permutation P takes $A_r = \langle a_1, a_2, \dots, a_r \rangle$ to B_r , $A_r P = B_r$, then $AP = B$ where B has B_r as an initial segment. That is, all solutions with B_r as initials will be isomorphic to solutions with A_r as initials.

There are three points with quite different procedures in each case at which rejection of isomorphs may be made: (1) Before the regular search routine. (2) Within this routine. (3) After completion. Of these, the latter is by far the simplest though apparently the most wasteful. All solutions are obtained in one type of operation, then sorted for isomorphism. This method can clearly be used alone only for very short problems; on the other hand in almost all problems of the type considered where a number of solutions actually exist it is a necessary final step since practical methods in the first two stages usually are not comprehensive. The simplest method is an exhaustive search for a permutation which will bring the candidate sequence into the form of an accepted one. This can often be supplemented or possibly replaced by the method of invariants to be described later.

In the first stage of rejection, the procedure is to give an initial sequence or a set of sequences so chosen that every complete solution will be isomorphic to one having a prescribed initial. This procedure may be simple or elaborate. In the latter case it again requires an exhaustive search presumably of more elementary character than the basic routine to prepare the initial segments.

The second stage seems the most reasonable but is also the most difficult at which to perform the rejection. In terms of data processing this is an on-line operation while the others are off-line. The obvious disadvantages are that rejection routines interrupt and slow down the normal search procedure. In high speed calculation excessive transfer from routine to routine involving references to lower speed storage units cannot be tolerated. On the other hand the saving in time expended on looking at unnecessary cases, on output, on rehandling of data and on final or stage three reduction, is frequently worth considerable interruption.

The ideal qualities of stage two testing are: (1) Simplicity and rapidity of application; (2) Relative completeness; (3) Capability of integration with the search routine. There are two basically different methods available. The first is the use of a selected set of permutations applied at selected intervals, the completion of a line of a latin square for example. The

permutations are used to attempt to produce a sequence of lower signature; that is, one which has already occurred. To satisfy the first condition of simplicity it is necessary to abandon completeness. One system is to use the method in connection with a fairly large scale use of stage one rejection and to use the automorphism groups of the initial segments as the subset of permutations. It is then only necessary to permute the portion of the sequence formed in the final search routine. This procedure is clearly incomplete since no allowance is made for permutations which might transform the new part of the sequence to a part earlier than the given initial segment. As for integrability, the method demands a search routine capable of storing an easily calculable and comparable signature. A number of standard permutation routines satisfy this criterion [5].

The second method is the employment of easily calculable invariants. That is, quantities which are the same for all isomorphic sequences but which may vary among non-isomorphic sequences. In this case there is a great deal of room for ingenuity. If the problem may be viewed geometrically, the invariants may be expressed as configurations. In a problem dealing with projective planes or partial planes we might count Fano configurations, Desarguesian perspectivities, etc. When the sequence may be viewed as an algebraic multiplication table, sets of elements satisfying identities such as the associative law may be listed, idempotents counted, etc. Frequency counts furnish a quick and easy invariant for many problems. Of course, combinations of these and other ideas may be used. The satisfaction of the completeness requirement now depends as much on the ingenuity of the investigator as on a willingness to sacrifice simplicity. The ideal case is that of a simple complete set of invariants. The catch here is in the third requirement of integrability. A search routine normally is concerned with a particular order of permutations and any ordering of the invariants is likely to be covered almost at random by the permutations. Hence, it is often difficult to say which sets of values have occurred before or, in the common case of an incomplete set of invariants, to say whether all possible cases coming under one value set have been examined. A way out of the difficulty is offered by combining the invariant process with a rather elaborate stage one reduction so designed that particular sets of values of the invariants will be examined in turn. An example will be given in the discussion of Steiner Triple Systems below.

The invariant calculations are also useful as preliminary sorting procedures before a final reduction by a permutation search in stage 3. In case the invariant set is known to be complete, it may, of course, be substituted for a search.

3. Some examples. We begin with some routines which depended primarily on stage one reduction:

- (a) 10×10 *orthogonal latin squares*. A routine was coded by Mr. Frank

Meek under the general supervision of C. B. Tompkins and the author with the basic intention of making a good try at the production of two squares of the type indicated, in contradistinction to the common procedure of attempting to prove the conjecture of Euler that no such squares exist. Rejection was stage one only and extremely simple at that. Initial columns $1, 2, \dots, 10$ were assigned to each square and an initial row $1, 2, \dots, 10$ to the first square. Attention was concentrated on rapidity of search and efficiency of rejection. Columns were added to each square in succession, testing for latinity and orthogonality. Success would have meant a total of 20 columns. Seventeen columns were achieved several times. A brief investigation showed that after a number of hours the bulk of computing time was spent redoing isomorphic cases. A few tries at restarting the routine with new additional columns not yet investigated were made with no greater success. There was, of course, never any intention that this routine would run to completion; it was one of those things which would have been wonderful if it worked, but which didn't.

(b) *Veblen-Wedderburn systems of order 16.* This routine by Addison and Kleinfeld is aimed at the production of a complete list of such systems. Rejection was again entirely stage one but this time very elaborate. With considerable ingenuity the investigators designed a two-dimensional algebraic structure which would simplify the calculation by making unnecessary the generation and testing of a complete multiplication table. It was, in fact, possible to distinguish each system by a single line of the table. About 1200 systems were found. Stage three reduction of these systems was carried out later entirely by hand calculation. The expeditious reduction should have been accomplished (in the present author's opinion) by a new set of routines which would take in each system, create its whole multiplication table, count idempotents and associativities and sort on these invariants, then attempt to match within classes by a permutation search. The initial assumption was equivalent to requiring a sub-plane of order 4 in any plane determined by a system of order sixteen. Until the necessity of such a plane is finally shown the results can not be regarded as complete. The present list contains about 120 systems.

(c) *Uniqueness of the projective plane of order 8 [4].* Again the stage one procedure was elaborate. The method involved completing a latin square of order 7 to the affine part of the plane. There was available a list of the non-isomorphic squares of order seven constructed by Norton and corrected by Sade. These could be further reduced by an argument on the number of times a triangle could be the set of diagonal intersections of a quadrilateral. (References are in the paper indicated.) The remaining squares were used as starts and one or two individual elements for the other lines were also designated. Only two outputs remained for stage 3 which was easily accomplished by hand. It is clear here that the whole success of the investigation depended on the existence of the list of squares; if it had not

existed, either it would have needed to be constructed or a totally different approach made.

We now turn to two cases in which stage two reduction has been successfully employed.

(d) *Semigroups of order 5.* This is an as yet unpublished investigation by T. Motzkin and J. Selfridge. It is an illustration of skillful employment of reduction by permutations. The only stage one procedure was the obvious one of a fixed format for the multiplication table. A partial permutation test was made for every row completed. Instead of permuting the whole table so far created, the newly formed elements were investigated by testing it against $\phi(\phi^{-1}(i) \cdot \phi^{-1}(j))$ where ϕ is the permutation. The subgroup of permutations leaving the first element fixed was used since most semigroups had the property $0 \cdot i = 0$, i.e. the first row was entirely zero. When it was not, a simple frequency count invariant was used rejecting a new table if the new row had more identical elements than the first had zeros. The process, while incomplete, was highly efficient; 80 per cent of the outputs were retained after stage three reduction which used all $5!$ permutations. A previous investigation of the semigroups of order 4 by Forsythe [2] had used a less sophisticated procedure. The present method appears so powerful that it is considered possible to go to the case of order six.

(e) *Steiner Triple Systems of order 15* [3]. This routine used elaborate reduction procedures at all three stages. In each stage a system of invariants was used which can perhaps best be described in algebraic terms. R. Bruck [1] has described a system called a totally symmetric loop which is essentially a triple system with addition of a unit element, e , giving triples of the form $[a, a, e]$. A triple $[a, b, c]$ represents six equations in the loop, $ab = c$, $ac = b$, etc. For each pair a, b of elements whose rows and columns in the table have been determined we calculate the number of elements x satisfying successive identities of the type $((((xa)b)a)b \dots) = (((xb)a)b)a \dots$. The set of these values gives a set of invariants which turned out to be complete for order 15. Attempts to prove or disprove completeness in general have failed so far.

In the case of order 15 there are just four different possibilities for the 12 elements x which do not appear in the triple $[a, b, c]$.

(A) All elements may satisfy the identity of length 2 (a and b associate with all elements); (B) four may satisfy the length 2 case and eight the length 4 only; (C) all elements may satisfy $((xa)b)a = (((xb)a)b)$; (D) all elements may require the full case of length six. If the initial starts are made on say 3 rows and columns of the quasigroup table we may order these starts AAA , AAB , etc. and refuse systems in a later start which have any three pairs with earlier characteristics.

In the actual run this system was not fully employed primarily because we were slow to realize the advantages of stage two reduction. It was used

in the later parts of the routine and the stage one portion was used throughout. For final reduction the systems were sorted on the invariants and a permutation sought. This is described in detail in the publication of the results.

The question arises of the possibility of isolating the systems of order 19. On any estimate there will be several thousand of them. Hence, production must be limited so that stage three is at an absolute minimum. At the present no better way suggests itself than an adaptation of the system involving a count of the identity types listed (there are six cases to replace the previous four) coupled with a trial of a subset of the automorphism group of the initial entry. It would be extremely helpful to know whether the identities characterize the system or to find a complete set of calculable invariants.

REFERENCES

1. R. Bruck, *Some results in the theory of quasigroups*, Trans. Amer. Math. Soc. vol. 55 (1944) pp. 19–52.
2. G. Forsythe, *SWAC computes 126 distinct semigroups of order 4*, Proc. Amer. Math. Soc. vol. 6 (1955) pp. 443–447.
3. M. Hall and J. Swift, *Determination of Steiner Triple Systems of order fifteen*, Math. Tables Aids Comput. vol. 9 (1955) pp. 146–152.
4. M. Hall, J. Swift, and R. Walker, *Uniqueness of the projective plane of order eight*, Math. Tables Aids Comput. vol. 10 (1956) pp. 186–194.
5. C. Tompkins, *Machine attacks on problems whose variables are permutations*, Proceedings of the Symposia in Applied Mathematics vol. 6, New York, McGraw-Hill, pp. 195–211.

UNIVERSITY OF CALIFORNIA,
LOS ANGELES, CALIFORNIA

SOME DISCRETE VARIABLE COMPUTATIONS

BY

OLGA TAUSSKY AND JOHN TODD

This is in the nature of a preliminary report on some computations and computational problems with which we have been connected. Most of these were carried out on the SEAC—the National Bureau Standards Eastern Automatic Computer and since this machine is no longer in use for mathematical purposes, we shall not enter deeply into the coding and programming details. We are indebted to various colleagues for advice and help, in particular to Dr. Morris Newman and Mrs. Ida Rhodes and to Mr. S. N. Alexander and his engineering staff. Much of the work was carried out with the support of the Office of Naval Research. Some of the computations were in the nature of pilot calculations and we expect to complete these (or at least continue, or check them) on more powerful machines. Of the six problems which we discuss, three are combinatoric and three are number theoretic.

1. **A problem of Sierpiński.** There has been much work on the idea of congruence and its generalization. The following problem was stated to us in 1949 by W. Sierpiński (cf. [1; 3]).

We say that two sets A, B , which we may suppose for simplicity, to be subsets of the Euclidean line, are *congruent by n -tuple decomposition* if there is a decomposition of A into n disjoint sub-sets A_1, A_2, \dots, A_n and one of B into n disjoint sub-sets B_1, B_2, \dots, B_n such that each pair of sets A_i and B_i are congruent in the sense of elementary Euclidean geometry. We shall denote this relation by

$$A \xrightarrow{n} B.$$

The problem is: If

$$A \xrightarrow{3} B$$

does there always exist a set C such that

$$A \xrightarrow{2} C, B \xrightarrow{2} C?$$

The answer to this, when the sets A and B are “general,” is “yes.” For if we have

$$A = A_1 + A_2 + A_3; \quad B = B_1 + B_2 + B_3; \quad A_i \cong B_i, i = 1, 2, 3$$

we can construct a set C by taking A_1 and A_2 as they lie in A and adjoining B_3 in such a position that $A_2 + B_3$ is congruent to $B_2 + B_3$ in their position in B . Diagrammatically:

A	\dots	\dots	\dots
B	\dots	\dots	\dots
C	\dots	\dots	\dots

This construction fails if the copy of B_3 in C has points in common with the copy of A_1 .

It has been shown by A. Sharma that the answer was always affirmative when A contained at most 7 points. The case of 8 points was studied by G. A. Dirac and John Todd in 1949. It was found possible, by elementary arguments, to dispose of all sub-cases except those where one of the sets A_i, B_i contained two points, and the other two three each. The following pair of sets was constructed so that they could not be dealt with using the criteria at our disposal :

$$\begin{array}{ll} A & 0, 1, 11; \quad 4, 5, 18; \quad 8, 14 \\ B & 0, 1, 11; \quad 7, 8, 21; \quad 4, 10 \end{array}$$

although it was clear that the question of the existence or nonexistence of a corresponding set C is a finite computational problem.

This problem remained untouched until it could be tried on SEAC. The detailed coding was carried out by R. T. Moore.

Conceptually, all partitions of A into two parts had to be considered and then all sets C_A formed by translating and reflecting these; and similarly for B . If no set C_A coincides with a set C_B then the pair A, B form a counterexample.

Considerations of the structure of the sets A, B indicated that it was just possible to do this in a geometrical way, plotting the points as binary ones in the 45 binary digit word of SEAC. As the reflection of a word seemed rather awkward to program, it was decided to begin with the case when only translations were considered. Later it was decided, again in the translation case, to use a more arithmetic approach.

After some 15 minutes running the program produced a set C :

$$C: \quad 1, 4, 14, 18, 19, 22, 25, 28.$$

The following decomposition of C

$$C: \quad 1, 4, 18; \quad 8, 14, 19, 22, 25, 28$$

is equivalent to the following decomposition of A

$$A: \quad 1, 4, 18; \quad 0, 5, 8, 11, 14$$

while this decomposition of C

$$C: \quad 1, 4, 14; \quad 18, 19, 22, 25, 28$$

is equivalent to the following decomposition of B

$$B: \quad 8, 11, 21; \quad 0, 1, 4, 7, 10.$$

The problem rests at this stage. A next step could be the examination of the set C just obtained and the setting of some general criteria which

cover this case, and an attempt to see whether these criteria would cover all the remaining cases. For current work see Cahn and Straus [2].

2. The Baker-Campbell-Hausdorff coefficients. Let $e^z = e^x e^y$, where x, y are noncommuting variables and the relation is interpreted as one between formal power series. It is possible to express

$$z = \log e^x e^y = \log \left[1 + x + y + \frac{x^2}{2!} + \frac{y^2}{2!} + \dots \right]$$

in terms of x, y and their commutators. Let us write

$$[AB] = AB - BA.$$

Then we have

$$\begin{aligned} z &= (x + y) + \frac{1}{2} [xy] + \frac{1}{12} \{ [[xy]y] + [[yx]x] \} + \frac{1}{24} [[[yx]x]y] \\ &\quad + \left\{ -\frac{1}{720} [[[xy]y]y]y - \frac{1}{720} [[[yx]x]x]x \right. \\ &\quad + \frac{1}{360} [[[xy]y]y]x + \frac{1}{360} [[[yx]x]x]y \\ &\quad \left. - \frac{1}{120} [[[xy]y]y]y - \frac{1}{120} [[[yx]x]y]x \right\} \\ &\quad + \dots \end{aligned}$$

No explicit formula for the coefficients seems to be known, and until recently only the coefficients of the commutators of order n , $n \leq 6$, were known.

This expansion has appeared in various chapters of pure and applied mathematics, e.g. group-theory, differential equations, statistical mechanics.

It is not difficult to see how these coefficients can be obtained mechanically. For instance it is possible to obtain all the early coefficients in the expansion of

$$\begin{aligned} \log[1 + (\exp x \exp y - 1)] &= (\exp x \exp y - 1) - \frac{1}{2} (\exp x \exp y - 1)^2 \\ &\quad + \frac{1}{3} (\exp x \exp y - 1)^3 + \dots \end{aligned}$$

and then group them as commutators.

Considerations of machine time suggested that it would be possible to do this easily in the high-speed memory of SEAC, for $n \leq 6$, and using the external tape memories, up to $n = 10$. One way of handling the basic operation of multiplication of two terms in the expansion say $axyx$ and $bxyxy$ where the a, b are rational numbers—which we can assume to be integers, with suitable scaling—is the following. Represent x by 0 and y

by 1 and assume we consider terms of degree at most 10. We use the sign bit and the first 33 bits of a SEAC word to represent a and the last 11 bits to represent the variables. We use a 1 as a flag to indicate the start of the word. Thus

$$\begin{array}{ll} xyxyxyxyxy & \text{is represented by } 1010101010 \\ xyx & \text{by } 00000001010 \\ xyxy & \text{by } 00000010101. \end{array}$$

It is possible to program a search for the flags and the true terms having been obtained it is easy to juxtapose these in the right order with a flag to obtain the product. The arithmetic product of the coefficients is obtained and placed in the correct position.

This problem was suggested to K. Goldberg, who used a slightly different representation, and the product up to order 6 was obtained after 40 seconds computing.

The program for the higher terms, using the external memory, was never written for the analysis of the simpler problem suggested to Goldberg [4: 5] an alternative approach to the problem which enabled the range originally planned to be covered by a 3-hour hand computation. It does not appear feasible to go much further by hand.

3. Covering theorems for groups. The source of the problems to be discussed was legal proceedings concerning copyright.

The basic idea of a (British) Football Pool is the successful forecast of the outcome (win, lose or draw) of a number m —a typical value of m is 13—of football matches. Each forecaster stakes a small sum, and the most successful wins the total amount staked. Even for small m , 3^m is large and so it is not feasible to cover all cases and so make certain of at least sharing the pool. As a matter of practice the forecaster assumes that the outcome of $m - n$ matches is certain and n doubtful and he “permutes” these results. If his original assumption is correct and he invests 3^n forecasts, then he will be successful. It has been observed that very often, no forecaster is completely correct and the pool is won with, e.g., $m - 1$ correct results. The question then arises as to what is the most efficient way of making forecasts of n games so that, no matter what happens, at least one of them will have $n - 1$ correct results.

A typical result is that it is possible to achieve this with 9 forecasts in the case of 3 games, when 27 would have been needed to be certain of being completely correct. In some cases the forecaster may assume that only two outcomes are possible—e.g. the home team cannot lose—and here, in the case of $n = 7$, while $2^7 = 128$ forecasts are necessary to insure a complete correct forecast, it is possible to choose 16 forecasts which ensure that at worst there will be one forecast with 6 results correct.

These problems can be set up in the following form. Let G be an Abelian group, with n base elements, each of the same order p (p not necessarily a

prime). Let S denote the set of $r = n(p - 1) + 1$ distinct powers of the base elements. Let AB , where A, B are two sub-sets of G , denote the set of all products ab where $a \in A, b \in B$. The problem is to determine the smallest integer $\sigma = \sigma(n, p)$, for which there exists at least one set H , consisting of σ elements of G , such that $G = HS$.

Various results have been obtained (cf. [6; 7; 8; 9; 10; 11]). We shall not discuss these now, but want to mention briefly a practically interesting case, which is not yet solved. Consider the case $n = 5, p = 3$. It is clear that we can choose 27 forecasts which will include at least one with at most one error: we take the 9 forecasts which are known to be satisfactory for the $n = 4, p = 3$, case and adjoin an additional coordinate with three values. It is also clear that at least 23 forecasts are required. Any particular forecast covers (itself) and 10 others. Hence, even assuming no overlapping, since $22 \times 11 = 242 < 243$, we cannot be certain with 22 forecasts. A little more complicated argument shows that 23 forecasts will not suffice. Whether the minimum is 24, 25, 26 or 27 does not seem to be known yet.

4. Legendre-Sophie St. Germain criterion. Fermat's "last theorem" asserts that the equation

$$(1) \quad x^l + y^l = z^l$$

has no solution in integers x, y, z , none zero if l is a prime number which exceeds 2. If x, y, z are all assumed prime to each other and to l then it is customary to refer to this as the first case of Fermat's last theorem.

Various criteria have been developed for testing the assertion. Among them is the famous criterion of Legendre and Sophie St. Germain (see e.g. [12]):

If there exists an odd prime p such that

$$(2) \quad x^l + y^l + z^l \equiv 0(p)$$

has no integral solution x, y, z each not divisible by p and such that the congruence

$$u^l \equiv l(p)$$

has no solution u then (1) has no solution prime to l .

Saying that (2) has no integral solution each not divisible by p is equivalent to saying that there are no two l th power residues mod p which are consecutive.

Not all primes p have to be tested when applying this criterion, only primes up to a certain bound. Professor H. Hasse suggested the tabulation of the four sets of p 's below this bound for which either both conditions of the criterion are satisfied or none or one or the other. Although the range we coded was only small the sets showed interesting properties. The criterion may not succeed as a proof or disproof of Fermat's last theorem, but the study of these four sets may reveal interesting occurrences.

The coding of the problem was mainly carried out by Dr. J. C. P. Miller. The code was only applied to values of $l < 100$, $p < 25,000$ and residues were only tested up to 10,000. The range $p < 25,000$ was necessary since it is the limit of the existing tables of primitive roots and without the use of primitive roots for p the work would be prohibitive. In the meantime primitive roots are being computed at various centers which will be helpful if computations are resumed.

Vandiver [12, Theorem VII] using the criterion of Legendre and St. Germain proved :

If (2) has no integral solutions prime to p for all $p = 1 + ml$ and $m < 10l$ then (1) has no solutions in integers each prime to l .

The range here is smaller and since the codes employed for the Legendre and St. Germain theorem examined the whole range instead of searching for one value of p it could be used for the test.

Another congruence also connected with Fermat's last theorem is

$$ax^n + 1 = by^l(p).$$

The solutions of this were computed by E. Lehmer and H. S. Vandiver [13], and also Vandiver [14].

5. Units in quadratic fields. The following conjecture was made by Ankeny, Artin and Chowla [15]. Let ϵ be the fundamental unit in the quadratic field $R(\sqrt{d})$, where d is a positive rational integer. It is known that ϵ is of the form $(t + u\sqrt{d})/2$ where $(t^2 - du^2)/4 = \pm 1$. The conjecture asserts that for p a prime we have $u \not\equiv 0 \pmod{p}$.

The conjecture was verified by its authors for all $p \equiv 5(8)$, $p < 2000$. It was subsequently tested on SEAC up to all $p < 100,000$. The fundamental unit of $R(\sqrt{d})$ was obtained from the continued fraction expansion of \sqrt{d} , see [16]. The actual coding was carried out by K. Goldberg. Since t and u can become very large the coding was so arranged that multiples of p were subtracted whenever possible.

Later L. Carlitz [17] showed that a value of u divisible by p can only happen if the Bernoulli number $B_{(p-1)/2} \equiv 0 \pmod{p}$, more precisely, he showed that $uh/t \equiv B_{(p-1)/2} \pmod{p}$ where h denotes the class number of $R(\sqrt{p})$.

6. Groups of classes of quadratic forms. It is known that there is a 1-1 correspondence between the ideal classes of a quadratic field $R(\sqrt{D})$ and the classes of quadratic forms $ax^2 + bxy + cy^2$ with $(a, b, c) = 1$ and a square free $D = b^2 - 4ac$. Quadratic forms were studied extensively by Gauss. He considered them, however, in the form $ax^2 + 2bxy + cy^2$ with $(a, b, c) = 1$ and $D = b^2 - ac$, while the above definition was used by Kronecker. The number of classes of these forms for a fixed negative $D \equiv 1(4)$ is equal in both cases unless $D \equiv 5(8)$ and $D \neq -3$, in which case

the number of Gauss form classes become 3 times as large, see [19, p. 243]. In particular, the number of Gauss form classes is always at least equal to 3 in the latter case.

It can be shown that the group of classes of Kronecker forms is isomorphic with a quotient group of the group of classes of all Gauss forms, namely, the quotient group with respect to the cyclic group of order 3 generated by the Gauss form class of $4x^2 + 2xy + ((-D + 1)/4)y^2$. (See [19, p. 400 and e.g. 18, p. 76 and 20]). In the two last mentioned references ring ideal classes are considered instead of form classes. A completely elementary proof for this can also be given. G. Pall [21, p. 791] shows that every Kronecker form corresponds to the following three Gauss form classes :

$$\begin{aligned} ax^2 + 2bxy + 4cy^2, \quad 4ax^2 + 2bxy + cy^2, \\ 4ax^2 + 2(2a + b)xy + (a + b + c)y^2. \end{aligned}$$

In particular the Kronecker principal class corresponds to the three classes

$$\begin{aligned} x^2 + 2xy + (-D + 1)y^2, \quad 4x^2 + 2xy + ((-D + 1)/4)y^2, \\ 4x^2 + 6xy + ((-D + 9)/4)y^2 \end{aligned}$$

which form a cyclic group of order 3. The first three mentioned classes are the products of one of them and the three classes corresponding to the principal class, i.e. they form a coset with respect to the subgroup of order 3. That this correspondence between the cosets and the Kronecker forms is an isomorphism follows from the following argument : Consider two Kronecker forms with the same D . It is known that they are equivalent to two "united" forms $ax^2 + bxy + a'cy^2$, $a'x^2 + bxy + acy^2$, $(a, a') = 1$ and that their product is $aa'x^2 + bxy + cy^2$. Since we may pick out any element in the coset we may assume that these forms correspond to the Gauss form classes

$$4ax^2 + 2bxy + a'cy^2, \quad 4a'x^2 + 2bxy + acy^2.$$

The first of these forms splits into the factors $(ax^2 + 2bxy + 4a'cy^2)$, $(4x^2 + 2bxy + aa'cy^2)$. The forms $ax^2 + 2bxy + 4a'cy^2$ and $a'x^2 + 2bxy + 4acy^2$ are again united and give the product $aa'x^2 + 2bxy + 4cy^2$ which lies in the coset corresponding to the Gauss form $aa'x^2 + bxy + cy^2$. The remaining two factors are equivalent with factors in the subgroup of order 3.

Let us call the subgroup of the Abelian group of form classes whose orders are powers of 3, the 3-class group. It is clear that either a new base class of order 3 is added or some class $\neq 1$ has its order tripled when going from the quotient group to the full group. It seemed of interest to study for what values of D the number of base classes is increased. Apart from the trivial case when the number of Gauss classes is 3, this seems to be the less frequent occurrence.

A table giving these values for negative square free D was constructed up to $|D| < 10,000$.

Various methods could be used for this problem. The method which was used was suggested by E. C. Dade who also carried out the coding.

The number of classes is easily obtained by using the fact that every form is equivalent to a uniquely determined reduced form. The set of reduced forms is then easily enumerated. The structure of the group is found by using a set of generators and a matrix of relations between them. This matrix can be reduced to triangular form by unimodular matrix factors in a routine process. The generators can be taken as the forms $ax^2 + 2bxy + cy^2$ with $a = 4$ or p_i where p_i are distinct odd primes. Since only reduced forms need to be considered the p_i are bounded.

It is known that -3299 is the first D with a noncyclic 3-class group of ideal classes. However $D = -307$ is the first D with a noncyclic 3-class group of Gauss forms.

BIBLIOGRAPHY

1. W. Sierpiński, *Sur quelques problèmes concernant la congruence des ensembles de points*, Elem. Math. vol. 5 (1950) pp. 1-4.
2. A. S. Cahn and E. G. Straus, *On piece-wise congruence*. Preliminary report, Notices American Math. Soc. vol. 6 (1958) p. 360.
3. W. Sierpiński, *On the congruence of sets and their equivalence by finite decomposition*, Lucknow University Studies, Lucknow, India, no. 20, 1954, 117 pp.
4. K. Goldberg, *The formal power series for $\log e^{xey}$* , Duke Math. J. vol. 23 (1956) pp. 13-22.
5. ———, *The formal power series for $\log e^{xey}$, II*. Ph.D. Thesis, American University, Washington, D.C., 1957.
6. O. Taussky and J. Todd, *Covering theorems for groups*, Ann. Soc. Polonaise de Math. vol. 21 (1948) pp. 303-305.
7. S. K. Zaremba, *A covering theorem for Abelian groups*, J. London Math. Soc. vol. 26 (1950) pp. 71-72.
8. ———, *Covering problems concerning Abelian groups*, J. London Math. Soc. vol. 27 (1952) pp. 242-246.
9. J. G. Mauldon, *Covering theorems for groups*, Quart. J. Math. Oxford ser. (2) vol. 1 (1950) pp. 284-287.
10. B. Kuttner, in a talk delivered to British Assoc. Adv. Sc., 1950.
11. E. Mattioli, *Sopra una particolare proprietà dei gruppi abeliani finiti*, Ann. Scuola Norm. Super. Pisa ser. (3) vol. 3 (1950) pp. 59-65.
12. H. S. Vandiver, *Fermat's last theorem*, Amer. Math. Monthly vol. 53 (1946) pp. 555-578.
13. Emma Lehmer and H. S. Vandiver, *On the computation of the number of solutions of a certain trinomial congruence*, J. Assoc. Comput. Mach. vol. 4 (1957) pp. 505-510.
14. H. S. Vandiver, *New types of trinomial congruence criteria applying to Fermat's last theorem*, Proc. Nat. Acad. Sci. vol. 40 (1954) pp. 284-252.
15. N. C. Ankeny, E. Artin and S. Chowla, *The class-number of real quadratic number fields*, Ann. of Math. vol. 56 (1952) pp. 479-493.
16. E. L. Ince, *Cycles of reduced ideals in quadratic fields*, British Assoc. Math. Tables vol. 2, London, 1934.
17. L. Carlitz, *Note on the class number of real quadratic fields*, Proc. Amer. Math. Soc. vol. 4 (1953) pp. 535-537.

18. R. Fueter, *Vorlesungen über die singulären Moduln und die komplexe Multiplikation der elliptischen Funktionen*, Leipzig-Berlin, 1924.
19. P. G. L. Dirichlet, *Vorlesungen über Zahlentheorie*, Braunschweig, 1879.
20. J. Sommer, *Vorlesungen über Zahlentheorie*, Leipzig-Berlin, 1907.
21. G. Pall, *Binary quadratic discriminants differing by square factors*, Amer. J. Math. vol. 57 (1935) pp. 789-799.

CALIFORNIA INSTITUTE OF TECHNOLOGY,
PASADENA, CALIFORNIA

This page intentionally left blank

SOLVING LINEAR PROGRAMMING PROBLEMS IN INTEGERS¹

BY

RALPH E. GOMORY

The problem of finding the best solution in integers to a linear programming problem arises naturally in several ways. In an allocation of resources problem involving indivisible items such as ships, a solution involving fractions is one that cannot be realized in practice. Sometimes this matters and sometimes not. If the numbers involved are large and especially if the data is as poor as it sometimes can be in practical problems, one can round off to the nearest integer and probably not make too great an error. However, when a combinatorial problem is formulated as a linear programming problem, as in Dantzig [1], the data is usually quite precise and the numbers in the solution are often restricted to be zero or one. An example would be the task of selecting the largest possible expedition from a group of available people, subject to certain restrictions such as "persons 7 and 8 cannot both be included in the expedition." This is converted into a linear programming problem by assigning variables x_i to each of the people. In a solution $x_i = 0$ means that a person is included in the expedition, $x_i = 1$ means he is excluded. The problem then is to minimize $\sum_i x_i$ subject to a series of restrictions such as

$$x_7 + x_8 \geq 1.$$

If a solution can be obtained in which the variables are 0 or 1, the inequality above implies that either person 7 or 8 has been excluded from the expedition. In a problem such as this one the numbers are automatically small (0 or 1) and a fractional solution is meaningless. In this particular example it is enough to demand a solution in integers. The minimization condition then assures that the solution contains only zeros and ones.

A close connection also exists between integer programming problems and problems involving piecewise linear, but not convex, domains or objective functions.

This paper outlines a finite algorithm for obtaining integer solutions to linear programs. The algorithm has been programmed successfully on an E101 computer and used to run off the integer solution to small (seven or less variables) linear programs completely automatically.²

¹ This work was supported in part by the Princeton-IBM mathematics research project.

² More recently a FORTRAN program on an IBM 704 has been used to run problems up to $m = n = 15$. The problems (only a few were run) ran rapidly.

The algorithm closely resembles the procedures already used by Dantzig, Fulkerson and Johnson [2], and Markowitz and Manne [3], to obtain solutions to discrete variable programming problems. Their procedure is essentially this. Given the linear program, first maximize the objective function using the simplex method, then examine the solution. If the solution is not in integers, ingenuity is used to formulate a new constraint that can be shown to be satisfied by the still unknown integer solution but not by the non-integer solution already attained. This additional constraint is added to the original ones, the solution already attained becomes non-feasible, and a new maximum satisfying the new constraint is sought. This process is repeated until an integer maximum is obtained, or until some argument shows that a nearby integer point is optimal.

What has been needed to transform this procedure into an algorithm is a systematic method for generating the new constraints. A proof that the method will actually give the integer solution in a finite number of steps is also important. This paper will describe an automatic method of generating new constraints. The proof of the finiteness of the process will be given separately.

Let us suppose that the original inequalities of the linear program have been replaced by equalities in nonnegative variables, so that the problem is to find nonnegative integers, $z, x_1, \dots, x_m, t_1, \dots, t_n$, satisfying

$$(1) \quad \begin{aligned} z &= a_{0,0} + a_{0,1}(-t_1) \cdots a_{0,n}(-t_n) \\ x_1 &= a_{1,0} + a_{1,1}(-t_1) \cdots a_{1,n}(-t_n) \\ &\vdots \\ x_m &= a_{m,0} + a_{m,1}(-t_1) \cdots a_{m,n}(-t_n) \end{aligned}$$

such that z is maximal. Using the method of pivot choice given by the simplex (or dual simplex) method, successive pivots result in leading the above array into the standard simplex form,

$$(2) \quad \begin{aligned} z &= a'_{0,0} + a'_{0,1}(-t'_1) \cdots a'_{0,n}(-t'_n) \\ x'_1 &= a'_{1,0} + \cdots \cdots a'_{1,n}(-t'_n) \\ &\vdots \\ x'_m &= a'_{m,0} + \cdots \cdots a'_{m,n}(-t'_n) \end{aligned}$$

where the primed variables are a rearrangement of the original variables and the $a'_{0,j}$ and $a'_{i,0}$ are nonnegative. From this array the simplex solution $t'_j = 0$, $x'_i = a'_{i,0}$ is read out.

An additional constraint can now be formulated. The constraint which will be generated is not unique, but is one of a large class that can be produced by a more systematic version of the following procedure.

Let us consider the equivalence relation, equivalence modulo 1.

We will write $a \equiv b$ (a equivalent to b) if and only if $a - b$ is an integer. This equivalence relation will be used to produce a new constraint.

If the $a'_{i_0,0}$ are not all integers, select some i_0 with $a'_{i_0,0}$ non-integer. For this i_0 we have the equation

$$(3) \quad x'_{i_0} = a'_{i_0,0} + \sum_{j=1}^{j=n} a'_{i_0,j}(-t'_j).$$

Any nonnegative *integer* solution $z'', x'', \dots, x''_m, t''_1, \dots, t''_n$ must satisfy (3). Since x''_{i_0} is an integer we have

$$(4) \quad x''_{i_0} \equiv 0.$$

So from (3) and (4) we have

$$(5) \quad \sum_{j=1}^{j=n} a'_{i_0,j}t''_j \equiv a'_{i_0,0}.$$

If $\bar{a}_{i_0,j}$ is any number equivalent to $a'_{i_0,j}$, then since the t''_j are integers,

$$(6) \quad \sum_{j=1}^{j=n} \bar{a}_{i_0,j}t''_j \equiv a'_{i_0,0}.$$

If the $\bar{a}_{i_0,j}$ chosen are all nonnegative, then the left hand side of (6) is also nonnegative since the t''_j are nonnegative by assumption. So the left hand side of (6) is both nonnegative and equivalent to $a'_{i_0,0}$. This implies

$$(7) \quad \sum_{j=1}^{j=n} \bar{a}_{i_0,j}t''_j \geq f'_{i_0,0}$$

where $f'_{i_0,0}$ is the fractional part of $a'_{i_0,0}$.

This inequality, although satisfied by any nonnegative *integer* solution to (2) is not satisfied by the present simplex solution since the simplex solution has $t'_j = 0, j = 1, n$.

The only restrictions placed on the $\bar{a}_{i_0,j}$ so far is that they should be nonnegative and equivalent to the $a'_{i_0,j}$. If any one of the chosen $\bar{a}_{i_0,j}$ is replaced by a smaller equivalent number which is still nonnegative, the result is a new inequality which is easily seen to imply the old one. A succession of such replacements then results in a series of increasingly strong inequalities. The strongest possible one, which implies all the others is

$$(8) \quad \sum_{j=1}^{j=n} f'_{i_0,j}t''_j \geq f'_{i_0,0}$$

where $f'_{i_0,j}$ are the fractional parts of the $a'_{i_0,j}$. That is $f'_{i_0,j} = a'_{i_0,j} - n'_{i_0,j}$ where $n'_{i_0,j}$ is the largest integer $\leq a'_{i_0,j}$.

To transform (8) into an equation we introduce the variable s_1 , required to be nonnegative, by

$$(9) \quad s_1 = -f'_{i_0,0} - \sum_{j=1}^{j=n} f'_{i_0,j}(-t'_j)$$

and add equation (9) to the set (2). The new set will be referred to as (2*). Since s_1 is the difference between the left and right sides of (8), and these

EXAMPLE

Maximize $w = 2x + 3y + z$

$$8x - 8y \leq 7$$

$$-x + 6y \leq 9$$

$$x + y + z \leq 6$$

Introduce slacks t_1, t_2, t_3

$$1 \quad -x \quad -y \quad -z$$

$w =$	0	-2	-3	-1
$t_1 =$	7	8	-8	0
$t_2 =$	9	-1	6*	0
$t_3 =$	6	1	1	1

$$1 \quad -x \quad -t_2 \quad -z$$

$w =$	$4\frac{1}{2}$	$-2\frac{1}{2}$	$\frac{1}{2}$	-1
$t_1 =$	19	$6\frac{2}{3}^*$	$1\frac{1}{3}$	0
$y =$	$1\frac{1}{2}$	$-\frac{1}{6}$	$\frac{1}{6}$	0
$t_3 =$	$4\frac{1}{2}$	$1\frac{1}{6}$	$-\frac{1}{6}$	1

$$1 \quad -t_1 \quad -t_2 \quad -z$$

$w =$	$11\frac{25}{40}$	$\frac{15}{40}$	1	-1
$x =$	$2\frac{34}{40}$	$\frac{6}{40}$	$\frac{8}{40}$	0
$y =$	$1\frac{39}{40}$	$\frac{1}{40}$	$\frac{8}{40}$	0
$t_3 =$	$1\frac{7}{40}$	$-\frac{7}{40}$	$-\frac{16}{40}$	1*

	+1	$-t_1$	$-t_2$	$-t_3$
$w =$	$12\frac{32}{40}$	$\frac{8}{40}$	$\frac{24}{40}$	1
$x =$	$2\frac{34}{40}$	$\frac{6}{40}$	$\frac{8}{40}$	0
$y =$	$1\frac{39}{40}$	$\frac{1}{40}$	$\frac{8}{40}$	0
$z =$	$1\frac{7}{40}$	$-\frac{7}{40}$	$-\frac{16}{40}$	1
$s_1 =$	$-\frac{39}{40}$	$-\frac{1}{40}$	$-\frac{8}{40}^*$	0



End of regular simplex method

$$+1 \quad -t_1 \quad -s_1 \quad -t_3$$

$w =$	$9\frac{7}{8}$	$\frac{1}{8}$	3	1
$x =$	$1\frac{7}{8}$	$\frac{1}{8}$	1	0
$y =$	1	0	1	0
$z =$	$3\frac{1}{8}$	$-\frac{1}{8}$	2	1
$t_2 =$	$4\frac{7}{8}$	$\frac{1}{8}$	-5	0
$s_2 =$	$-\frac{7}{8}$	$-\frac{1}{8}^*$	0	0



$$+1 \quad -s_2 \quad -s_1 \quad -t_3$$

$w =$	$\frac{9}{1}$	1	3	1
$x =$	$\frac{1}{1}$	1	1	0
$y =$	$\frac{1}{1}$	0	1	0
$z =$	$\frac{4}{1}$	-1	2	1
$t_2 =$	$\frac{4}{1}$	1	-5	0
$t_1 =$	$\frac{7}{1}$	-8	0	0

Integer solution

two sides are equivalent for any integer solution, s_1 will always be integer whenever the other variables are, so we still require all the variables appearing in (2*) to be integers.

The procedure now is to maximize z over the solutions to (2*). This is done using the dual simplex method because all the $a'_{0,j}$ and $a'_{i,0}$ are already nonnegative, and $-f'_{i_0,0}$ is the only negative entry in the zero column of the equations (2*). This fact usually makes remaximization quite rapid. The process is then repeated if the new simplex maximum is non-integer.

Of course the equations (2*) involve one more equation than the equations (2), and an equation is added after each remaximization. However, the total number need never exceed $m + n + 2$. For if an s -variable, added earlier in the computation reappears among the variables on the left hand side of the equations after some remaximization, the equation involving it can simply be dropped, as the only equations that must be satisfied by a solution are the original ones. This limits the total number of s -variables present at one time to $n + 1$ or less.

Of course even the process just described involves an element of choice, any of the rows i of (2) with $a'_{i,0}$ non-integer might be chosen to generate the new relation. Some choices are better than others. A good rule of thumb based on the idea of "cutting" as deeply as possible with the new relation, and borne out by limited computational experience, is to choose the row with the largest fractional part $f_{i,0}$ in the zero column.

This class of possible additional constraints is not limited to those produced by the method described here since it is easily seen that some simple operations on and between rows preserve the properties needed in the additional relations. These operations can be used to produce systematically a family of additional relations from which a particularly effective cut or cuts can be selected. A discussion of this class of possible additional constraints together with a rule of choice of row which can be shown to bring the process to an end in a finite number of steps—thus providing a finite algorithm—requires some space and will be given as part of a more complete treatment in another place.

A small example, illustrating the method, is on the preceding page.

REFERENCES

1. George B. Dantzig, *Discrete-variable extremum problems*, Operations Res. vol. 5 no. 2 (1957).
2. G. Dantzig, R. Fulkerson, and S. Johnson, *Solution of a large-scale traveling salesman problem*, J. Operations Res. Soc. Amer. vol. 2 no. 4 (1954).
3. Harry M. Markowitz and Alan S. Manne, *On the solution of discrete programming problems*, Econometrica vol. 25 no. 1 (1957).

This page intentionally left blank

COMBINATORIAL PROCESSES AND DYNAMIC PROGRAMMING

BY

RICHARD BELLMAN

1. Introduction. The purpose of this paper is to discuss the application of dynamic programming techniques to a class of problems which for want of a better term we call combinatorial. The essential difficulty of these problems, from the standpoint of the analyst, lies in their apparent lack of complexity. Usually, it is either a question of performing a finite set of arithmetic operations or determining the largest of a finite set of numbers.

If there are one hundred elements in the finite set, we can classify the problem as trivial. If, however, the finite set possesses a million members, or a hundred million, it is worthwhile to ask whether or not there are more efficient techniques than just an element-by-element examination.

Problems of this nature arise in the following ways:

1. Solving linear systems of equations of the form

$$\sum_{j=1}^N a_{ij}x_j = b_i, \quad i = 1, 2, \dots, N.$$

2. Maximization of a linear form $L(x) = \sum_{i=1}^N c_i x_i$ subject to constraints of the form

$$\sum_{j=1}^N a_{ij}x_j \leq b_i, \quad i = 1, 2, \dots, M.$$

3. Maximization of functions over finite sets, such as permutations, paths along a grid, and so on.

At the present time, there is no systematic theory of problems of this genre, nor is it likely that there ever will be, considering the many varieties and sources. There are, however, some categories of problems recognized as tractable. Some are soluble explicitly in traditional analytic terms, some by means of algorithms that can be carried out by hand, and some requiring the most powerful computers available.

In what follows, we shall discuss various ways in which the functional equation technique of dynamic programming can be applied. We shall use only those portions of the general theory required for our present purposes, referring the reader interested in further aspects to [1].

Although we shall not present any specific numerical results here, we shall furnish references to extensive computational studies carried out by S. Dreyfus and the author.

2. An allocation problem. Let us begin with the following simple allocation problem. Suppose that we have a quantity x of a resource which we are going to subdivide into N parts, x_1, x_2, \dots, x_N , corresponding to N different activities. To make a mathematical problem of this, let us suppose that we are given functions $g_i(x_i)$ which measure the return from the i th activity due to an allocation x_i .

The question of most efficient allocation of resources leads to the analytic problem of maximizing the function

$$(1) \quad F_N(x_1, x_2, \dots, x_N) = g_1(x_1) + g_2(x_2) + \dots + g_N(x_N)$$

subject to the constraints

$$(2) \quad \begin{aligned} (a) \quad & x_1 + x_2 + \dots + x_N = x, \\ (b) \quad & x_i \geq 0. \end{aligned}$$

Although this may seem like a most prosaic problem, and hardly worth any attention at this late date in the history of calculus, as we shall see it has its hidden pitfalls.

The run-of-the-mill approach to this problem converts it by way of a Lagrange multiplier into that of maximizing the new function

$$(3) \quad G_N(x_1, x_2, \dots, x_N) = \sum_{i=1}^N g_i(x_i) - \lambda \sum_{i=1}^N x_i,$$

where λ is a parameter that will be determined from (2a).

The variational equations are

$$(4) \quad \frac{\partial G_N}{\partial x_i} = 0 = g'_i(x_i) - \lambda, \quad i = 1, 2, \dots, N.$$

Solving these N equations for the x_i in terms of λ , $x_i = x_i(\lambda)$, the parameter λ is determined by means of the relation

$$(5) \quad \sum_{i=1}^N x_i(\lambda) = x.$$

Although this approach is infallible in textbook problems, a number of difficulties arise in applications. Let us enumerate them.

In the first place, the functions $g_i(x)$ may not have a derivative. Although, as we shall see below, we do possess an efficient technique for solving a maximization problem of this type, we shall not insist upon this point. It is, however, reasonable to expect that the individual functions need not possess derivatives at various points.

Let us ignore these bizarre possibilities and assume that it is sufficient to examine the solutions of (4) and (5). If each of the functions $g'_i(x_i)$ is monotone, which is to say that $g_i(x)$ is either convex or concave, then the inverse function $x_i(\lambda)$ is uniquely defined and it becomes relatively easy to study the solutions of (5).

Since it is quite common for utility functions $g_i(x)$ to have points of inflection, if we wish to resolve general problems of this nature we must consider situations in which the equations in (4) have a multiplicity of solutions. Assuming, for the sake of moderate complication, that each equation of the form $g'_i(x_i) = \lambda$ possesses two solutions for any particular value of λ , we see that the solution of any equation such as (5) leads to a consideration of 2^N cases.

This number 2^N arises by counting all possible cases. The problem thus appears to have unpleasant combinatorial overtones.

We can amplify these overtones in a number of ways. In the first place, we can insist that the endpoints of the x_i -intervals be tested. The value $x_i = 0$ has a very important interpretation. It means that the i th activity is not engaged in at all. The problem of taking into account all possibilities of end-point extrema greatly complicates the enumeration of cases.

Secondly, we can impose additional constraints of the type

$$(6) \quad x_i x_{i+1} = 0.$$

The meaning of a constraint of this type is that the use of one activity effectually prevents the use of another.

So far, we have been complaining about the limitations of an approach based on calculus. Let us further curtail this technique by allowing the x_i to range only over the elements of a discrete set. Thus, we may impose the restriction

$$(7) \quad x_i = 0, 1, 2, \dots, \quad i = 1, 2, \dots, N.$$

At this point the analyst is tempted to feel that the cards have been too thoroughly stacked against him. Let the computing machines take over; let them solve the problem by the trivial method of examining all possibilities.

The people in charge of the computers, however, may become a bit aggrieved at this attitude. They will concede that they possess fantastic machines operating at phenomenal speeds that can resolve in a matter of hours problems that would have consumed lifetimes even twenty years ago. But these problems must be carefully chosen. Even rudimentary problems of other types cannot be solved by enumeration of cases.

Since this statement may come as a shock to anyone who has not taken the trouble to compute the total number of possible cases arising from simple combinatorial problems, let us illustrate this point by means of a question involving permutations.

Take the problem of placing N objects in N pigeonholes, assuming that we are given a function which measures the value of each assignment of objects. To resolve the problem by examining all cases, we must evaluate $N!$ different cases, corresponding to all different permutations.

Accustomed as we are to the familiar function $N!$, we seldom realize its

rapidity of growth. For $N = 10$, we have 3,628,000 possibilities, a formidable but not incredible number. For $N = 20$, it will amuse the reader to calculate how long it would take a computing machine which could evaluate one permutation a microsecond to examine all cases.¹

It follows that the mathematician cannot abdicate. He is obligated to develop algorithms which can handle these strange, new problems. Our aim will be to present some simple algorithms which are particularly suited to digital computers. This is not to be considered our ultimate objective, but merely a preliminary step on the way to understanding.

3. Functional equations. Let us now present an approach to these problems quite different from that of the calculus. Introduce the function $f_N(x)$, defined for $x \geq 0$ and $N = 1, 2, \dots$, by the relation

$$(1) \quad f_N(x) = \underset{R_N}{\text{Max}} [g_1(x_1) + g_2(x_2) + \dots + g_N(x_N)],$$

where R_N is the region in (x_1, x_2, \dots, x_N) -space defined by the relations

$$(2) \quad \begin{array}{ll} (a) & x_1 + x_2 + \dots + x_N = x, \\ (b) & x_i \geq 0. \end{array}$$

The only assumption we need make concerning the $g_i(x)$ is that they are continuous for $x_i \geq 0$. In the cases we shall treat below where the x_i assume only a finite set of values, even this restriction will not be necessary.

Let us write²

$$(3) \quad f_N(x) = \underset{0 \leq x_N \leq x}{\text{Max}} \underset{R_{N-1}(x_N)}{\text{Max}} [g_1(x_1) + g_2(x_2) + \dots + g_N(x_N)],$$

where $R_{N-1}(x_N)$ is defined by the inequalities

$$(4) \quad \begin{array}{ll} (a) & x_1 + x_2 + \dots + x_{N-1} = x - x_N, \\ (b) & x_i \geq 0. \end{array}$$

Hence, for $N = 2, \dots$,

$$(5) \quad f_N(x) = \underset{0 \leq x_N \leq x}{\text{Max}} [g_N(x_N) + \underset{R_{N-1}(x_N)}{\text{Max}} [g_1(x_1) + \dots + g_{N-1}(x_{N-1})]].$$

Thus, referring to the original definition of the sequence $\{f_k(x)\}$,

$$(6) \quad f_N(x) = \underset{0 \leq x_N \leq x}{\text{Max}} [g_N(x_N) + f_{N-1}(x - x_N)], \quad N = 2, 3, \dots$$

For $N = 1$, we have

$$(7) \quad f_1(x) = g_1(x).$$

¹ A simple lower bound is half a million years!

² This is a particular application of the "principle of optimality", see [1, p. 83].

4. Discussion. The preceding formalism reduces the original multi-dimensional maximization problem to a sequence of one-dimensional problems. The practical significance of this fact is that we now do possess a feasible technique for solving these problems by direct search methods using digital computers.

In this way, we can treat a number of problems arising in mathematical economics, engineering, and operations research. The computational solutions of these questions, joint work with S. Dreyfus, will appear in book form in the near future.

If we attribute some structure to the functions, such as linearity, quadratic character, convexity or concavity, the recurrence relations in (3.5) can be used to determine the analytic character of the sequence $\{f_i(x)\}$ and the maximizing x_i as functions of x ; cf. [1; 2], Karush, [40].

Alternatively, a structural feature such as concavity can be used to accelerate greatly the machine search for a maximum, cf. Kiefer, [41], Johnson, [36], Johnson and Gross, [33], Kiefer, [42]. Applications of the technique are contained in [18; 19].

5. An imbedding process. It is important to point out what we have accomplished by means of the functional equation technique. We have taken a particular problem with a specific value of x and N and made it a member of a family of problems, continuous in x and discrete in N .

In other words, we have imbedded a particular process within a family of processes. Oddly, it is easier to treat the particular process by consideration of the whole family of processes, than it is to treat the process by itself.

Further discussion of this point will be found in [1; 23].

6. Constraints. Let us now examine the effect of constraints upon the method outlined in §4. Suppose that we impose the additional constraints

$$(1) \quad 0 < a_i \leq x_i \leq b_i, \quad i = 1, 2, \dots, N,$$

in addition to those of (2.2).

It is easy to see that the relations of (4.5) are replaced by

$$(2) \quad f_N(x) = \underset{S_N}{\text{Max}} [g_N(x_N) + f_{N-1}(x - x_N)],$$

where S_N is the x_N -region determined by the new conditions

$$(3) \quad \begin{aligned} (a) \quad & a_N \leq x_N \leq b_N, \\ (b) \quad & x_N \leq x - (a_1 + a_2 + \dots + a_{N-1}). \end{aligned}$$

In the definition of $f_N(x)$, x is restricted by the lower bound $a_1 + a_2 + \dots + a_N$.

The interesting thing to note is that whereas in the usual approach to maximization problems additional constraints cause difficulties, here the more constraints, the simpler the computational task. Additional constraints restrict the region over which each variable can roam, and thus simplify the search for a maximum. We shall mention this again below.

7. Constraints-discreteness. As an example of a nastier type of constraint, consider the problem of maximizing

$$(1) \quad F(x_1, x_2, \dots, x_N) = \sum_{i=1}^N g_i(x_i)$$

subject to the constraints

$$(2) \quad \begin{aligned} (a) \quad & x_i = 0 \text{ or } 1, \\ (b) \quad & \sum_{i=1}^N x_i \leq x. \end{aligned}$$

As before, we obtain the recurrence relation

$$(3) \quad \begin{aligned} f_N(x) &= \underset{x_N=0,1}{\operatorname{Max}} [g_N(x_N) + f_{N-1}(x - x_N)] \\ &= \operatorname{Max} [g_N(1) + f_{N-1}(x - 1), g_N(0) + f_{N-1}(x)]. \end{aligned}$$

The computation can now be carried out by hand in a very simple fashion. Observe that this is the simplest type of maximization problem that a machine can perform.

8. Mutually exclusive activities. Let us now complicate matters still further. In addition to the restrictions in the preceding section, let us impose the constraint

$$(1) \quad x_i x_{i+1} = 0, \quad i = 1, 2, \dots, N - 1.$$

To treat this problem, introduce the sequence of functions of two variables, $f_N(x, y)$, defined by the relations

$$(2) \quad f_N(x, y) = \underset{R_N}{\operatorname{Max}} [g_1(x_1) + g_2(x_2) + \dots + g_N(x_N)],$$

where R_N is now the region in (x_1, x_2, \dots, x_N) -space defined by

$$(3) \quad \begin{aligned} (a) \quad & x_1 + x_2 + \dots + x_N \leq x, \\ (b) \quad & x_i = 0, 1, \\ (c) \quad & x_i x_{i+1} = 0, \quad i = 1, 2, \dots, N - 1, \\ (d) \quad & x_N y = 0. \end{aligned}$$

The quantity y is allowed to take only two values, 0 or 1.

Then we have the recurrence relation

$$(4) \quad f_N(x, y) = \underset{x_N}{\operatorname{Max}} [g_N(x_N) + f_{N-1}(x - x_N, x_N)],$$

where x_N is subject to the conditions

$$(5) \quad \begin{aligned} (a) \quad & x_N = 0, 1, \\ (b) \quad & x_N \leq x, \\ (c) \quad & x_N y = 0. \end{aligned}$$

In order to resolve the original problem we must compute the two sequences $f_N(x,0)$, $f_N(x,1)$.

It is easily seen that

$$(6) \quad \begin{aligned} f_N(x,0) &= \text{Max } [g_N(1) + f_{N-1}(x - 1,1), g_N(0) + f_{N-1}(x,0)], \\ f_N(x,1) &= g_N(0) + f_{N-1}(x,0). \end{aligned}$$

The two sequences $\{f_N(x,0)\}$, $\{f_N(x,1)\}$, can thus be determined quite simply.

9. More constraints. Returning to the simpler problem discussed in §2, let us consider the problem of maximizing

$$(1) \quad F(x_1, x_2, \dots, x_N) = g_1(x_1) + g_2(x_2) + \dots + g_N(x_N),$$

subject to the constraints

$$\begin{aligned} (a) \quad & x_1 + x_2 + \dots + x_N \leq x, \\ (2) \quad (b) \quad & a_1 x_1 + a_2 x_2 + \dots + a_N x_N \leq y, \quad a_i \geq 0, \\ (c) \quad & x_i \leq 0. \end{aligned}$$

Observe that we have replaced equality signs by inequalities, since this avoids some unimportant consistency requirements.

Introducing the sequence of functions, $f_N(x,y)$, defined by

$$(3) \quad f_N(x,y) = \underset{R_N}{\text{Max}} F(x_1, x_2, \dots, x_N)$$

for $N = 1, 2, \dots, x, y \geq 0$, it is easy to see that, as in the preceding sections, we obtain the recurrence relation

$$(4) \quad f_N(x,y) = \underset{s_N}{\text{Max}} [g_N(x_N) + f_{N-1}(x - x_N, y - y_N)],$$

$N \geq 2$, with

$$(5) \quad f_1(x,y) = \text{Max } [g_1(x_1)],$$

for $0 \leq x_1 \leq \text{Min } [x, y/a_1]$.

10. General formulation. There is no difficulty in formulating the problem of maximizing

$$(1) \quad F^N(x) = \sum_{i=1}^N g(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)})$$

subject to the constraints

$$\begin{aligned} (2) \quad (a) \quad & \sum_{i=1}^N k_{ij}(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)}) \leq y_j, \quad j = 1, 2, \dots, M, \\ (b) \quad & x_i^{(j)} \geq 0, \end{aligned}$$

in the same fashion. Setting

$$(3) \quad f_N(y_1, y_2, \dots, y_M) = \underset{x}{\operatorname{Max}} F_N(x),$$

we see that

$$(4) \quad f_N(y_1, y_2, \dots, y_M) = \underset{S_N}{\operatorname{Max}} [g(x_N^{(1)}, x_N^{(2)}, \dots, x_N^{(k)}) \\ + f_{N-1}(y_1 - k_{N1}(x_N^{(1)}, \dots, x_N^{(k)}), \dots)].$$

Prior to any discussion of the computational feasibility of an algorithm of this type for general values of M , let us turn to a particular problem of this type in which large values of M enter in a most natural way.

11. The Hitchcock-Koopmans transportation problem. Let us now discuss one of the most interesting models in mathematical economics, the Hitchcock-Koopmans model of the flow of commodities.

Suppose that at N different locations, which we shall call *sources*, there are quantities of an item which must be transported to M other locations which we shall call *sinks*.

Let x_{ij} denote the quantity of the item at the i th source, y_j denote the demand for this item at the j th sink, and a_{ij} denote the cost of transporting a unit quantity from the i th source to the j th sink. Furthermore, assume that the total supply at the sinks is equal to the total demand from the sources.

The problem is to determine a shipping policy which minimizes the cost of supplying the demand. To reduce this problem to analytic form, let

$$(1) \quad x_{ij} = \text{the quantity sent from the } i\text{th source to the } j\text{th sink.}$$

Then we are required to minimize the linear function

$$(2) \quad L(x) = \sum_{i=1}^N \sum_{j=1}^M a_{ij} x_{ij}$$

over all x_{ij} satisfying the linear constraints

$$(a) \quad \sum_{j=1}^N x_{ij} = x_i,$$

$$(3) \quad (b) \quad \sum_{i=1}^M x_{ij} = y_j,$$

$$(c) \quad x_{ij} \geq 0.$$

This problem is one that can be treated very successfully by the "simplex technique" of G. Dantzig, [27], or by the newer methods of Fulkerson and Ford, [30]; cf. also Prager, [46]. Both of these methods depend strongly upon the linearity of the various equations.

It can easily be shown that the linearity of all the functions involved prevents the existence of any internal maximum. The region defined by the relations of (3) is the interior of a multi-dimensional polyhedron. To determine the maximum of $L(x)$, it is sufficient to examine the values of $L(x)$ at the vertices of this region.

It follows that we have a problem of combinatorial type. The methods described above furnish efficient search techniques.

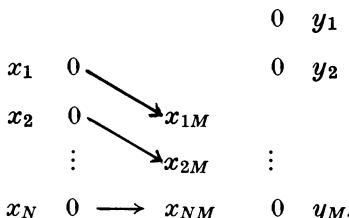
These methods fail in general if we introduce nonlinear cost functions. We shall employ functional equation techniques.

12. The nonlinear transportation problem. Let us examine the problem of minimizing

$$(1) \quad g(x) = \sum_{i=1}^M \sum_{j=1}^N g_{ij}(x_{ij})$$

over all x_{ij} satisfying the constraints of (10.3), where the $g_{ij}(x)$ are not necessarily linear.

To treat the question by means of functional equations, we can proceed in one of two ways. To begin with, assume that we satisfy the total demand in the following fashion: first, the demand of the M th sink, then, having satisfied this, the demand of the $(M - 1)$ st sink, and so on.



For fixed demands, y_i , let

- (2) $f_M(x_1, x_2, \dots, x_N) =$ the minimum cost to satisfy the demands of M remaining sinks, starting with quantities x_1, x_2, \dots, x_N at the N sinks.

Then the same reasoning as we have used above yields the equation

$$(3) \quad f_M(x_1, x_2, \dots, x_N) = \min_{x_{iM}} \left[\sum_{i=1}^N g_{iM}(x_{iM}) + f_{M-1}(x_1 - x_{1M}, x_2 - x_{2M}, \dots, x_N - x_{NM}) \right]$$

where the x_{iM} vary over the region determined by

$$(4) \quad \begin{aligned} (a) \quad & \sum_{i=1}^N x_{iM} = y_M, \\ (b) \quad & 0 \leq x_{iM} \leq x_i, \quad i = 1, 2, \dots, N. \end{aligned}$$

The function $f_1(x_1, x_2, \dots, x_N)$ is given by

$$(5) \quad f_1(x_1, x_2, \dots, x_N) = \sum_{i=1}^N g_{i1}(x_{i1}).$$

We have thus transformed the original problem into that of computing the sequence $\{f_M(x_1, x_2, \dots, x_N)\}$.

It is clear that we could obtain an alternative formulation by using first all the resources of the N th source to satisfy some of the demands at the M sinks, and so on.

13. Feasibility. Let us now see whether or not the recurrence relations presented in (12.3) actually lead to a feasible computational scheme. At each stage of the computation we have to tabulate a function of N variables, and perform a minimization over an N -dimensional region.

Although both of these are formidable procedures if M and N are large, the tabulation problem is at the moment the most difficult. Suppose that we allow each x_i to assume one hundred values, say $x_i = 0, \Delta, \dots, 99\Delta$. Then the total number of grid-points required to tabulate $f_M(x_1, x_2, \dots, x_N)$ will be 10^{2N} . If $N = 1$, this is 100, a trivial number; if $N = 2$, this is 10,000, a respectable number; and if $N = 3$, this is 1,000,000, an impossible number at the present time.

It follows that the foregoing method in its straightforward form cannot be used to handle problems of this nature unless N or $M \leq 2$.

Two facts save this from being an academic exercise. In the first place, there are a number of important situations in which N or M is one, two or three. In the second place, as we shall see below, there are a number of devices we can combine with the functional equation technique in order to treat higher dimensional problems.

These are

- (a) Lagrange multipliers,
- (1) (b) Functional approximation,
- (c) Successive approximations.

We shall discuss these ideas in turn.

14. Reduction by one variable. In view of the tremendous difference between the memory requirements for functions of two variables and functions of three variables, it is of interest to point out that transportation processes involving N sources can be treated by means of functions of $N - 1$ variables. Hence, problems involving two sources are easily resolved, while problems involving three sources can be treated with the best of current machines.

To obtain this reduction in dimensionality, we observe that as yet we have made no use of the fact that supply is equal to demand,

$$(1) \quad \sum_{i=1}^N x_i = \sum_{j=1}^M y_j.$$

From this it follows that the values of x_1, x_2, \dots, x_{N-1} determine the value of x_N , once we have specified the y_i . Hence

$$(2) \quad f_M(x_1, x_2, \dots, x_N) \equiv f_M(x_1, x_2, \dots, x_{N-1}).$$

In much of analysis, dimensionality plays an inessential role. In computational work, it is a basic consideration.

15. Lagrange multipliers and dynamic programming.³ We have another very powerful way of evading the curse of dimensionality. Returning to the allocation problem discussed initially, consider the problem of maximizing

$$(1) \quad F(x_1, x_2, \dots, x_N) = \sum_{i=1}^N g_i(x_i)$$

subject to the constraints

$$\begin{aligned} (a) \quad & \sum_{i=1}^N x_i \leq x, \\ (2) \quad (b) \quad & \sum_{i=1}^N a_i x_i = y, & a_i \geq 0, \\ (c) \quad & x_i \geq 0. \end{aligned}$$

Observe that we have kept one constraint an equality, one an inequality. Again there are some reasons of convenience.

As we know, this problem can be treated by means of functions of two variables. However, there is a great incentive for reducing the problem to one that can be handled by functions of one variable.

What we do is combine the functional equation technique with the classical Lagrange multiplier formalism. Consider the problem of maximizing the new function

$$(3) \quad G(x_1, x_2, \dots, x_N) = \sum_{i=1}^N g_i(x_i) - \lambda \sum_{i=1}^N a_i x_i$$

subject to the constraints

$$\begin{aligned} (a) \quad & \sum_{i=1}^N x_i \leq x, \\ (4) \quad (b) \quad & x_i \geq 0, \end{aligned}$$

where λ is an as yet undetermined parameter.

For fixed λ , introduce the sequence of functions

$$(5) \quad f_N(x) = \underset{R_N}{\text{Max}} \left[\sum_{i=1}^N g_i(x_i) - \lambda \sum_{i=1}^N a_i x_i \right],$$

³ First presented in [3].

where R_N is defined only by (4). Then, as before, we readily compute the sequence $\{f_k(x)\}$ by means of the relations

$$(6) \quad f_N(x) = \underset{0 \leq x_N \leq x}{\text{Max}} [g_N(x_N) - \lambda x_N + f_{N-1}(x - x_N)].$$

Let $x_1(\lambda), x_2(\lambda), \dots, x_N(\lambda)$ be a set of values yielding the maximum of $G(x_1, x_2, \dots, x_N)$. Then we assert that these values yield the solution to the problem of maximizing (1) subject to the constraint in (2) where y is determined by

$$(7) \quad y = \sum_{i=1}^N a_i x_i(\lambda).$$

To prove this, proceed by contradiction. Suppose that there existed values (z_1, z_2, \dots, z_N) satisfying (2) such that

$$(8) \quad F(z_1, z_2, \dots, z_N) > F(x_1(\lambda), x_2(\lambda), \dots, x_N(\lambda)).$$

Then

$$(9) \quad \begin{aligned} F(z_1, z_2, \dots, z_N) - \lambda \sum_{i=1}^N a_i z_i &= F(z_1, z_2, \dots, z_N) - \lambda y \\ &> F(x_1(\lambda), x_2(\lambda), \dots, x_N(\lambda)) - \lambda y \\ &= F(x_1(\lambda), x_2(\lambda), \dots, x_N(\lambda)) - \lambda \sum_{i=1}^N a_i x_i(\lambda). \end{aligned}$$

This, however, yields a contradiction, since the $x_i(\lambda)$ were obtained as a solution to (5), subject to the constraints of (4).

Although there is no difficulty in letting the results justify the method in any particular application, there are a number of important facts which remain to be verified. We suspect that as λ varies from $-\infty$ to $+\infty$ that the value of $\sum_{i=1}^N a_i x_i(\lambda)$ will vary between its maximum and minimum, and, furthermore that this variation will be monotone and continuous.

The monotonicity is not only of theoretical importance, but of practical significance in determining the value of λ for which $\sum_{i=1}^N a_i x_i(\lambda) = y$; cf. Gross and Johnson, [33]. Some applications of this technique will be found in [17; 18].

16. Discussion. The importance of the procedure outlined above resides in the fact that it enables us to partition a problem originally requiring a sequence of functions of two variables into a sequence of problems requiring functions of one variable.

There is no difficulty in extending these techniques to treat the case where there are M constraints. What we gain in reducing dimensionality on one hand, we must pay for in multi-dimensional search on the other.

As we know, the introduction of Lagrange multipliers is equivalent to introducing dual variables; cf. Kuhn and Tucker, [43]. What we have done above is to operate partially in the original space and partially in the dual space; partly in the space of "resources" and partly in the space of "prices".

17. Functional approximation. In the previous sections, when we have discussed the computational solution of functional equations, we have tacitly equated the concept of a function $f(x)$ defined over an interval $[0, a]$ with a set of values $\{f(k\Delta)\}$, where $N\Delta = a$, and Δ is some grid size. The finer the grid, the more values that must be computed. Similarly, a function of two variables, $f(x, y)$, is equivalent to a sequence of values $\{f(k\Delta, l\Delta)\}$.

As we increase the number of independent variables, the number of grid-points goes up at an exponential rate. It is this fact that defeats the effective use of the algorithms presented above in a number of significant processes.

It follows that one way to defeat this exponential growth in the information required to specify a function is to use a different description.

Consider, for example, a power series expansion

$$(1) \quad f(x) = \sum_{n=0}^{\infty} a_n x^n,$$

convergent for $0 \leq x \leq a$. If we truncate the series and use the polynomial $\sum_{n=0}^N a_n x^n$ as an approximation to the function, we see that $f(x)$ is determined for all x in $[0, a]$ by the $N + 1$ coefficients, a_i , $i = 0, 1, \dots, N$, and thus by $(N + 1)$ quantities.

Power series expansions have the drawback of being associated with analyticity and, in addition, of not providing uniformly good approximation over the entire interval. Let us then use instead an orthonormal expansion

$$(2) \quad f(x) \sim \sum_{n=0}^{\infty} a_n \phi_n(x)$$

where the functions $\phi_i(x)$ are elements of a complete orthonormal system. For a finite interval, two particularly important choices are those of trigonometric functions, $\{\sin kx, \cos kx\}$, and of Legendre polynomials.

We write

$$(3) \quad f(x) = \sum_{n=0}^N a_n \phi_n(x),$$

where

$$(4) \quad a_n = \int_0^a f(x) \phi_n(x) dx.$$

In evaluating this integral, we don't wish to use a Riemann sum, say

$$(5) \quad a_n \simeq \sum_{k=0}^M f(k\Delta) \phi_n(k\Delta) \Delta,$$

since this will involve the calculation of $f(k\Delta)$ for all k , precisely the type of computation we wished to avoid.

Consequently, we employ a numerical integration formula of the form

$$(6) \quad \int_0^a g(x)dx \cong \sum_{i=1}^M c_i g(x_i),$$

where x_i are fixed points in $[0, a]$, independent of $g(x)$ but dependent on M , and the c_i are likewise fixed coefficients independent of $g(x)$, but dependent on M .

Thus

$$(7) \quad a_n \cong \sum_{i=1}^M c_i f(x_i) \phi_n(x_i) \cong \sum_{i=1}^M d_i n f(x_i),$$

since the quantities $\phi_n(x_i)$ can be calculated once and for all.

Observe the interesting fact about this formula that the value of a_n , and thus of $f(x)$ is made to depend upon the values of $f(x_i)$ at a fixed set of points $\{x_i\}$.

Let us see then how the calculation proceeds. Turning to the recurrence relation

$$(8) \quad f_N(x) = \underset{0 \leq x_N \leq x}{\text{Max}} [g_N(x_N) + f_{N-1}(x - x_N)],$$

suppose that x is restricted to an interval $[0, a]$. Starting with the function $f_1(x) = g_1(x)$, we reduce $f_1(x)$ to a sequence of coefficients

$$(9) \quad f_1(x) \sim [a_0^{(1)}, a_1^{(1)}, \dots, a_N^{(1)}].$$

We see from the above discussion that to determine $f_2(x)$, we need only calculate the values $f_2(x_i)$. Hence, we compute these from the relations

$$(10) \quad f_2(x_i) = \underset{0 \leq x \leq x_i}{\text{Max}} [g_N(x_2) + f_1(x_i - x_2)].$$

The values of $f_1(x_i - x_2)$ are determined from (7).

Having computed $f_2(x_i)$, $i = 1, 2, \dots, M$, in this fashion, we determine the new coefficients $a_i^{(2)}$ using (7). The function $f_2(x)$ is thus reduced to a sequence of coefficients

$$(11) \quad f_2(x) \sim [a_0^{(2)}, a_1^{(2)}, \dots, a_N^{(2)}].$$

We now repeat this process to determine as many elements of the sequence $\{f_N(x)\}$ as desired.

A number of questions remain before this technique can be applied. We must determine N in (3) and M in (7), and the type of orthonormal sequence. The choice of N and M depend upon the accuracy desired and the facilities available.

Both the trigonometric functions, $\{\sin Nx, \cos Nx\}$, and the Legendre polynomials possess simple recurrence relations which permit the N th member of the sequence to be computed from the values of the first members.

As far as quadratic formulas are concerned, it is probably best to use Gaussian quadrature, which, as we know, is exact for polynomials up to degree $2M - 1$ if M points are used in (6). For an application of this technique, see [16].

18. Cebycev approximation.

The approximation

$$(1) \quad f(x) \cong \sum_{n=0}^N a_n \phi_n(x)$$

is equivalent to choosing the coefficients $\{a_i\}$ according to mean-square approximation. If the b_i are determined so as to minimize the mean-square deviation

$$(2) \quad \int_0^a \left[f(x) - \sum_{n=0}^N b_n \phi_n(x) \right]^2 dx,$$

we find that $b_i = a_i$.

However, mean-square deviation is less desirable than Cebycev approximation,

$$(3) \quad \text{Min}_{b_i} \text{Max}_{0 \leq x \leq a} \left| f(x) - \sum_{n=0}^N b_n \phi_n(x) \right|.$$

Unfortunately, no simple representation for the minimizing b_i , corresponding to (17.4) exists. Nevertheless, there are available feasible computational techniques for determining the minimizing b_i in (3).

19. Functions of several variables. In the previous sections, we have shown how a function defined over $[0, a]$ may be described by a relatively small set of parameters. The same process can be applied to a function of two variables, $f(x, y)$, defined over $0 \leq x, y \leq a$,

$$(1) \quad f(x, y) \cong \sum_{m,n=0}^N a_{mn} \phi_m(x) \phi_n(y).$$

We see that functions of two variables will require $(N + 1)(N + 2)/2$ coefficients, while those of three variables will require roughly $N^3/6$ coefficients.

Taking $N = 10$, we have functions of two variables determined by 66 quantities, and functions of three variables determined by approximately 200 quantities. These numbers compare very favorably with 10^4 and 10^6 arising from $10^2 \times 10^2$ grids and $10^2 \times 10^2 \times 10^2$ grids.

In any particular problem, a certain amount of experimentation will be required.

Again an important point to stress is that these techniques allow us to treat problems which cannot be treated by straightforward tabulation of functional values at grid-points.

20. Successive approximations. Let us now discuss an approach of entirely different type to the problem of solving multi-dimensional problems in terms of functions of a small number of variables. We wish to employ the classical method of *successive approximations*.

To illustrate the workings of the method, let us give two examples of its use, one in connection with the allocation problem described above, and one in connection with the Hitchcock-Koopmans problem.

Consider the problem of maximizing

$$(1) \quad h(x_1, \dots, x_N; y_1, \dots, y_N) = \sum_{k=1}^N g_k(x_k, y_k)$$

subject to the constraints

$$(a) \quad \sum_{k=1}^N x_k = x, \quad x_k \geq 0,$$

$$(2) \quad (b) \quad \sum_{k=1}^N y_k = y, \quad y_k \geq 0.$$

As we know, this problem can be treated by means of sequences of functions of two variables, §9, and by means of sequences of functions of one variable using Lagrange multiplier techniques. Let us now treat it by means of successive approximations.

Let $(y_1^{(0)}, y_2^{(0)}, \dots, y_N^{(0)})$ be an initial guess concerning the choice of the y_i , and consider the problem of maximizing the function

$$(3) \quad h(x_1, \dots, x_N; y_1^{(0)}, \dots, y_N^{(0)}) = \sum_{k=1}^N g_k(x_k, y_k^{(0)})$$

subject to the constraint of (2a). This problem can, as we know, be resolved via functions of one variable.

Call a maximizing set of x_k $\{x_k^{(0)}\}$. Now consider the problem of maximizing

$$(4) \quad h(x_1^{(0)}, \dots, x_N^{(0)}; y_1, \dots, y_N) = \sum_{k=1}^N g_k(x_k^{(0)}, y_k),$$

subject to the constraints of (2b). This again is a one-dimensional problem in our terms. Call a maximizing set $\{y_k^{(1)}\}$. The pattern of procedure is now set. We obtain alternately maximizing sequences $\{y_k^{(i)}\}$ and $\{x_k^{(i)}\}$, with corresponding approximations to the desired maximum value,

$$(5) \quad h(x_1^{(i)}, \dots, x_N^{(i)}; y_1^{(i)}, \dots, y_N^{(i)}), \quad h(x_1^{(i)}, \dots, x_N^{(i)}; y_1^{(i+1)}, \dots, y_N^{(i+1)}).$$

We have thus once again reduced a problem originally requiring functions of two variables to one requiring sequences of functions of one variable.

21. Monotonicity of approximation. Let

$$(1) \quad \begin{aligned} u_{2i+1}(x, y) &= \underset{x}{\text{Max}} \ h(x, y^{(i)}), & i &= 0, 1, \dots, \\ u_{2i}(x, y) &= \underset{y}{\text{Max}} \ h(x^{(i)}, y), & i &= 1, 2, \dots \end{aligned}$$

It is clear that

$$(2) \quad u_1(x,y) \leq u_2(x,y) \leq \dots$$

Hence, the sequence $\{u_k(x,y)\}$ converges monotonically. It is, however, not clear that it converges to the absolute maximum. This requires a separate discussion which we shall present elsewhere.

22. Approximation in policy space. This monotonicity of approximation is not accidental. The type of approximation we have been employing is a particular type of approximation in “policy space”, which necessarily yields monotonicity.

In place of approximating to the return functions, the $f_k(x,y)$ as defined in §9, the usual method of successive approximation, we operate partially in the space of policy functions.

For a further discussion of approximation in policy space, see [1, Chapter 3], and for some further applications of successive approximations, see [4; 24].

23. Application to Hitchcock-Koopmans transportation problem. One way of applying these ideas to either the linear or nonlinear transportation is the following. Let the shipments from the 3rd to N th source be assigned arbitrarily, and consider the problem of determining the shipments from the first two sources which will minimize the cost of supplying the remaining demand.

This, as we know, can be done using functions of one variable. Having obtained the optimal shipments from the first and second sources, we use this shipping policy from the first source, retain the shipments previously used from the fourth to N th source, and consider the new problem of determining the shipments from the second and third sources which will minimize the cost of supplying the remaining demand.

Continuing in this way, it is clear that we obtain a sequence of costs which decrease monotonically and thus converge. Again it is necessary to determine when there is convergence to the actual minimum.

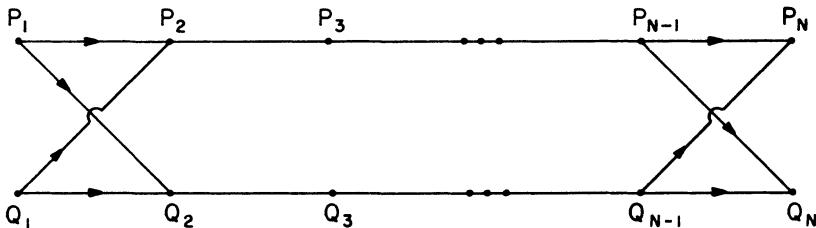
The process can be speeded up, by considering three sources at a time and one Lagrange multiplier to retain the use of functions of one variable.

24. The Harris transportation problem. Consider the network, shown at the top of the following page, which can be used to describe certain types of flow of information or flow of commodities.

As the arrows indicate, the only permissible flows are from P_i to P_{i+1} , P_i to Q_{i+1} , Q_i to P_{i+1} , and Q_i to Q_{i+1} .

We introduce quantities, which we can call *capacities*, defined as follows:

- (1) a_i = maximum rate of flow over the link between P_i and P_{i+1} ,
- b_i = the same quantity for P_i and Q_{i+1} ,
- c_i = the same quantity for Q_i and P_{i+1} ,
- d_i = the same quantity for Q_i and Q_{i+1} .



Assuming that we are given fixed rates of flow, x and y , into P_1 and Q_1 , we wish to maximize a prescribed function of the rate of flow into P_N and Q_N . At each of the terminals, we have a choice of dividing the input flows into two output flows.

To treat this problem, we introduce the sequence of functions,

- (2) $f_i(x,y) = \text{the maximum of the prescribed function of the flow into } P_N \text{ and } Q_N, \text{ given rates of flow } x \text{ and } y \text{ into } P_i \text{ and } Q_i \text{ respectively,}$

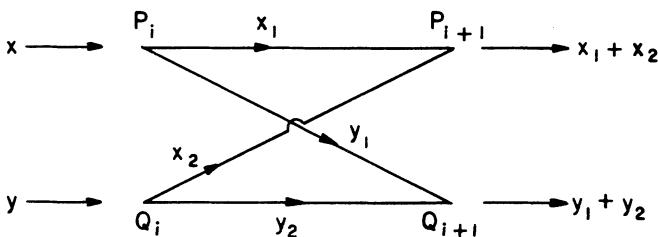
for $i = 1, 2, \dots, N - 1$, $x, y \geq 0$.

Then, as above, we obtain the recurrence relations

$$(3) \quad f_i(x,y) = \underset{R_i}{\text{Max}} f_{i+1}(x_1 + x_2, y_1 + y_2),$$

where $R_i = R_i(x_1, x_2, y_1, y_2)$ is determined by the constraints

- (4) (a) $x_1 + y_1 \leq x, \quad x_2 + y_2 \leq y,$
 (b) $0 \leq x_1 \leq a_i, \quad 0 \leq x_2 \leq b_i,$
 (c) $0 \leq y_1 \leq c_i, \quad 0 \leq y_2 \leq d_i.$



25. **General network problems.** Networks of general type, where the points are irregularly arranged with great freedom of intercommunication cannot be treated in the foregoing fashion. The problem of determining optimal flow over systems of this type was first posed by Harris, [34], and has given rise to a good deal of research.

The general problem has been most successfully attacked by Ford and

Fulkerson, [30], who have developed techniques for the treatment of this specific process which have proved to have much wider applicability.

On the other hand, stimulated by the same process, Boldyreff, [24], has developed a very interesting and flexible relaxation technique, the "flooding technique", which also has a wide range of applications.

26. A routing problem. Let us now consider the following problem. Suppose that we have a set of N points in the plane or in space, with the property that every two points, P_i and P_j , have an associated number, d_{ij} , which we can call the time required to travel from P_i to P_j . It is not necessary to assume that $d_{ij} = d_{ji}$.

To take account of the fact that in any particular situation two points may not be mutually accessible, we can let $d_{ij} = \infty$.

Given the matrix (d_{ij}) , where $d_{ii} = 0$, the problem is to trace a route of minimum time from P_1 to P_N .

To treat this, introduce the $N - 1$ quantities, f_i , defined by

$$(1) \quad f_i = \text{the minimum time to travel from } P_i \text{ to } P_N.$$

It is easy to see that

$$(2) \quad f_i = \underset{j \neq i}{\text{Min}} [d_{ij} + f_j].$$

Since this system of nonlinear equations does not determine the sequence $\{f_i\}$ recurrently, we must use some method of successive approximations to obtain the f_i .

Perhaps the simplest is one based upon approximation in policy space. Let

$$(3) \quad f_i^{(0)} = d_{iN}, \quad i = 1, 2, \dots, N - 1,$$

and

$$(4) \quad f_i^{(k)} = \underset{j \neq i}{\text{Min}} [d_{ij} + f_j^{(k-1)}],$$

for $k = 1, 2, \dots$.

Since this process corresponds to examining first all direct routes from i to N , then all that make one stop, and so on, it is clear that we will have monotone decreasing convergence. Further discussion concerning the equations will be found in [5].

27. The traveling salesman problem. As an example of a problem of closely related type where the direct functional equation technique fails, consider the problem of drawing a polygonal path of minimum length through N given points in the plane, P_1, P_2, \dots, P_N .

It is clear that the principle of optimality is still valid. No matter what part of the path has already been traversed, the remainder of the path must be of minimum length. What prevents a simple application of recurrence

relations is the fact that we must keep track of where we have been. The problem thus has certain features in common with a variety of "excluded volume" problems arising in mathematical physics.

Furthermore, it is a very nice example of the virtual impossibility of gauging the level of difficulty of a simply stated combinatorial problem. Recently, linear programming techniques plus ingenuity have proved successful in solving particular questions of this nature, see Dantzig and Johnson, [28].

To treat this problem by means of functional equations and successive approximations, let us begin with the question of tracing a path through N given points which must start at P_1 and end at P_N .

Introduce to this end the sequence of functions

$$(1) \quad f(Q_1, Q_2, \dots, Q_k) = \text{the minimum length of the remaining path from } Q_k \text{ to } P_N, \text{ having been through } Q_1, Q_2, \dots, Q_k.$$

Here the Q_i are particular elements of the P_j -set.

It is easy to see that we obtain the following relations :

$$(2) \quad \begin{aligned} f(P_1) &= \text{Min}_{j \neq i} [d_{ij} + f(P_1, P_j)], \\ f(P_1, P_j) &= \text{Min}_{k \neq j} [d_{jk} + f(P_1, P_j, P_k)], \end{aligned}$$

and so on.

It remains to discuss the computational feasibility of a solution based upon this algorithm. The tabulation of $f(P_1, P_j)$ requires $(N - 1)$ values ; that of $f(P_1, P_j, P_k)$ requires $(N - 1)(N - 2)/2$. What is of great help is the fact that the order of Q_i in (1) is of no importance, so that the tabulation of $f(Q_1, Q_2, \dots, Q_k)$ requires $(N - 1)(N - 2) \cdots (N - k)/k!$ entries, rather than $(N - 1)(N - 2) \cdots (N - k)$.

The maximum number of entries will be required when $k = [N/2]$. To illustrate the order of magnitude of these quantities, we have

$$(3) \quad 10C_5 = 252, \quad 16C_8 = 12,970.$$

It follows that with current machines it would be possible to solve problems of this type in a direct fashion for $N \leq 17$.

Let us note in passing that the same technique can be applied to problems of optimal trajectory arising in rocket problems, and in the study of general variational processes ; cf. Cartaino and Dreyfus, [25], and [6 ; 1].

28. Successive approximations. To treat problems of larger magnitude, we can combine functional equation techniques with the method of successive approximations.

Let $P_1Q_2Q_3 \cdots P_N$ be a proposed route. To test this, let us examine the sub-route $P_1Q_2Q_3 \cdots Q_9$, and pose the problem of pursuing a path from P_1 to Q_9 , going through the points Q_2, Q_3, \dots, Q_8 , and of minimal length.

This problem can easily be resolved computationally, using the recurrence relation method of the preceding section. Let $P_1R_2R_3\cdots Q_9$ be the new path of minimum length through these ten points, and let us test in this way the set of points $R_5R_6R_7R_8Q_9P_{10}P_{11}P_{12}P_{13}P_{14}$. Continuing in this way we obtain a monotone sequence of decreasing lengths. Again it will be necessary to study whether or not this scheme of successive approximations converges to the absolute minimum.

29. A class of scheduling problems—the book-binding problem. Let us now turn to another class of problems which require maximization over permutation.

Suppose that we have N manuscripts which must be printed and bound, in that order, before being published. Suppose further that we have one printing press and one binding machine. Given the quantities

- (1) (a) a_i = the time required to print the i th book,
- (b) b_i = the time required to bind the i th book,

the problem is to determine the order in which the N manuscripts should be processed so as to minimize the total time required for their printing and binding.

This problem is representative of a large class of questions of this nature which arise in scheduling theory. A very simple solution of this problem was given by Johnson, [37], and a derivation of this solution by functional equation techniques was given in [7].

If we add a third operation, say typing, the problem appears to enter the hopeless case. At the present time, we do not even possess any reasonable approximate policies.

30. The Caterer problem. It is possible, in a number of cases, to reduce scheduling problems to maximization problems of the type encountered in the Hitchcock-Koopmans transportation problem. Having done this, we can introduce functional equations by various artifices. One example of this is in connection with “warehousing” problems, cf. [9], Dreyfus, [29], and also [13]. Here, we shall discuss another example, the “caterer” problem.

Let us state the version of this problem given by Jacobs, [35], or Prager, [46].

“A caterer knows that in connection with the meals he has arranged to serve during the next n days, he will need r_j fresh napkins on the j th day, for $j = 1, 2, \dots, n$. There are two types of laundry service available. One type requires p days and costs b cents per napkin; a faster service requires q days, $q < p$, but costs c cents per napkin, $c > b$. Beginning with no usable napkins on hand or in the laundry, the caterer meets the demands by purchasing napkins at a cents per napkin. How does the caterer purchase and launder napkins so as to minimize the total cost for n days?”

This problem can be resolved by linear programming techniques in some cases, see the above references and also Gaddum, Hoffman, and Sokolowsky, [32].

What is interesting about this problem from the standpoint of dynamic programming and functional equations is that it appears upon first glance to be a problem requiring functions of q variables. As it turns out, however, the linearity of the process permits us to make a certain type of preliminary transformation which reveals the true dimensionality of the problem. Surprisingly, this is $p - q$. The same type of transformation has proved of great service in connection with a number of variational problems arising in the theory of control processes and elsewhere, see [4; 10].

In the present case, these transformations reduce the problem to that of determining the maximum of the linear form

$$(1) \quad L(v) = v_1 + v_2 + \cdots + v_n,$$

subject to the constraints

$$(2) \quad \begin{aligned} (a) \quad & r_i \geq v_i \geq 0, \\ (b) \quad & v_1 \leq b_1, \\ & v_1 + v_2 \leq b_2, \\ & \vdots \\ & v_1 + v_2 + \cdots + v_k \leq b_k, \\ & v_2 + v_3 + \cdots + v_{k+1} \leq b_{k+1}, \\ & \vdots \\ & v_{n-k+1} + \cdots + v_n \leq b_n. \end{aligned}$$

A further surprising fact about this problem is that we can exhibit an explicit analytic solution, a property that is quite rare in this domain.

31. Bottleneck problems. Let us, without going into any detail, mention a class of problems which we have called "bottleneck" problems because the operation of the system depends at each stage upon the scarcest resources.

The general question is that of utilizing a complex of interdependent industries to produce one or two essential items. Using a "lumped" model of economic interaction we consider the state of the system at any time to be specified by a capacity vector $x(t)$ and a stockpile vector $y(t)$.

Considering first a continuous process, since these are usually more amenable to solution, we meet the problem of determining vectors $z(t)$ and $w(t)$ so as to maximize the inner product

$$(1) \quad I(x,y) = (x(T),a) + (y(T),b),$$

where x and y are determined by linear equations

$$(2) \quad \begin{aligned} \frac{dx}{dt} &= A_1x + B_1y + C_1z + D_1w, \quad x(0) = c, \\ \frac{dy}{dt} &= A_2x + B_2y + C_2z + D_2w, \quad y(0) = d, \end{aligned}$$

under appropriate proportionality assumptions, and the vectors z and w are subject to further feasibility constraints

$$(3) \quad Ez + Fw \leq G + Hx + Iy$$

for $0 \leq t \leq T$.

The novel features of the problem are introduced by the combination of linear equations and linear constraints. The continuous version can be resolved explicitly in a number of cases, see [1; 11]. In addition, Lehman has devised a continuous version of the simplex technique which seems quite promising, [44].

The discrete version can be simply treated by computational techniques if the number of state variables does not exceed 3. A transformation of the problem enables us to reduce the number of variables by one, and simultaneously to keep all variables within a uniformly bounded region; cf. [1; 15].

If the constraints in (3) are of the simpler form

$$(4) \quad Ez + Fw \leq G,$$

the techniques sketched in [10] can be used to reduce drastically the dimensionality of the problem.

A detailed discussion of problems of this type, together with solutions of some special cases is given in [1]. Some further results are found in Lehman, [45].

32. Slightly intertwined matrices. The functional equation technique is designed to take advantage of structural features of the process. In the preceding discussion we have utilized the multistage aspects in a very natural way. Let us now indicate another way in which functional equations can be employed. These results were presented in [12].

The general linear programming problem is that of maximizing a linear form $L(x) = (x, a)$ subject to linear constraints of the form $Ax \leq b$.

If A is block-diagonal,

$$(1) \quad A = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_r \end{bmatrix}$$

the problem breaks up into a set of r independent problems of smaller dimension. Once again we are concerned with dimensionality in connection with computational feasibility.

Similarly, if A is block-triangular, which is to say that there are zeros above or below the diagonal matrices, the problem reduces to a sequence of problems of lesser difficulty.

Here we wish to show how functional equation techniques can be utilized

to treat a class of problems in which A has approximately a block-diagonal form.

Specifically, let us consider the question of determining the maximum of

$$(2) \quad L_N(x) = \sum_{i=1}^{3N} x_i,$$

subject to the constraints

$$(3) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &\leq c_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &\leq c_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + b_1x_4 &\leq c_3, \\ a_{44}x_4 + a_{45}x_5 + a_{46}x_6 &\leq c_4, \\ a_{54}x_4 + a_{55}x_5 + a_{56}x_6 &\leq c_5, \\ a_{64}x_4 + a_{65}x_5 + a_{66}x_6 + b_2x_7 &\leq c_6, \\ &\vdots \\ a_{3N-2,3N-2}x_{3N-2} + a_{3N-2,3N-1}x_{3N-1} + a_{3N-2,3N}x_{3N} &\leq c_{3N-2}, \\ a_{3N-1,3N-2}x_{3N-2} + a_{3N-1,3N-1}x_{3N-1} + a_{3N-1,3N}x_{3N} &\leq c_{3N-1}, \\ a_{3N,3N-2}x_{3N-2} + a_{3N,3N-1}x_{3N-1} + a_{3N,3N}x_{3N} &\leq c_{3N}, \end{aligned}$$

and

$$(4) \quad x_i \geq 0.$$

It is assumed that $a_{ij} \geq 0$, $b_i > 0$, with sufficiently many $a_{ij} > 0$ so that the maximum is not infinite.

Let us now define a sequence of functions of z ,

$$(5) \quad f_N(z) = \underset{x_i}{\text{Max}} \ L_N(x),$$

where the x_i are subject to the constraints given above, with the exception that the last constraint is now

$$(6) \quad a_{3N,3N-2}x_{3N-2} + a_{3N,3N-1}x_{3N-1} + a_{3N,3N}x_{3N} \leq z.$$

Employing the principle of optimality, we see that the sequence $\{f_N(z)\}$ satisfies the recurrence relation

$$(7) \quad f_N(z) = \underset{\left[x_{3N-2}, x_{3N-1}, x_{3N}\right]}{\text{Max}} [x_{3N-2} + x_{3N-1} + x_{3N} + f_{N-1}(c_{3N-3} - b_{N-1}x_{3N-2})],$$

$$N \geq 1,$$

with the variables x_{3N-2} , x_{3N-1} , x_{3N} subject to the constraints

$$(8) \quad \begin{aligned} a_{3N-2,3N-2}x_{3N-2} + a_{3N-2,3N-1}x_{3N-1} + a_{3N-2,3N}x_{3N} &\leq c_{3N-2}, \\ a_{3N-1,3N-2}x_{3N-2} + a_{3N-1,3N-1}x_{3N-1} + a_{3N-1,3N}x_{3N} &\leq c_{3N-1}, \\ a_{3N,3N-2}x_{3N-2} + a_{3N,3N-1}x_{3N-1} + a_{3N,3N}x_{3N} &\leq z, \\ b_{N-1}x_{3N-2} &\leq c_{3N-3}, \\ x_{3N-2}, x_{3N-1}, x_{3N} &\geq 0. \end{aligned}$$

the function $f_0(z)$ is identically zero.

33. Reduction in dimensionality. Let us write the recurrence relation of (32.7) in the form

$$(1) \quad f_N(z) = \operatorname{Max}_{x_{3N-2}} \left[\operatorname{Max}_{x_{3N-1}, x_{3N}} [\dots] \right]$$

$$= \operatorname{Max}_{x_{3N-2}} \left[\operatorname{Max}_{R_N} (x_{3N-1} + x_{3N}) + x_{3N-2} + f_{N-1}(c_{3N-3} - b_{N-1}x_{3N-2}) \right]$$

where R_N is the region in (x_{3N-1}, x_{3N}) space defined by

$$(2) \quad \begin{aligned} a_{3N-2, 3N-1}x_{3N-1} + a_{3N-2, 3N}x_{3N} &\leq c_{3N-2} - a_{3N-2, 3N-2}x_{3N-2}, \\ a_{3N-1, 3N-1}x_{3N-1} + a_{3N-1, 3N}x_{3N} &\leq c_{3N-1} - a_{3N-1, 3N-2}x_{3N-2}, \\ a_{3N, 3N-1}x_{3N-1} + a_{3N, 3N}x_{3N} &\leq z - a_{3N, 3N-2}x_{3N-2}, \\ x_{3N-1}, x_{3N} &\geq 0. \end{aligned}$$

Thus we can write

$$(3) \quad f_N(z) = \operatorname{Max}_{x_{3N-2}} [g_N(x_{3N-2}, z) + f_{N-1}(c_{3N-3} - b_{N-1}x_{3N-2})]$$

where x_{3N-2} is constrained by

$$(4) \quad 0 \leq x_{3N-2} \leq \operatorname{Min} \left[\frac{c_{3N-3}}{b_{N-1}}, \frac{c_{3N-2}}{a_{3N-2, 3N-2}}, \frac{c_{3N-1}}{a_{3N-1, 3N-2}}, \frac{z}{a_{3N, 3N-2}} \right].$$

The function $g_N(y, z)$ is readily determined, since the maximum over R_N is attained at a vertex of the region.

34. Slightly intertwined symmetric matrices. Let us now consider the problem of resolving a set of linear equations of the forms

$$(1) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= c_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= c_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + b_{14}x_4 &= c_3, \\ b_{13}x_3 + a_{44}x_4 + a_{45}x_5 + a_{46}x_6 &= c_4, \\ a_{54}x_4 + a_{55}x_5 + a_{56}x_6 &= c_5, \\ a_{64}x_4 + a_{65}x_5 + a_{66}x_6 + b_{27}x_7 &= c_6, \\ &\vdots \\ b_{N-1}x_{3N-3} + a_{3N-2, 3N-2}x_{3N-2} + a_{3N-2, 3N-1}x_{3N-1} + a_{3N-2, 3N}x_{3N} &= c_{3N-2}, \\ a_{3N-1, 3N-2}x_{3N-2} + a_{3N-1, 3N-1}x_{3N-1} + a_{3N-1, 3N}x_{3N} &= c_{3N-1}, \\ a_{3N, 3N-2}x_{3N-2} + a_{3N, 3N-1}x_{3N-1} + a_{3N, 3N}x_{3N} &= c_{3N}. \end{aligned}$$

A matrix of the type appearing above, we shall call *slightly intertwined*. It arises in a variety of physical, engineering, and economic problems involving multicomponent systems with weak coupling.

Let us introduce the matrices

$$(2) \quad A_k = (a_{i+3k-3, j+3k-3}), \quad i, j = 1, 2, 3,$$

for $k = 1, 2, \dots, N$, and the vectors

$$(3) \quad x^k = (x_{3k-2}, x_{3k-1}, x_{3k}), \quad c^k = (c_{3k-2}, c_{3k-1}, c_{3k}).$$

Since the matrix of coefficients is, by assumption, positive definite, the solution of the linear system (1) is equivalent to determining the minimum of the inhomogeneous quadratic form

$$(4) \quad (x^1, A_1 x^1) + (x^2, A_2 x^2) + \cdots + (x^N, A_N x^N) \\ - 2(c^1, x^1) - 2(c^2, x^2) - \cdots - 2(c^N, x^N) \\ + 2b_1 x_3 x_4 + 2b_2 x_6 x_7 + \cdots + 2b_{N-1} x_{3N-3} x_{3N-2}.$$

For $N = 1, 2, \dots$, and $-\infty < z < \infty$, let us introduce the sequence of functions of the variable z defined by

$$(5) \quad f_N(z) = \min_{x_i} \left[\sum_{i=1}^N (x^i, A_i x^i) - 2 \sum_{i=1}^N (c^i, x^i) + 2 \sum_{i=1}^{N-1} b_i x_{i+3} x_{3i} + 2zx_{3N} \right].$$

We then have the following recurrence relation:

$$(6) \quad f_N(z) = \min_{(x_{3N}, x_{3N-1}, x_{3N-2})} [(x^N, A_N x^N) + 2zx_{3N} \\ - 2(c^N, x^N) + f_{N-1}(b_{N-1} x_{3N-2})].$$

35. Computational aspects—I. Since the function $f_1(z)$ is readily determined, we can compute the sequence $\{f_k(z)\}$ at the expense of a minimization over a 3-dimensional region. This minimization may be greatly speeded up upon using the convexity properties of the functions involved. Although no optimal methods are known for multidimensional problems, the one-dimensional method presented in [36] may be employed in an iterative manner.

Writing (34.6) in the form

$$(1) \quad f_N(z) = \min_{x_{3N-2}} \left[\min_{x_{3N}, x_{3N-1}} [(x^N, A_N x^N) + 2zx_{3N} - 2(c^N, x^N)] \\ + f_{N-1}(b_{N-1} x_{3N-2}) \right],$$

we see that it reduces to

$$(2) \quad f_N(z) = \min_y [g_N(z, y) + f_{N-1}(b_{N-1} y)],$$

where

$$(3) \quad g_N(z, y) = \min_{x_{3N}, x_{3N-1}} [(x^N, A_N x^N) + 2zx_{3N} - 2(c^N, x^N)],$$

upon identifying x_{3N-2} as y . This new relation is now well-suited to the technique described in [36].

The computation of the functions $\{g_N(z, y)\}$ is independent of the computation of the sequence $\{f_N(z)\}$. Observe that this computational approach involves no divisions.

36. Computational aspects—II. Another approach to the computational solution reposes upon the easily established fact that $f_N(z)$ is a quadratic in z for each N , i.e.

$$(1) \quad f_N(z) = u_N + 2v_N z + w_N z^2,$$

where u_N , v_N and w_N are independent of z . This is the same device used in connection with Jacobi matrices, see [13].

Substituting in (34.6), we obtain the equation

$$(2) \quad u_N + v_N z + w_N z^2 = \underset{(x_{3N}, x_{3N-1}, x_{3N-2})}{\text{Min}} \quad [(x^N, A_N x^N) + 2zx_{3N} - 2(c^N, x^N) \\ + u_{N-1} + 2b_{N-1} x_{3N-2} v_{N-1} + b_{N-1}^2 x_{3N-2}^2 w_{N-1}].$$

Upon performing the minimization and determining the minimum value of the right-hand side, we obtain recurrence relations connecting the triple (u_N, v_N, w_N) with the triple $(u_{N-1}, v_{N-1}, w_{N-1})$.

This affords an alternative computational technique.

The problem of determining the largest and smallest characteristic values may be approached in a similar fashion, cf. [13; 1].

37. Reliability of multi-component systems. A fundamental problem in the design of electronic systems, switching networks, computing devices and automata, is that of constructing more reliable devices from less reliable components. Essentially, it becomes a question of minimal duplication. Some general discussions of these problems may be found in [17].

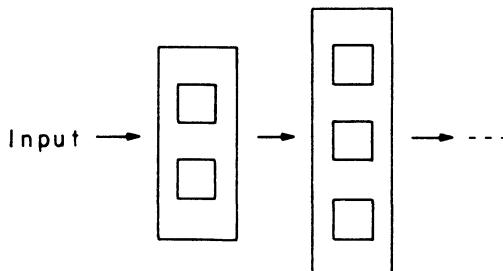
As a sample of the type of problem that can be treated using functional equation techniques, let us consider the following.

Let us suppose that the device we wish to design will consist of a number of stages each of which feeds its successor. Thus



The reliability of the device will be interpreted to be the probability that it operates successfully, and this in turn will be taken to be the product of the reliabilities of the individual stages.

In many cases, the overall reliability is too low for its intended use. One way to increase the reliability is to introduce a number of duplicate components in parallel at various stages. Thus



We assume that we possess switching techniques which will automatically introduce a new component into the circuit at any stage if the first component used is faulty. The reliability of the entire system can now be improved by duplication of components in this fashion.

The process cannot be carried to any logical extreme because of physical constraints of size and cost. Consequently, we have the problem of determining optimal duplication subject to given weight and size constraints. In addition to permitting first choice of the number of components at each stage, we shall subsequently also allow a choice of the type of component. This latter feature introduces combinatorial aspects, although, of course, these are already present in the condition of discreteness.

Assuming that at least one component must be used at each stage, let

- (1) $p_j(x_j)$ = the probability of successful performance of the j th stage if $1 + x_j$ components are used at the j th stage.

Let the cost of a component at the j th stage be c_j and the weight be w_j . Due to the increase in complexity of switching circuits as x_j is increased, there is no reason to assume proportional cost. However, we shall do so here to simplify the notation, since the method we present is equally applicable to the general problem.

The variational problem is then that of determining the maximum of

$$(2) \quad \prod_{j=1}^N p_j(x_j),$$

subject to the constraints

$$(a) \quad \sum_{j=1}^N c_j x_j \leq c,$$

$$(3) \quad (b) \quad \sum_{j=1}^N w_j x_j \leq w,$$

$$(c) \quad x_j = 0, 1, \dots$$

It is clear that this problem may be transformed into that of computing a sequence of functions of two variables, as outlined in §9, and by use of the Lagrange multiplier technique reduced to a problem involving a sequence of functions of one variable. The details, and the results of some computations, may be found in [17].

38. Different types of components. Let us now suppose that at each stage we have a choice of two types of components, those of an *A*-type and those of a *B*-type. Let $c_j(A)$, $w_j(A)$, be respectively the costs and weights for components of the *A*-type at the j th stage, and $c_j(B)$, $w_j(B)$ denote the corresponding quantities for *B*-types.

Given the overall restrictions on weight and cost, we wish, as above, to determine the types of components, and the quantities, which maximize the total reliability. We shall suppose that the switching requirements are

such that at any particular stage it is impossible to combine any number of A -components with any number of B -components.

Defining the sequence of functions, $f_i(c,w)$, in the usual fashion, with the functions $p_j(x_j, A)$, $p_j(x_j, B)$ corresponding to $p_j(x_j)$, we obtain the relations

$$(1) \quad f_1(c,w) = \operatorname{Max} \begin{bmatrix} \operatorname{Max}_{x_1} p_1(x_1, A) \\ \operatorname{Max}_{x_1} p_1(x_1, B) \end{bmatrix},$$

where $1 \leq x_1 \leq \operatorname{Min}[[c/c_1(A)], [w/w_1(A)]]$ in the expression containing A , and x_1 satisfies the analogous constraint in the expression containing B .

Generally,

$$(2) \quad f_N(c,w) = \operatorname{Max} \begin{bmatrix} \left[\operatorname{Max}_{x_N} p_N(x_N, A) f_{N-1}(c - c_N(A)x_N, w - w_N(A)x_N) \right] \\ \left[\operatorname{Max}_{x_N} p_N(x_N, B) f_{N-1}(c - c_N(B)x_N, w - w_N(B)x_N) \right] \end{bmatrix},$$

where x_N satisfies corresponding constraints,

$$(3) \quad \begin{aligned} 1 \leq x_N &\leq \operatorname{Min}[[c/c_N(A)], [w/w_N(A)]], \\ 1 \leq x_N &\leq \operatorname{Min}[[c/c_N(B)], [w/w_N(B)]], \end{aligned}$$

in the two maximizations.

39. Sequential search. In the remainder of the paper we wish to discuss a number of interesting problems in which we encounter the general question of finding in minimum time an element of a finite set possessing certain distinguishing characteristics.

Considering the great importance of the problem and the fascinating nature of the questions that arise, it is amazing how little work has been done in the field.

40. Determining the maximum value of a function. Let us begin with a simple deterministic problem. Given a continuous function $f(x)$ defined over the interval $[0,1]$, we wish to determine the value of x which maximizes $f(x)$.

For a variety of reasons, some of which we have discussed above, we do not wish to use calculus, but wish rather to employ a search method. To make the problem of determining the value of a maximizing x in an efficient fashion precise, let us pose the following problem.

"Given the continuous function $f(x)$ defined over $[0,1]$, determine the quantity $N(d)$ and the associated search policy so that one can guarantee that a maximum value can be located within an interval of length d in at most $N(d)$ steps."

If the function is taken merely to be continuous, with no additional properties, it is clear that $N(d) = 1/d$. If, however, we add that $f(x)$ is concave, then this number can be considerably reduced, and the problem possesses a very elegant solution.

This solution was found first by Kiefer, [41] and then, independently, and in a simpler fashion by Johnson, [36], using functional equations.

A similar problem can be posed with reference to locating the unique zero of a continuous, monotone, concave function. This has been resolved, using functional equations, by Gross and Johnson, [33].

A detailed discussion of this type of problem will be found in Kiefer, [42].

41. Sequential testing. The problem we have discussed in the preceding section is a particular case of the general problem of sequential testing. Let us discuss two particular problems which will illustrate the difficulties in this domain.

Suppose that we have a piece of equipment which has N different parts to be examined if there is loss of function. Given a priori probability distributions associated with the individual parts, and a set of testing devices which furnish various indications, how should we proceed so as to locate all sources of malfunction in a minimum time?

One version of this problem arises in connection with designing a machine for the diagnosis and treatment of medical ailments. There is no reason why this vital problem should be left to intuition of individuals. Other aspects arise in connection with writing manuals for mechanics in automobile and airplane establishments.

Particular parts of the general problem have been treated by means of the theory of sequential analysis; see Wald, [49].

The other problem we wish to discuss is well-known in its simplest version.

Given a number of coins which are indistinguishable in appearance and the information that exactly one is heavier or lighter than all the others which are of the same weight, we wish to use an equal arm balance so as to determine in the smallest number of weightings the distinctive coin.

This problem has been discussed by a number of authors, using a number of different techniques. It possesses a very simple and elegant solution.

The following extension of the problem seems to be of quite a different level of complexity. In place of assuming that there is one distinctive coin, suppose that we are given the information that there are k distinctive coins. At the moment, we may assume that these are all of the same weight.

Although some work has been done on this problem; cf. Cairns, [26], there exists no solution at the present time.

42. Discussion. If we attempt to apply functional equation techniques to these problems, we encounter very much the same difficulty that we met in the traveling salesman problem. As we proceed in our testing, the information pattern becomes tremendously complicated, and it appears to be impossible to describe the state of the system in any simple way.

43. Design of experiments. The problems presented in the foregoing sections are in turn special cases of what may be called the general problem of the design of experiments.

We have considered initially the relatively simple case in which the structure of the system is assumed known. If the structure is taken to be partially unknown, we first encounter situations in which we hypothesize a stochastic structure, and then the very much more difficult situations in which we have to determine the structure on the basis of observation as we proceed.

The information we possess determines the decisions we make, and the decisions we make determine the new information pattern. The problem of determining optimal policies in situations of this type is very much more difficult than any of the problems we have previously discussed.

Not only is the analysis much more intricate because of the stochastic structure of the process, but it is no longer easy to make precise what we mean by an optimal policy.

For a detailed discussion of matters of this nature, we refer to the papers by Robbins, [47], Bellman and Kalaba, [22], Bellman, [14], and Karlin, [39], Karlin and Johnson, [38].

BIBLIOGRAPHY

1. R. Bellman, *Dynamic programming*, Princeton, Princeton University Press, 1957.
2. ———, *A functional equation arising in allocation theory*, J. Soc. Indust. Appl. Math. vol. 3 (1955) pp. 129–132.
3. ———, *Dynamic programming and lagrange multipliers*, Proc. Nat. Acad. Sci. U.S.A. vol. 42 (1956) pp. 767–769.
4. ———, *Some new techniques in the dynamic programming solution of variational problems*, Quart. Appl. Math. vol. 16 (1958) pp. 295–305.
5. ———, *On a routing problem*, Quart. Appl. Math. vol. 16 (1958) pp. 87–90.
6. ———, *On the application of the theory of dynamic programming to the study of control processes*, Proc. Symposium on Control Processes, Polytechnic Institute of Brooklyn, April, 1956, pp. 199–213.
7. ———, *Mathematical aspects of scheduling theory*, J. Soc. Indust. Appl. Math. vol. 4 (1956) pp. 168–205.
8. ———, *On the theory of dynamic programming—a warehousing problem*, Management Sci. vol. 2 (1956) pp. 272–276.
9. ———, *Dynamic programming and the smoothing problem*, Management Sci. vol. 3 (1956) pp. 111–113.
10. ———, *Terminal control, time lags, and dynamic programming*, Proc. Nat. Acad. Sci. U.S.A. vol. 43 (1957) pp. 927–930.
11. ———, *Bottleneck problems, functional equations and dynamic programming*, Econometrica vol. 22 (1954).
12. ———, *On the computational solution of linear programming problems involving almost block diagonal matrices*, Management Sci. vol. 3 (1957) pp. 403–406.
13. ———, *On some applications of dynamic programming to matrix theory*, Illinois J. Math. vol. 1 (1957) pp. 297–301.
14. ———, *A problem in the sequential design of experiments*, Sankhya vol. 16 (1956) pp. 221–229.
15. R. Bellman and S. Dreyfus, *On the computational solution of dynamic programming processes—VIII: A bottleneck situation involving interdependent industries*, The RAND Corporation, Paper P-1282, April 17, 1957.

16. R. Bellman and S. Dreyfus, *Approximations and dynamic programming*, The RAND Corporation, Paper P-1176, September 13, 1957.
17. ———, *Dynamic programming and the reliability of multi-component devices*, J. Operations Res. Soc. Amer. vol. 6 (1958) pp. 200–206.
18. ———, *On the computational solution of dynamic programming processes—X: The flyaway kit problem*, The RAND Corporation, Research Memorandum RM-1889, April 5, 1957.
19. ———, *On the computational solution of dynamic programming processes—V: A smoothing problem*, The RAND Corporation, Research Memorandum RM-1749, April 2, 1957.
20. ———, *On the computational solution of dynamic programming processes—II: On a cargo loading problem*, The RAND Corporation, Research Memorandum RM-1746, November 5, 1956.
21. R. Bellman and R. Kalaba, *On the role of dynamic programming in statistical communication theory*, IRE Transactions of the Professional Group on Information Theory vol. IT-3 (1957) pp. 197–203.
22. ———, *Communication processes involving learning and random duration*, IRE National Convention Record, Section on Information Theory, Part 4 (1958) pp. 16–20.
23. ———, *On the principle of invariant imbedding and propagation through inhomogeneous media*, Proc. Nat. Acad. Sci. U.S.A. vol. 42 (1956) pp. 629–632.
24. A. Boldyreff, *An iterative technique for determining rail capacity*, J. Operations Res. Soc. Amer. (1957).
25. H. Cartaino and S. Dreyfus, *Application of dynamic programming to the minimum time-to-climb problem*, Aeronautical Engineering Review (1957).
26. S. Cairns, *Balance scale sorting*, The RAND Corporation, Paper P-736, September 7, 1955.
27. G. Dantzig, *Application of the simplex technique to a transportation problem, activity analysis of production and allocation*, New York, Wiley and Sons, 1953.
28. G. Dantzig, D. R. Fulkerson and S. Johnson, *Solution of a large-scale traveling salesman problem*, J. Operations Res. Soc. Amer. (1954).
29. S. Dreyfus, *An analytic solution of the warehousing problem*, Management Sci. (1957).
30. L. R. Ford, Jr. and D. R. Fulkerson, *A simple algorithm for finding maximal network flows and an application to the Hitchcock problem*, Canad. J. Math. vol. 9 (1957) pp. 210–218.
31. D. R. Fulkerson, *A network flow feasibility theorem with applications to incidence matrices and the subgraph problem*, The RAND Corporation, Paper P-1278, February 12, 1958.
32. J. W. Gaddum, A. J. Hoffman and D. Sokolowsky, *On the solution to the caterer problem*, Naval Res. Logist. Quart. vol. 1 (1954) pp. 154–165.
33. O. Gross and S. Johnson, *Sequential minimax search for a zero of a convex function*, The RAND Corporation, Paper P-935, September 11, 1956.
34. T. E. Harris and F. S. Ross, *Fundamentals of a method for evaluating rail net capacities*, Unpublished.
35. W. Jacobs, *The caterer problem*, Naval Res. Logist. Quart. vol. 1 (1954) pp. 154–165.
36. S. Johnson, *Best exploration for maximum is fibonaccian*, The RAND Corporation, Paper P-856, May 4, 1956.
37. ———, *Optimal two- and three-stage production schedules with set-up times included*, Naval Res. Logist. Quart. (1954).
38. S. Johnson and S. Karlin, *A Bayes model in sequential design*, Ann. Math. Statist. (1956).

39. S. Karlin and R. Bradt, *On the design and comparison of certain dichotomous experiments*, Ann. Math. Statist. vol. 27 (1956) pp. 390-409.
40. W. Karush, *On a class of minimum-cost problems*, Management Sci. vol. 4 (1958) pp. 136-155.
41. J. Kiefer, *Sequential minimax search for a maximum*, Proc. Amer. Math. Soc. vol. 4 (1953) pp. 502-506.
42. ———, *Optimum sequential search and approximation methods under minimum regularity assumptions*, J. Soc. Indust. Appl. Math. vol. 5 (1957) pp. 105-136.
43. H. Kuhn and A. W. Tucker, *Nonlinear programming*, Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1951.
44. R. S. Lehman, *On the continuous simplex method*, The RAND Corporation, Research Memorandum RM-1386, November 24, 1954.
45. ———, *Studies in bottleneck problems in production processes*, Part II, The RAND Corporation, Paper P-492, 1954.
46. W. Prager, *On the caterer problem*, LBM-13, Division of Applied Mathematics, Brown University, 1956.
47. H. Robbins, *Some aspects of the sequential design of experiments*, Bull. Amer. Math. Soc. vol. 58 (1952) pp. 527-536.
48. J. von Neumann, *A certain zero-sum two-person game equivalent to the optimal assignment problem*, Contributions to the Theory of Games, Annals of Mathematics Studies, no. 28, Princeton University Press, 1953.
49. A. Wald, *Statistical decision functions*, New York, John Wiley and Sons, 1950.

THE RAND CORPORATION,
SANTA MONICA, CALIFORNIA

This page intentionally left blank

SOLUTION OF LARGE SCALE TRANSPORTATION PROBLEMS

BY

MURRAY GERSTENHABER

In the classical Hitchcock-Koopmans transportation problem (sometimes also associated with Kantorovitch) a homogeneous commodity is available at m points ("producers") in quantities a_1, \dots, a_m and is required at n points ("consumers") in quantities b_1, \dots, b_n with $\sum a_i = \sum b_j$; the cost c_{ij} of transporting a unit quantity from the i th producer to the j th consumer is given and one must find the amount x_{ij} that should be delivered to the j th consumer from the i th producer in order that the total cost shall be a minimum, i.e., one must find quantities $x_{ij} \geq 0$ such that (1) $\sum_j x_{ij} = a_i$, (2) $\sum_i x_{ij} = b_j$, and (3) $\sum_{ij} x_{ij} c_{ij} = \text{minimum}$. More algorithms have been published for this problem than for any other in linear programming, a circumstance probably resulting from the simplicity of its statement and its intuitive appeal, but none has been given which works best on all computers, let alone on problems of all possible sizes and special characteristics. (It is highly unlikely that such an algorithm exists.) The purpose of this article is to describe some of the considerations that entered into the programming of an algorithm due to the author [1] and suitable for very large problems. Some empirical results giving indications of its speed are also given. This paper contains little theory, and if it proves anything it is that computation is still more nearly an art than a science, particularly when very large problems are involved.

Description of the algorithm. The author's algorithm stems from the observation that if the sources of supply actually represented distinct and independent producers and the points of demand, similarly, independent consumers, then in practice each consumer would place an order for the amount he requires with that producer who could quote the lowest delivered price, i.e., concentrating attention on the j th consumer we would have $x_{ij} = 0$ unless $c_{ij} = \min_h c_{hj}$. For x_{ij} 's so determined, however, we would have, in general, $\sum_j x_{ij} \neq a_i$, so they would not define even a "feasible" solution to the problem, i.e., one satisfying conditions (1) and (2); if by extraordinary chance $\sum_j x_{ij} = a_i$, then the solution would be both feasible and optimal, i.e., would also satisfy (3). (Detailed proofs of these statements, if any are needed, can be found in [1].)

Now if the costs c_{ij} are replaced by $c'_{ij} = c_{ij} - \lambda_i$ then the problem is not changed in the sense that a solution is optimal for these costs if and only if it was so for the original ones. The behaviour of the independent consumers will change, however, and the orders they place will be functions $x_{ij}(\lambda)$ of

the vector of "subsidies" $\lambda = (\lambda_1, \dots, \lambda_m)$. Were it possible to find subsidies such that $\sum_j x_{ij}(\lambda) = a_i$ then the problem would again be at an end, but this can happen only by extraordinary chance even now since $x_{ij}(\lambda)$ is a discontinuous function of λ ; a consumer will place his entire order with the lowest-bidding producer, no matter how small the difference in price. The matter would be much improved if the consumer's behavior could be smoothed, and to this end a "consumer ordering policy" is introduced. This consists of a set of m continuous non-negative functions $p_1(t_1, \dots, t_m), \dots, p_m(t_1, \dots, t_m)$ of m real variables with the properties (i) there exists a "threshold" r such that $t_i - t_j \geq r$ implies $p_i(t_1, \dots, t_m) = 0$ and (ii) $\sum p_i = 1$ for all values of the variables. Here p_i is to be interpreted as the probability that a consumer will order from the i th producer if the prices quoted by the m producers for delivering a unit quantity of the commodity are t_1, \dots, t_m , respectively. Setting $x_{ij}(\lambda) = b_j p_i(c_{ij} - \lambda_1, \dots, c_{mj} - \lambda_m)$, the x_{ij} become continuous functions of λ and it is possible to choose the subsidies λ so that $\sum_j x_{ij}(\lambda) = a_i$ (for proof see [1]). Since $\sum_i x_{ij}(\lambda) = b_j$ for any λ , these x_{ij} define a feasible solution to the problem, although no longer necessarily an optimal one. However, the difference between the cost of this solution and that of an optimal one is not more than $r \sum a_i$ and for sufficiently small r may be counted as negligible. (Algorithms, including those for the general linear programming problem, which derive from propositions similar to these have come to be known as "threshold methods" after the threshold r ; cf. Kelley [3].) The principal problem now is to compute a subsidy vector λ for which $\sum_j x_{ij}(\lambda) = a_i$.

The choice of the term "subsidy" for the components of the vector λ is not intended to prejudice the economic interpretation attached to them. Koopmans has pointed out that we may in fact regard the λ 's as profits and deal with the present model as one of a free economy. We speak of subsidies because the person computing a solution has the bureaucratic power to assign them at will.

It seemed at first remarkable to the author that a solution to the equations $\sum_j x_{ij}(\lambda) = a_i$ should exist regardless of the choice of "consumer ordering policy" (probability functions) p_1, \dots, p_m satisfying the given conditions. Various new proofs having been given (cf., e.g., [5]), one may now take the matter as intuitively obvious. It is only for special classes of functions p_1, \dots, p_m , however, that an actual algorithm for the solutions of the equations $\sum_j x_{ij}(\lambda) = a_i$ has so far been given, specifically for functions with $p_i(t_1, \dots, t_m)$ monotone increasing in t_j for $j \neq i$ and monotone decreasing in t_i . (The probability of buying from the i th producer is a monotone increasing function of the prices quoted by the other producers and a monotone decreasing function of the price quoted by the i th producer.) For such functions, the following is a process converging to a solution vector λ . (For details of the proof see [2].)

For simplicity we shall henceforth set $\sum_j x_{ij}(\lambda) - a_i = d_i(\lambda)$; this may be

interpreted as the excess of orders received by the i th producer over the supply he has available when the subsidy vector is fixed at λ . Our task is then to solve the “order equilibrium problem” $d_i(\lambda) = 0$ for $i = 1, \dots, m$. Denoting the vector with 1 in the i th place and zeros elsewhere by e_i , we can define a sequence of vectors $\lambda^{(0)}, \lambda^{(1)}, \lambda^{(2)}, \dots$ in the following way. Set $\lambda^{(0)} = 0$, and having given $\lambda^{(k)} = \lambda$, define real non-negative quantities c_i so $c_i = 0$ if $d_i(\lambda) \geq 0$; otherwise choose c_i so that $d_i(\lambda + c_i e_i) = 0$. By the assumed monotonicity properties of the functions p_1, \dots, p_m , such a c_i always exists. Then set $\lambda^{(k+1)} = \lambda^{(k)} + \sum c_i e_i$. It is then the case that the sequence $\lambda^{(0)}, \lambda^{(1)}, \lambda^{(2)}, \dots$, is monotone increasing and converges to a limit λ for which $d_i(\lambda) = 0$, $i = 1, \dots, m$. For those values of i for which $d_i(\lambda) < 0$ it is, in fact, not necessary to choose c_i so that $d_i(\lambda + c_i e_i) = 0$; having $d_i(\lambda + c_i e_i) \leq 0$ and $|d_i(\lambda + c_i e_i)| \leq \theta |d_i(\lambda)|$, where θ is a fixed quantity less than one, would do. Since $d_i(\lambda_1, \dots, \lambda_m)$ is monotone increasing in λ_i and decreasing in λ_j for $j \neq i$, an admissible value of c_i can certainly be found in a finite number of steps simply by searching for it. The construction of a $\lambda^{(k+1)}$ from $\lambda^{(k)}$ can therefore be carried out effectively in a finite number of steps. It follows that in a finite number of steps one can by this method obtain an arbitrarily good approximation to a solution of the original transportation problem in the sense that it is arbitrarily close to feasible and its cost is arbitrarily close to minimal. In that sense the method may be called an algorithm. It does not, however, yield an exact solution in any finite number of steps, a reflection of the fact that while the original transportation problem was strictly combinatorial, the problem as modified here is analytic in that the existence of a solution depends on the properties of continuous functions. (Note, however, that if the threshold r is made sufficiently small and an exact solution to the order-equilibrium problem $d_i(\lambda) = 0$ is found, then the corresponding $x_{ij}(\lambda)$ are in fact a true optimal solution to the original problem, cf. Kelley [3].)

Properties of the algorithm. Anyone familiar with the already classical work on linear programming will recognize the subsidies λ which solve the order-equilibrium problem $d(\lambda) = 0$ (d stands for the vector (d_1, \dots, d_m)) as being closely related to the solution to the dual of the original transportation problem. (For full exposition see [4].) A significant property of λ which is not enjoyed by the classical dual is that the m values $\lambda_1, \dots, \lambda_m$ already completely determine the $m \times n$ quantities x_{ij} which constitute the solution to the problem, for $x_{ij} = x_{ij}(\lambda)$. This has the following consequences: (1) It makes the algorithm especially useful for very large problems. (2) It permits easy recomputation of a solution when the problem is perturbed. (3) It causes endless headaches when one actually tries to program the algorithm for existing computers in such a way that it converges with reasonable rapidity. Why these ensue will be clearer after a gross description of how the algorithm is actually carried out by a computer.

The machine for which all programming, and on which all numerical

experiments were done, is the Univac I. This machine has a small rapid access internal memory of 1000 12-digit words, each digit being any of 63 symbols, and a large slow access memory in the form of 10 servo-driven magnetic tapes. Each tape can hold 2187 60-word blocks. At the start of actual computation, i.e., after certain initial editing and data handling routines have been performed, the internal memory of the machine contains the a_1, \dots, a_n and initial estimates $\lambda_1^{(0)}, \dots, \lambda_m^{(0)}$ for the subsidies (which in advance of other information may all be taken to be zero). The cost matrix c_{ij} is written on one magnetic tape by columns, that is, the costs c_{1j}, \dots, c_{mj} of transporting a unit quantity from the m producers to the j th consumer are all written together, and together with this is written the name (or number) of the consumer and his demand, b_j . (The order in which the consumers are written is immaterial.) Each of these columns and corresponding values of b_j are written on one block of the magnetic tape (more blocks could be used but problem sizes so far have not required this) and we shall speak henceforth of these as consumer blocks. The consumer blocks are read one at a time into the internal memory. When the j th block is read in, the internal memory contains all information necessary to compute $x_{1j}(\lambda^{(0)}), \dots, x_{mj}(\lambda^{(0)})$ and this is done. These values are then written together with c_{1j}, \dots, c_{mj} and b_j on one block of magnetic tape (there is space in each consumer block to do this), and $x_{ij}(\lambda)$ is subtracted from a_i for each i . The next consumer block is then read in. When all consumer blocks have been so processed, the internal memory contains $\lambda^{(0)}$ as before, $-d(\lambda^{(0)})$, and some magnetic tape contains the n processed consumer blocks. From the values of $\lambda^{(0)}$ and $d(\lambda^{(0)})$ contained in the internal memory a new approximation $\lambda^{(1)}$ is computed and substituted for the old. The process may now be repeated of reading in consumer blocks one at a time, computing $x_{ij}(\lambda^{(1)}), \dots, x_{mj}(\lambda^{(1)})$ and tallying $-d(\lambda^{(1)})$ when all the blocks have been processed. In this way a sequence of pairs of vectors $\lambda^{(0)}, d(\lambda^{(0)}) ; \lambda^{(1)}, d(\lambda^{(1)}) ; \dots$, is computed, and these results are themselves written on another magnetic tape as soon as they are found in order that the work should not be lost in the event of machine breakdown. When $\lambda^{(0)}, d(\lambda^{(0)}), \dots, \lambda^{(k)}, d(\lambda^{(k)})$ have been found, $\lambda^{(k+1)}$ is computed on the basis of the information already generated.

The data in the internal memory at the time the consumer blocks are being processed consist of a subsidy vector (λ_i), a vector ($d_i(\lambda)$) representing $(-a_i + \text{a subtotal of } \sum_j x_{ij}(\lambda))$ —the subtotal depends on which consumer blocks have already been processed—and a consumer block. The first two vectors fill $2m$ words and the consumer block theoretically contains $2m + 2$ more, but in practice, by a special device, never more than 60 are used for the latter. The limitation on the size of m is therefore that the program for processing consumer blocks shall not require more than $1000 - (2m + 60)$ words. With the present program on Univac I, values of m up to 180 can be handled. Since the computation of the new subsidy vector from the

preceding ones occupies generally only a small part of the time needed for the whole computation, one can there trade time for space—always possible on a machine having a very large slow access memory—and we may therefore assume for the moment that it imposes no further limitation on m . (Actually, for a fixed value of m it imposes a limitation on the complexity of the procedure which may be used to compute the new approximation $\lambda^{(k+1)}$.) The value of n (number of consumers) is limited only by the number of blocks that Univac I can write on a magnetic tape, namely 2187. The maximum problem size that can be handled by the routine as presently programmed is therefore approximately 180×2100 ; the author has never seen data gathered for a problem even approaching this size. An increase in the size of the internal memory by any number of words would allow an increase of m by half that number. At present m is crowded by the size of the program but on, for example, the LARC presently under construction it would be feasible to have values of m up to 40 thousand! The value of n could easily be increased by using more magnetic tapes to carry the data.

The device used to write a whole column of the cost matrix—which may in the present routine contain as many as 180 items—in a fraction of one block of 60 words consists simply of leaving most of it out. This is done by selecting, for every consumer, those producers by whom that consumer is most likely to be supplied in an optimal solution, i.e., those for which the costs of transportation to that consumer are least. Only these costs are written on the consumer block, the others are discarded entirely. In the present program there is space to enter up to 19 costs for every consumer. Discarding the other costs is equivalent to setting them equal to $+\infty$. It is easy to give examples showing that this may make the problem impossible, let alone force the adoption of a solution which is not optimal. One can test, however, to see if an optimal solution to the problem with certain of the costs replaced by $+\infty$ is actually an optimal solution to the original problem as follows:

Introduce a fictitious $(m + 1)$ st producer and a fictitious $(n + 1)$ st consumer, the supply of the former being equal to the demand of the latter. Set the cost of transporting from any producer (including the fictitious one) to the $(n + 1)$ st consumer equal to zero, and the cost of transporting from the $(m + 1)$ st producer to the j th consumer equal to the least of those c_{ij} 's which have been suppressed, i.e., replaced by $+\infty$. Call this the augmented problem, and consider now the problem with suppressed costs, the augmented problem (which still has suppressed costs) and the original problem. Suppose a feasible solution to the suppressed problem has been found. One can trivially extend it to a solution of the augmented problem by having the fictitious producer supply only the fictitious consumer. (On the other hand, if a feasible solution x_{ij} to the original problem is given, then one can obtain one to the augmented problem so: For fixed i set $x_{i,n+1}$ equal to the sum of those x_{ij} for which c_{ij} is suppressed, for fixed j set $x_{m+1,j}$ equal to the sum of

those x_{ij} for which c_{ij} is suppressed; $x_{m+1,n+1}$ is then uniquely determined and positive if the supply assigned to the fictitious producer is sufficiently large.) It is now the case that an extension of a solution to the problem with suppressed costs (henceforth the reduced problem) is an optimal solution to the augmented problem if and only if the solution to the reduced problem is an optimal solution both to the reduced problem and the original. To see this it is most convenient to use the dual. From the fundamental duality theorem it follows that a feasible solution x_{ij} to the transportation problem with costs c_{ij} is optimal if and only if there exist real numbers λ_i, μ_j such that $(\alpha) c_{ij} - \lambda_i - \mu_j \geq 0$ for all i and j , and $(\beta) x_{ij} \neq 0$ implies $c_{ij} - \lambda_i - \mu_j = 0$. Considering the λ_i as producer subsidies and the μ_j as consumer subsidies, we may say that x_{ij} is optimal if and only if producer and consumer subsidies can be introduced in such a way as to make the new costs $c'_{ij} = c_{ij} - \lambda_i - \mu_j$ all non-negative while reducing the cost $\sum c'_{ij}x_{ij}$ of the solution x_{ij} to zero. Suppose now a solution $x_{ij}, i = 1, \dots, m, j = 1, \dots, n$ to the reduced problem is given whose extension is optimal. That x_{ij} is an optimal solution to the reduced problem is trivial; what we must show is that x_{ij} is an optimal solution to the original. By the duality theorem, there exist $\lambda_i, \mu_j, i = 1, \dots, m + 1, j = 1, \dots, n + 1$ satisfying (α) and (β) above. For certain pairs (i,j) , however, (α) is vacuous in the augmented problem since c_{ij} has been replaced by $+\infty$. To show that we have an optimal solution to the original problem, it is sufficient to show that (α) continues to hold with c_{ij} restored to its original value. If c_{ij} has been suppressed, however, we have $c_{ij} \geq c_{m+1,j}$ by construction of the augmented problem. Further, since $c_{i,m+1} = 0$ we have $\lambda_i + \mu_{m+1} \leq 0$, and since the extension of the solution is optimal, $\lambda_{m+1} + \mu_{m+1} = 0$. Now (α) holds in the augmented problem for $i = m + 1$: $c_{m+1,j} - \lambda_{m+1} - \mu_j \geq 0$. For the (i,j) under consideration, therefore, $c_{ij} - \lambda_{m+1} - \mu_j \geq 0$. This and the preceding relations yield $c_{ij} - \lambda_i - \mu_j \geq 0$, showing that (α) holds for the original problem and proving the optimality of the solution. The remaining part of the assertion says essentially that if a solution to the reduced problem is an optimal solution to the original then its extension is an optimal solution to the augmented problem; this one may prove similarly. One may further show that if for every j the least suppressed c_{ij} is strictly larger than the greatest non-suppressed one, and if the extension of an optimal solution to the reduced problem is still optimal, then the augmented problem has no optimal solutions other than those obtained by extending an optimal solution to the reduced problem. Since by arbitrarily small change in the costs one can always cause this strict inequality to hold, the augmented problem may be used as a strict test to see if too many costs have been suppressed. If after solving the augmented problem with $c_{m+1,j}$ replaced by $c_{m+1,j} + \epsilon$ with small ϵ whenever necessary, one finds that the solution obtained is the trivial extension of a solution to the reduced problem then the matter is at an end. Otherwise, one must suppress fewer costs to get a strict solution.

In the unfortunate case when a solution to the augmented problem is found which is not an extension of one to the reduced problem, one nevertheless has a lower bound on the total cost of transportation. For the procedure given of constructing a solution to the extended problem from one of the original clearly does not increase the cost, and one sees therefore that the total cost of an optimal solution to the augmented problem is not greater than that of the original. One may therefore still be willing to accept a solution to the reduced problem instead of restoring more costs if the cost of that solution is not too much more than the bound obtained by solving the extended problem.

It seems unlikely that except for very small m one should have to consider more than $m^{1/2}$ costs for any one consumer—as long as the problem is a real one and not created by a malicious programmer. (This is not true, however, if the “transportation problem” actually arose as an assignment problem.)

It should be remarked at this point that the algorithm under discussion is fast but can easily be beaten by any of several which carry more information in the internal memory at one time. If, for example, there were sufficient internal memory space to contain the whole cost matrix—and some room to spare for the program—then it would probably be better to use any of the very fast algorithms which involve scanning of the cost matrix both by rows and by columns, for example the algorithms of Ford-Fulkerson, Kuhn, or the variants due to Flood. The present program sees the cost matrix only one column at a time and does not make use of any information which must be obtained by scanning the rows. In fact, the availability of m^2 words in the internal memory would probably induce one to abandon the author's program for one which sees more of the cost matrix at once. However, given a fixed memory size large enough so that one can neglect the space occupied by the program, the present algorithm handles larger problems than any other which is fast enough to be practical.

The advantage of the present algorithm in recomputing a solution after the supplies, demand or costs have been perturbed is clear if one recalls that the solution $x_{ij}(\lambda)$ is completely determined by λ . Of the strictly combinatorial methods, those which first compute the solution to the dual probably come closest to sharing this advantage.

In view of what has already been said, one may well ask why threshold methods have not immediately supplanted all others, at least for the solution of large-scale problems. There is a serious difficulty, which stems from a lack of theoretical information at the most crucial point: How should one compute $\lambda^{(k+1)}$ from $\lambda^{(0)}, d(\lambda^{(0)}), \dots, \lambda^{(k)}, d(\lambda^{(k)})$? The rate of convergence depends on how one does this and it has taken much experimentation to develop certain workable rules. Of those which have so far proved feasible to program within the memory space available, that which has given best results is, surprisingly, a search technique. Not all of these, however, are equally good. The one currently in use records the changes in sign of

$d_i(\lambda^{(p)})$ as p varies from 0 to k , and computes $\lambda_i^{(k+1)}$ as a function of the times and frequencies of these changes. It has the property, therefore, of using a small part of the information generated at each of a large number of preceding iterations. Possibly techniques using more of the information from fewer iterations may be better, but not enough work has yet been done to tell. In any case, the simplest search routines which discard too much of the previous information are impossibly slow.

Empirical results and prognosis. Despite the programming difficulties encountered in adapting the algorithm here presented to a computer, it is feasible and problems of various sizes up to 15×488 have been run as tests. Solving the latter consumed 2.5 hours of computer time. The same problem had originally been solved using a program written when only the simplex method (in its crudest form) was available and consumed 60 hours. More recently it has been used to test an algorithm programmed by the Remington-Rand Systems Development Department in New York, which requires that an $m \times m$ submatrix of the cost matrix be carried in the internal memory. The latter program solved the problem in 72 minutes, a slow time resulting from unsophisticated choice of the initial feasible solution. Better choices reduced the time to 15 minutes; the mean time assignable to the program for a problem of this size probably lies somewhere between these limits. For problems of small size the present algorithm is highly inefficient, requiring almost an hour for a 10×10 problem. (The New York program solves this internally in a few seconds, but a 30×30 problem is beyond its limits.) To date, no larger problems than 15×488 have been run, but extrapolating from admittedly meager results, the efficiency for large problems should improve considerably as proportionately less time is consumed by motion of the program tape and similar "housekeeping" items. More important, better techniques for computing new approximate subsidy vectors will probably result in much shorter computation times. Some of these are currently being programmed but have not yet been tested.

Threshold methods have so far been applied successfully only to the transportation problem, although the theory can be carried through [3] for the general linear programming problem. The programming for the general problem, however, presents such formidable difficulties that it is unlikely that it could be successfully carried out on any presently available commercial machine, principally because of the limitations of internal memory. Among machines now under construction, however, are some for which the method would certainly be feasible, making possible in turn considerable expansion in the size of problems one can handle. In general we know at the present time too little about the possible usefulness of analytic techniques in linear programming (or for that matter, in other combinatorial problems). The present work indicates that at least in certain cases they compare favorably or surpass combinatorial methods when the problem size is at the limits of the machine's capacity.

REFERENCES

1. M. Gerstenhaber, *A solution method for the transportation problem*, J. Soc. Indust. Appl. Math. vol. 6 no. 4 (1958) pp. 321-334.
2. ———, *A procedure for finding a zero of a vector valued function with certain monotonicity properties*, To appear in J. Soc. Indust. Appl. Math.
3. J. E. Kelley, Jr., *Threshold methods for linear programming*, Naval Res. Logist. Quart. vol. 4 no. 4 (1957) pp. 35-45.
4. T. C. Koopmans (editor), *Activity analysis of production and allocation*, New York, Wiley, 1951.
5. H. W. Kuhn, *Methods for solving transportation problems*, mimeographed, Presented at the Techniques of Industrial Operations Research Seminar, Chicago, 1957.

INSTITUTE FOR ADVANCED STUDY,
PRINCETON, NEW JERSEY
UNIVERSITY OF PENNSYLVANIA,
PHILADELPHIA, PENNSYLVANIA

This page intentionally left blank

ON SOME COMMUNICATION NETWORK PROBLEMS

BY

ROBERT KALABA

I. INTRODUCTION

The general field of communication provides a rich source of problems in applied mathematics. These embrace fundamental considerations of the communication process itself [30], a wide spectrum of scientific and technological problems, and still others involving the design and utilization of large-scale networks. The rather modest objective of this paper is to draw attention to several classes of communication network problems, of some importance in the applications, which lead to combinatorial problems of varying degrees of complexity. Generally speaking, these problems are concerned with the optimal design and utilization of communication networks in which the complex interactions among users' demands for service, system capacities, and economic factors must be resolved. Pioneering efforts along these lines are associated with the names of A. K. Erlang [7], T. C. Fry [18], E. C. Molina [27], and R. I. Wilkinson [33], among others.

Problems of the type mentioned have been assuming increasing importance in recent years due to the rapid expansion of communication systems, involving large capital investments. Wide-sweeping technological improvements in both switching and transmission facilities will vastly alter the nature of the networks. Lastly, the advent of the high-speed large-memory digital computing machine has forced a re-evaluation of the very methods of analysis and design which are in current use. Though the models which we shall discuss are highly simplified, their analysis may point the way toward the treatment of more refined and realistic ones.

The problems to be discussed lead, from the mathematical point of view, to the determination of extrema from among finitely many choices, so that no questions concerning existence of solutions arise. Interest centers rather on obtaining algorithms which lead to efficient computational schemes for obtaining solutions, and which shed light on the structure of the solutions. Finding solutions in these problems through the mere enumeration of cases, as remarked by Euler in his famous paper on the Königsberg bridge problem, is at best onerous and unsatisfying and in many situations impossible (even with the aid of a high-speed computing machine), as will become evident.

The first type of problem which we shall consider is that of determining minimal cost connecting networks. Given a network, each link of which has a cost assigned to it, find a connected network which includes all the stations and has least total cost. Solutions have been proposed by Kruskal

[24], Prim [29], and Kalaba in the forms of algorithms which lend themselves well to hand and machine computation and which provide much insight into the nature of the solution. This problem can be generalized along various lines.

The second type of problem is that of determining an optimal chain connecting two points in a network. Perhaps the simplest version of this type is to find a shortest chain connecting two terminals in a given network, each link of which has a prescribed time of transit. Solutions have been provided by Bellman [2], through the use of functional equations, Dantzig [10], using a linear programming approach, and Ford [14]. The problem may be modified by requiring that the chain pass through several specified intermediate points, and it still remains amenable to treatment. Furthermore, Bellman has proposed a method for finding the n th shortest chain leading from one point to another in a network.

The methods to be discussed make possible the solution of certain optimal chain problems involving probabilistic considerations. In particular, the problem of determining a path through a network which maximizes the probability that the time of transit between two given points be no greater than a prescribed time t is solved, using functional equations, under the assumption that the times of traverse of the various branches are independent random variables with known probability densities. In addition some applications to the theory of blocking in networks are provided [25].

The last type of problem discussed involves the optimal routing of messages in networks [20]. Under certain conditions one may formulate this as a linear programming problem for which Dantzig's simplex method is available for numerical solution, provided the network is not too large. A method of solution based on an idea of Ford and Fulkerson [16], makes possible the numerical solution of problems involving about 150 links. Finally, some related problems involving interoffice trunking and the augmentation of networks to meet increased demands for service are discussed [21].

Acknowledgement. During the preparation of this paper the author has had the benefit of many discussions with R. Bellman, G. Dantzig, D. R. Fulkerson, and M. L. Juncosa and expresses his sincere thanks for their friendly interest and suggestions.

II. MINIMAL COST CONNECTING NETWORKS

1. Formulation. A television broadcasting company wishes to lease video links so that its stations in various cities may be formed into a connected network. Assuming that the costs for the individual links, all different, are known, we wish to show how to construct the network at minimal cost [1]. (Continuity considerations enable one to remove the restriction that the costs be different, but, as will become evident, uniqueness of the solution may be lost.)

Various solutions for this problem will now be discussed and some extensions will be indicated.

2. Solution I. Kruskal [24], has proposed the following solution, the simplicity of which is quite remarkable. Perform the following step as often as possible: Among the links not yet included in the connecting network, choose the lowest priced link which does not form any loops with the links already chosen. The proof, which follows, is by contradiction.

If there are N stations in the network, it is evident that a minimal cost connecting network, denoted by K , contains no loops and consists of exactly $N - 1$ links. Let the links chosen according to the above algorithm be denoted by e_1, e_2, \dots, e_{N-1} ; since the costs are all different from each other, this sequence is uniquely determined. This set of links is denoted by L_{N-1} .

If $K \neq L_{N-1}$, let e_i be the link of lowest index of L_{N-1} which is not in K . If e_i is added to the set K , a loop is formed of which e_i is one link. This loop also contains a link, f , which is not in L_{N-1} but which is in K . Furthermore, the link f does not close a loop when added to the set e_1, e_2, \dots, e_{i-1} , for all these links, including f , lie in the set K , which contains no loops. But according to the algorithm e_i is the lowest priced such link; consequently

$$(1) \quad \text{price}(f) > \text{price}(e_i).$$

This implies that the network consisting of the union of K and e_i from which f has been deleted, which also contains $N - 1$ links and does not contain any loops, is available at lower cost than K , contrary to assumption. Hence the Kruskal tree $L_{N-1} = K$ is the unique minimal cost connecting network.

3. Solutions II and III. In the same paper referred to above, Kruskal also proposes two additional constructions. Let S be an arbitrary, but fixed and non-empty subset of all the N stations to be joined into a connected network. Perform the following step as often as possible: Among the links not yet chosen, but which are connected either to a station in S or to a link already chosen, choose the link of lowest price which does not form any loops with the links already chosen. This reduces to the construction of §2 if S consists of all the stations in the network.

The other consists in determining the links not in K by choosing as many times as possible, from among the links not yet chosen, the most expensive link which does not disconnect the network. The set of links not eventually chosen forms the minimal cost connecting network K . This may be established by showing that it is always possible to remove a link from consideration for membership in K if the link is the most costly link whose removal from the network does not disconnect it. Let A be the set of links which can be removed without disconnecting the network, and let e be the one of greatest cost. Suppose e to be in K . The removal of the link e

from the set K disconnects this network, which can, however, be reconnected by the addition of a link f which is contained in the set A and is different from e ; for if this were not the case, e could not be an element of the set A . Consequently, the union of K and f , from which e is deleted, would be available at lower cost than K , which results in a contradiction.

4. Solution IV. Still another algorithm is available in which we proceed from one connecting network containing no loops to another of lesser cost until the optimal network K is attained. Select any connecting network with precisely $N - 1$ links. Add another link to this network so that a loop is formed and eliminate from the loop the most costly link. Repeat until no further changes in the connecting network are possible. The resulting network T is the optimal network K . For suppose $T \neq K$ and that e_m is the link of smallest index in the Kruskal construction of §2 which is not in T . Add e_m to T to form a loop. This loop contains at least one link e'_m which is in T but is not in K . Furthermore, adding e'_m to the set of Kruskal links e_1, e_2, \dots, e_{m-1} cannot complete a loop since all these links, including e'_m , lie in the network T , which is free of loops. Therefore

$$(2) \quad \text{price}(e_m) < \text{price}(e'_m),$$

so that the algorithm calls for adding e_m to T and eliminating e'_m from T , which is contrary to the assumption, under the rules of the algorithm, that no further changes in T are possible.

5. Solution V. Though the algorithms mentioned above rather clearly show the structure of the minimizing network, they are not the best insofar as rapid computation of the solution is concerned. In this regard, a suggestion of Prim [29], involving a combination of algorithms I and II, is probably the best, for it avoids considerations of loops and connectedness, and makes rather modest memory requirements on a computing machine.

It is a simple matter to determine the most costly connecting network using similar procedures. Prim has also called attention to the fact that the minimizing connecting network K also minimizes all increasing symmetric functions and maximizes all decreasing symmetric functions of the link costs, among all connecting networks with no loops.

III. OPTIMAL PATHS THROUGH NETWORKS

In this section we shall discuss a variety of problems involving the determination of optimal paths through networks. The first of these, and perhaps the simplest, which will be attacked in several ways, involves the determination of a path of minimal time of transit between two points of a network. It is assumed that the time of transit of each link is known. It is then shown how the n th shortest path (or paths) can be determined. The former problem is closely related to finding a path between two points in a network which has minimum probability of being blocked, given the probabilities

that the individual links are blocked, and the fact of the independence of the individual link's being blocked. Lastly an interesting extension is indicated which consists in determining a path between two points in a network which maximizes the probability of being traversed in a time t or less, being given the probability densities for the times of transit of the individual links and the information that the times of transit are independent.

It is clear that problems of the types just mentioned are of importance in the study of networks where the possibility of alternate routing exists. This will become even more apparent in Part IV in which it is shown that these problems are intimately connected with the general problem of optimal routing of messages in networks.

6. Formulation and solution. Given a network consisting of N stations and interconnecting links, where the time to traverse link (i,j) is $t_{ij} \geq 0$, $t_{jj} = 0$, find a shortest path from point 1 to point N . Note that t_{ij} need not equal t_{ji} and that t_{ij} need not be proportional to the distance between points i and j . Our first approach is based upon that given by Bellman [2], in which the original problem is imbedded within the class of problems of determining the shortest paths from any point i in the network to the point N .

The problem is a combinatorial one in which we seek the minima of times of transit of a finite number of paths. For N of the order of twenty, the enumerative approach becomes quite onerous, so that we seek more efficient methods of obtaining the extrema.

Denote the time of transit from i to N via an optimal path by u_i . Employing the principle of optimality [1], we are led to the system of nonlinear equations

$$(1) \quad \begin{cases} u_i = \min_{j \neq i} \{t_{ij} + u_j\}, & i = 1, 2, \dots, N - 1, \\ u_N = 0. \end{cases}$$

To resolve this system, following Bellman, we resort to the method of successive approximations. As an initial approximation we set

$$(2) \quad u_i^{(0)} = t_{iN}, \quad i = 1, 2, \dots, N,$$

which corresponds physically to traversing the direct links from points i to N . The higher approximations are then obtained through use of the formulas

$$(3) \quad \begin{cases} u_i^{(k+1)} = \min_{j \neq i} \{t_{ij} + u_j^{(k)}\}, & i = 1, 2, \dots, N - 1, \\ u_N^{(k+1)} = 0, \end{cases}$$

for $k = 0, 1, 2, \dots$. It is readily seen that $u_i^{(k)}$ is the minimal time of transit from point i to point N via k intermediate points. Since the sequence $u_i^{(k)} \geq 0$ is monotone non-increasing in k , the sequence converges to a solution of equation (1) in no more than $N - 2$ iterations beyond the initial

one. Furthermore, as Bellman has shown, the solution is unique, though an optimal path need not be.

This furnishes a feasible method for machine calculation with N of the order of several hundred. Since only additions and comparisons are required, the computation proceeds rapidly. Moreover, the memory requirement for the computation of $u_i^{(k+1)}$ is modest, since for each value of i only the i th row of the matrix (t_{mn}) is required in addition to the previously computed values $u_j^{(k)}$.

It is also possible to obtain a monotone increasing sequence of approximations. Let

$$(4) \quad \begin{cases} u_i^{(0)} = \min_{j \neq i} t_{ij}, & i = 1, 2, \dots, N - 1, \\ u_N^{(0)} = 0, \end{cases}$$

be the initial approximation, and let the additional approximations be determined by the relations in equation (3). We can see inductively that the sequence is monotone non-decreasing and furthermore that

$$(5) \quad u_i^{(k)} \leq u_i, \quad i = 1, 2, \dots, N, \quad k = 0, 1, 2, \dots,$$

where u_i is the solution of equation (1). For $k = 0$ the inequality (5) is valid. Hence if we assume it holds for $k = m$, we obtain

$$(6) \quad u_i^{(m+1)} = \min_{j \neq i} \{t_{ij} + u_j^{(m)}\} \leq \min_{j \neq i} \{t_{ij} + u_j\} \leq u_i,$$

which completes the induction and establishes the monotone convergence of the sequence defined by equations (3) and (4).

Observe that if a shortest path connecting 1 to N through one intermediate point, m , is required, the solution is given by the sum of the shortest chains connecting 1 to m and m to N . Should two intermediate points, m and n , be specified, the solution is the shorter of the chains $(1, m, n, N)$ and $(1, n, m, N)$ where each pair of nodes, $(1, m)$, (m, n) , (n, N) , etc. is joined by a shortest chain. If the number of intermediate points is small, then a shortest path can be determined through enumeration of cases, the computation of a shortest path between two specified points being effected as above.

7. Solution II. Another technique for solving the problem posed at the beginning of §6 is contained in an algorithm described by Ford [14], and others. It is, of course, simply another way of solving equations (6.1) and runs as follows. Assign the value $0 = u_N$ to the node N and $u_i = \infty$ to the nodes $i \neq N$. Hunt through the network until a pair of points i and j with the property that

$$(1) \quad u_i > t_{ij} + u_j$$

is found, should there be any such. Then replace u_i at the node i with the smaller value $t_{ij} + u_j$. Repeat this step until no pairs fulfilling the inequality

(1) remain. The numbers u_i then assigned to the nodes i represent the minimal times of transit from these nodes to the node N . This will now be proved.

Let i, i_1, i_2, \dots, N be an optimal chain from i to N . We have

$$(2) \quad u_i - u_{i_1} \leq t_{ii_1},$$

with similar inequalities holding for every link in the chain. Through addition of all these inequalities we find that

$$(3) \quad u_i \leq \text{minimal time of transit from } i \text{ to } N.$$

On the other hand, for every node $m \neq N$ there is a link from m to a node n for which

$$(4) \quad u_m = t_{mn} + u_n.$$

All nodes except N were initially assigned the values ∞ , and these values have been monotone decreasing (or else have not changed at all). At the last decrease in u_m there is an n which still has the same value. We can trace a chain from i to N composed of links for which equalities such as that in equation (4) hold. The values at the nodes are decreasing. Eventually the point N must be attained. Along this chain

$$(5) \quad u_j - u_k = t_{jk}.$$

A summation yields that

$$(6) \quad u_i = \text{time of transit of this chain.}$$

Consequently this is a shortest chain.

An elegant version of this algorithm has been suggested by Dantzig. It enables one to determine the minimal times of transit from i to N and the paths to be traversed through use of a constructive procedure reminiscent of Kruskal's algorithm. First determine a closest point to N , say P_1 , and record the time of transit from P_1 to N . Then determine a closest point to P_1 , say Q , and also a point which is second closest, via a direct path, from N , say R . Determine the smaller of t_{RN} and $t_{QP_1} + u_{P_1}$. This yields P_2 , the second closest point to N , and an optimal path from P_2 to N . A comparison among the times to travel to N from the closest unchosen point to N , via a direct path, and the closest points to P_1 and P_2 , continuing from P_1 or P_2 along the paths already selected, yields P_3 , and so on.

If there are N stations in the network, this procedure will result in solution after at most $1 + 2 + (N - 1) = ((N - 1)/2)N$ comparisons. This assumes that for each node in the network the remaining ones have been arranged in order according to the times of transit from the given node to each of the others.

8. The n th shortest chains. It has been noted by Bellman that the n th shortest paths can also be conveniently determined through use of functional

equations. The importance of this resides in the fact that this enables us to see how sensitive to change the times of transit are for paths in neighborhoods of optimal paths. This has implications for the general theory of multi-stage decision processes which will be discussed elsewhere [5].

We define u_i , $i = 1, 2, \dots, N$, as in the previous section and introduce the quantities

$$(1) \quad v_i = \text{time of transit of a second shortest path from } i \text{ to } N,$$

for $i = 1, 2, \dots, N - 1$.

Next we observe that if the first link in a second shortest route is the link (i, j) then the continuation from j to N must be along either a path which minimizes the time of transit from j to N or which is a second shortest path from j to N , no others being possible. These lead to total durations of the routes from i to N of $t_{ij} + u_j$ and $t_{ij} + v_j$ respectively. Hence v_i is equal to the smaller of the following two values : the second smallest value of $t_{ij} + u_j$, $j \neq i$, and the smallest value of $t_{ij} + v_j$, $j \neq i$. If Min_k refers to the operation of taking the k th smallest value of a given set, with $\text{Min}_1 = \text{Min}$, the resulting equations are

$$(2) \quad v_i = \text{Min} \left\{ \text{Min}_{j \neq i} (t_{ij} + v_j), \text{Min}_{j \neq i} (t_{ij} + u_j) \right\},$$

for $i = 1, 2, \dots, N - 1$.

The generalization to the calculation of the n th shortest paths is evident, though various questions concerning the numerical solution of the equations arise.

9. Solutions by analogue computation. The problem of determining the shortest path between two points in a network may also be solved by constructing a string model [34] in which inextensible strings of lengths proportional to the times of transit are connected between all pairs of nodes in a network. A path of minimal time of transit between two nodes is then determined by separating the selected pair of nodes to the greatest extent possible. The links in chains which are stretched taut form optimal paths, and the distance of separation of the points measures the time of transit over an optimal path.

Electrical analogues can also be employed. Each branch of the network is replaced by gas tubes whose breakdown voltage is proportional to the times of transit, and the terminals of a current source are connected to points under consideration. The paths over which current flows are optimal.

See also [26] for a discussion of related matters, including use of soap-film models.

10. Some stochastic problems. We now turn our attention to some extensions in which various probabilistic elements are introduced. Consider a switching network in which the probability that a link from m to n is

available for service is p_{mn} . The problem is to determine a path from i to N which has the greatest probability of being available for service (i.e., unblocked).

We introduce a set of variables P_i , $i = 1, 2, \dots, N$, defined by the relation

- (1) P_i = the probability of no blocking on an optimal path from i to the point N .

This leads to the relations

$$(2) \quad \begin{cases} P_i = \underset{j \neq i}{\text{Max}} p_{ij}P_j, & i = 1, 2, \dots, N - 1, \\ P_N = 1, \end{cases}$$

which, similarly to the equations discussed earlier, can be resolved through use of the successive approximations

$$(3) \quad \begin{cases} P_i^{(k+1)} = \underset{j \neq i}{\text{Max}} p_{ij}P_j^{(k)}, & i = 1, 2, \dots, N - 1, \\ P_N^{(k+1)} = 1, \end{cases}$$

for $k = 0, 1, 2, \dots$, along with the initial approximation

$$(4) \quad \begin{cases} P_i^{(0)} = p_{iN}, \\ P_N^{(0)} = 1. \end{cases}$$

The sequence is clearly monotone increasing.

Now let us suppose that the time to traverse the link from i to j is a random variable t_{ij} with probability density function $p_{ij}(s)$, $i \neq j$, $s \geq 0$, and that the times of transit of the various links are independent. The treatment of the problem in which we seek a path from i to N for which the average time of transit is minimum is evident. Let us therefore turn to the problem in which we require a path connecting the point i to the point N which maximizes the probability that the time of transit is no greater than a given time t . Again using the principle of optimality, after introducing the functions $u_i(t)$, $i = 1, 2, \dots, N$, to be the probability that the time of transit from i to N is no greater than t , using an optimal path, we find

$$(5) \quad \begin{cases} u_i(t) = \underset{j \neq i}{\text{Max}} \int_0^t p_{ij}(t-s)u_j(s)ds, & i = 1, 2, \dots, N - 1, \\ u_N(t) = 1. \end{cases}$$

Once again we may resort to the method of successive approximations to resolve this nonlinear system :

$$(6) \quad \begin{cases} u_i^{(k+1)}(t) = \underset{j \neq i}{\text{Max}} \int_0^t p_{ij}(t-s)u_j^{(k)}(s)ds, & k = 0, 1, 2, \dots \\ u_N^{(k+1)}(t) = 1. \end{cases}$$

As initial approximations we take

$$(7) \quad \begin{cases} u_i^{(0)}(t) = \int_0^t p_{iN}(s)ds, & i = 1, 2, \dots, N - 1, \\ u_N^{(0)}(t) = 1, \end{cases}$$

which yields approximations that are monotone increasing. The initial approximation

$$(8) \quad \begin{cases} u_i^{(0)}(t) = \text{Max}_{j \neq i} \int_0^t p_{ij}(s)ds, & i = 1, 2, \dots, N - 1, \\ u_N^{(0)}(t) = 1, \end{cases}$$

yields monotone decreasing approximations.

IV. OPTIMAL ROUTING PROBLEMS

A problem of considerable importance in the operation of communication systems is that of the determination of the routing doctrine to be used in handling the messages. Large systems frequently employ a central traffic control unit for this purpose. Information concerning backlogs of messages is periodically sent to this control unit, as is information concerning the state of the communication system itself (wires may be down, equipment may be malfunctioning, etc.). On the basis of this information plus predictions concerning the new demands for service, decisions are made concerning the way the messages are to be routed through the network. Inefficiencies in the routing of the messages are reflected in the need for greater quantities of equipment for a fixed grade of service.

The papers referred to in the introduction, some of which contain extensive bibliographies, indicate a mathematical treatment of these problems based on probability theory. Interest centers on fluctuations in the traffic. Here we shall consider a steady state formulation for these problems which leads to a linear programming setting. A further discussion can be found in [20].

Even for moderately sized networks of about thirty stations the problems become so large that solution is not feasible through use of the general simplex method of George Dantzig [11]. Instead we resort to use of a modification of the simplex method which was originally proposed for multi-commodity flow problems and which is due to Ford and Fulkerson [16]. First the general approach is sketched and then a simple optimal routing problem is worked in detail for illustrative purposes.

11. Problem formulation and method of solution. We now reduce this version of the problem of the routing of messages in a network to mathematical form. Introduce the quantities

- (1) d_{ij} = the number of messages available at i which are destined for j ,

- (2) c_{ij} = the number of messages which can be sent over the direct link from i to j .

All action is assumed to take place during a given time interval. Next label all the directed links in the network L_1, L_2, \dots, L_m , and label all the directed routes in the network which lead from a source to a destination R_1, R_2, \dots, R_n . We describe the composition of the routes in terms of the links through use of the $m \times n$ incidence matrix (a_{ij}) , where

$$(3) \quad a_{ij} = \begin{cases} 1, & \text{if link } i \text{ lies in route } j \\ 0, & \text{otherwise.} \end{cases}$$

If the link from i to j is labeled s , we set

$$(4) \quad c_{ij} = c_s.$$

At each source S_i it is convenient to modify the original network by introducing a set of fictitious sources, $S_i^{(r)}$, which are connected to S_i by fictitious directed links, each fictitious source $S_i^{(r)}$ corresponding to messages originated at i which are destined for the station r . The capacity of each fictitious link $(S_i^{(r)}, i)$ is d_{ir} . If d_{ir} is zero, then the fictitious source and link are not introduced. In this way all messages are conceived of as arising at fictitious sources; the messages flow over the fictitious links and then the actual links to their destinations.

Hence all constraints, as in relations (6) and (7) below, appear as capacity constraints, including those that are due to the limited supplies of messages available for delivery. All routes lead from fictitious sources to their destinations, and we shall assume that the incidence matrix (a_{ij}) has reference to the modified network. In particular m is the sum of the numbers of actual and fictitious links. We shall henceforth not distinguish between fictitious stations and actual stations. Lastly we let x_j be the number of messages which flow over the route j , $j = 1, 2, \dots, n$.

The problem involves the maximization of the number of delivered messages

$$(5) \quad d = \sum_{j=1}^n x_j,$$

subject to the constraints

$$(6) \quad x_j \geq 0, \quad j = 1, 2, \dots, n + m,$$

$$(7) \quad \sum_{j=1}^n a_{sj} x_j + x_{n+s} = c_s.$$

Here we have denoted the amount of unused capacity in link s by x_{n+s} , $s = 1, 2, \dots, m$.

If the problem is to maximize the revenue derived from the operation

and r_j is the return from sending a message over the j th route, then the objective form becomes

$$(5') \quad r = \sum_{j=1}^n r_j x_j.$$

As was remarked earlier, n , the number of routes becomes so large, even in moderately sized networks, that it is not possible to determine an optimal linear program through use of the simplex method in its most general form, even if use of a high-speed digital computer is contemplated. The memory requirements for storage of the matrix (a_{ij}) alone become excessive. Hence we resort to a modification of the simplex method in which only m columns of the matrix, the basic vectors, need be stored simultaneously. At each stage of the simplex algorithm the new vector to be brought into the basis is determined by several applications of one of the algorithms described earlier for determining a shortest path connecting two points in a network.

From the general theory of linear inequalities we know that there is an optimal routing of the messages in which no more than m of the activities of sending a message over a route or storing capacity on a link are raised above the zero-level. Using this fact we can place the entire algorithm on quite an intuitive basis. We start by storing all capacity on all links, so that $x_{n+s} = c_s$, $s = 1, 2, \dots, m$. We then show how to improve the routing doctrine at any stage of the process by raising some favorable activity from zero-level to some positive level which is high enough to drive the level of some other formerly nonzero-level activity down to zero-level. To determine which new activity to introduce we consider the "shadow" prices which are induced on the capacities as a result of nonzero activities which are carried out at a particular stage. See [12] for a general discussion. The prices that are assigned to the fictitious links may be thought of as being franchise prices, that is, unit prices of the right to accept messages at one station destined for another. Prices assigned to the actual links are unit prices for the equipment.

Let p_j be the value of each unit of capacity in link j . For each additional sending of a message over a route j , $j = 1, 2, \dots, n$, the number of messages delivered is increased by unity. The unit prices of the capacities in the links along the routes used must therefore sum to unity,

$$(8) \quad \sum_{s=1}^m a_{sj} p_s = 1, \quad j \text{ such that } x_j > 0, j \leq n.$$

On the other hand there is a zero return for storing capacity so that

$$(9) \quad p_{j-n} = 0, \quad x_j > 0, \quad j > n.$$

The equations (8) and (9) determine the m prices p_1, p_2, \dots, p_m .

Let us now introduce an entrepreneur who examines the system capacities,

the users' demands and the price structure in an effort to determine whether or not it is possible to buy capacity from the communication network operating company, according to the price schedule $\{p_i\}$ and deliver messages himself at a profit, a delivered message being worth one unit. That is, the entrepreneur wishes to ascertain whether or not there is a route j for which

$$(10) \quad \sum_{s=1}^m a_{sj} p_s < 1, \quad j = 1, 2, \dots, n.$$

If there is such a route, though, it would be advantageous for the operating company to send messages over that route to the greatest extent possible. In general sending messages over route j will use capacity that was being used for sending other messages, so that some other activities may have to be curtailed, until finally at least one is reduced to zero-level, and so is eliminated. In any event, capacity constraints prevent x_j from increasing indefinitely.

Should the price of a certain link be negative, then this may be interpreted to mean that the communication system operating concern would be willing to pay the entrepreneur a subsidy to take this capacity from it. Rather than do this, this capacity should be sent to storage, so that if p_j is negative it is advantageous to raise x_{n+j} above the zero-level.

Assuming that all the prices are positive, how can the entrepreneur determine a route for which condition (10) is fulfilled? Since each unit of capacity is assigned a price, including the capacity of the fictitious links, he has merely to determine a lowest-price route from each source to destination. As soon as one is found for which the price is less than unity, as many messages as possible should be sent from this source to destination. If there is no such route, then the routing doctrine being employed is optimal, as one sees from the duality theorem of linear programming. This idea constitutes the essence of the delightfully simple suggestion of Ford and Fulkerson.

To summarize, the steps in the algorithm are:

1. Under the current price schedule determine a favorable activity to introduce. If a price is negative, store as much as possible of the corresponding capacity ; otherwise determine, via one of the algorithms discussed in Part III, a route having cost less than unity, and introduce this activity. If there is none, the routing doctrine is optimal.
2. Increase the level of the favorable activity until some activity which was previously at a nonzero-level is driven to zero-level. This determines the new routing doctrine and the number of messages which are thereby delivered.
3. Determine the new schedule of unit prices on the capacities and return to Step 1.

An illustrative example is provided below to illustrate this technique.

If the total number of links, including the fictitious ones, is of the order of 150, the steps of the algorithm are possible for implementation on a high-speed computing machine. It is difficult to try to estimate the rate at which the approximations converge to an optimal solution, since the number of chains might be numbered in the tens of thousands. Some numerical experimentation is undoubtedly called for. In actual computations, great advantages might be realized by being very selective with regard to which favorable activity is to be introduced at each stage.

12. Solution of an illustrative optimal routing problem. Consider the four-station network shown below in Figure 1, in which the capacities of

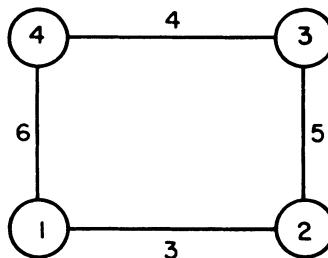


FIGURE 1. A capacitated four-station network

the links are as shown. We assume that the link capacities are undirected rather than directed, a matter of no importance insofar as the method is concerned. Consider that station 4 has 5 messages destined for station 2

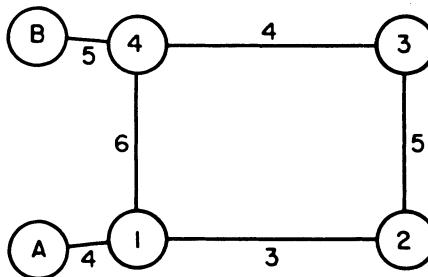


FIGURE 2. The network including the artificial elements

and station 1 has 4 messages destined for station 3. This is accounted for in Figure 2 in which the appropriate artificial stations and links are intro-

duced. Station B has messages destined for 2 and station A has messages destined for 3. Since there are six links, there is an optimal solution with no more than six activities raised above the zero-level.

To start the algorithm we put all capacity in storage, which corresponds to backlogging all messages. Since no messages are delivered, all capacities have prices of zero. By inspection, we see that the unit cost for the route $(A, 1, 2, 3)$ is zero; consequently three messages are sent over this route. This eliminates the activity of storing capacity in link $(1, 2)$. Letting the unit price of capacity in link $(B, 4)$ be p_1 , that in $(A, 1)$ be p_2 , and so on, as shown in Figure 3, we find that the prices satisfy the equations

$$(1) \quad p_1 = p_2 = p_3 = p_4 = p_6 = 0,$$

$$p_2 + p_5 + p_4 = 1.$$

The situation is shown in Figure 3 in which the amounts of capacity used are shown above the horizontal lines on the links and the capacities below them. The prices implied by equations (1) are also indicated.

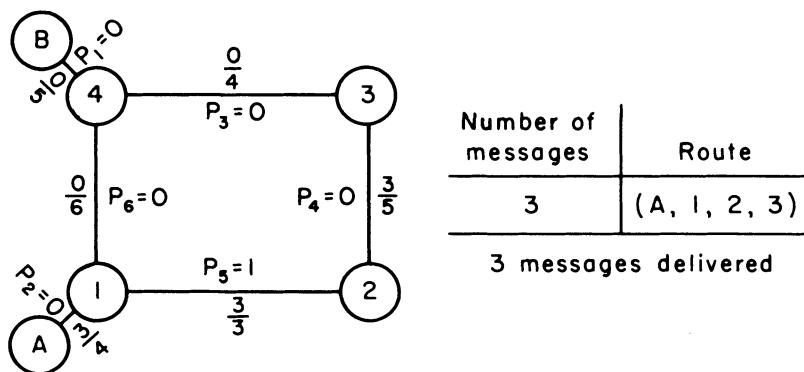


FIGURE 3. The first stage

The unit price of the route $(A, 1, 4, 3)$ is zero. One message is sent over this route, which eliminates the activity of storing capacity along the fictitious link $(A, 1)$ (i.e., the activity of backlogging messages at A is eliminated). With this routing schedule the equations for the new prices become

$$(2) \quad p_1 = p_3 = p_4 = p_6 = 0,$$

$$p_2 + p_5 + p_4 = 1,$$

$$p_2 + p_6 + p_3 = 1.$$

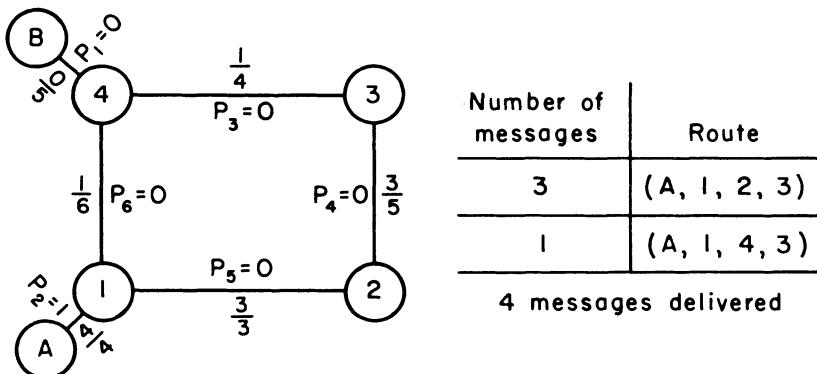


FIGURE 4. The second stage

The unit price of the route $(B, 4, 3, 2)$ is zero. Two messages are sent over this route which saturates link $(3, 2)$. The new prices are determined from the equations

$$(3) \quad \begin{aligned} p_1 &= p_3 = p_6 = 0, \\ p_2 + p_5 + p_4 &= 1, \\ p_2 + p_6 + p_3 &= 1, \\ p_1 + p_3 + p_4 &= 1, \end{aligned}$$

which lead to the situation of Figure 5.

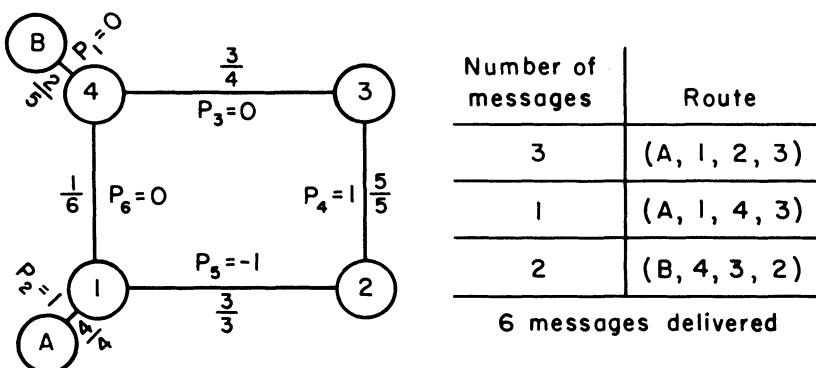


FIGURE 5. The third stage

Since the unit price p_5 is negative, the activity of storing capacity on the link $(1, 2)$ is reintroduced in the amount z . To find which activity is

eliminated we note that $3 - z$ messages are then sent over the route $(A, 1, 2, 3)$, which causes $2 + z$ to be sent over $(B, 4, 3, 2)$, to avoid introducing the storage of capacity on the link $(2, 3)$. The number of messages sent over $(A, 1, 4, 3)$ must be increased to $1 + z$ to avoid introducing the backlogging of messages at station A . Then an examination of the flows over each link shows that z can be increased to $1/2$, at which point the link $(4, 3)$ is saturated, so that storage of capacity on this link is eliminated. The new routing and price schedule are shown in Figure 6.

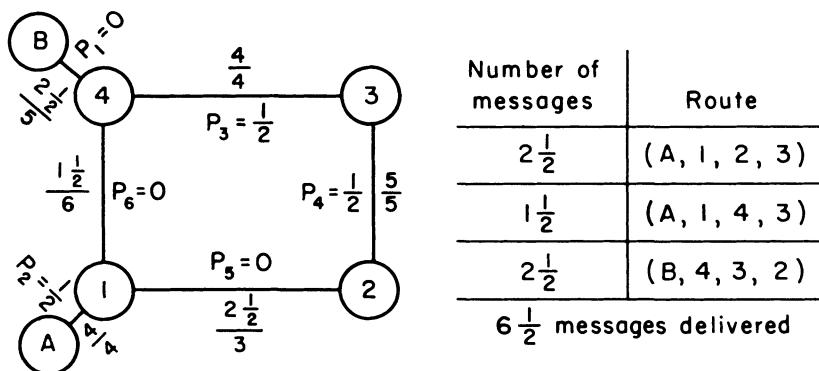


FIGURE 6. The fourth stage

The prices are determined from the equations

$$(4) \quad \begin{aligned} p_1 &= p_5 = p_6 = 0, \\ p_2 + p_5 + p_4 &= 1, \\ p_2 + p_6 + p_3 &= 1, \\ p_1 + p_3 + p_4 &= 1. \end{aligned}$$

The route $(B, 4, 1, 2)$ now has zero unit price, so that $1/2$ message is sent along this route, which eliminates storage of capacity on the link $(1, 2)$. This leads to the situation of Figure 7. The prices are determined from the equations

$$(5) \quad \begin{aligned} p_1 &= p_6 = 0, \\ p_2 + p_5 + p_4 &= 1, \\ p_2 + p_6 + p_3 &= 1, \\ p_1 + p_3 + p_4 &= 1, \\ p_1 + p_6 + p_3 &= 1. \end{aligned}$$

The solution is optimal, for, as is seen from the figure, no prices are negative and no paths from origin to destination exist which have unit costs of less than one.

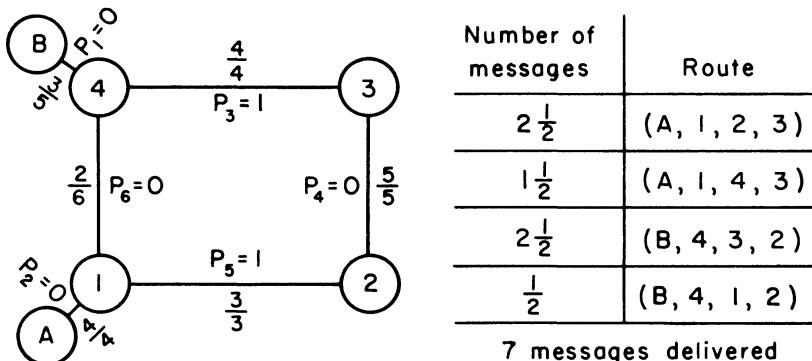


FIGURE 7. The last stage

By the way of comment it should be pointed out that in solution of large scale problems much more efficient methods of computing the prices and determining the new activity levels, as we proceed from stage to stage, are available. The method adopted here is for illustrative purposes only.

The unit prices shown in Figure 7 show where the real bottlenecks in the system are. Thus for z sufficiently small, if the capacity of link $(1, 2)$ is increased by the amount z , then z additional messages can be sent over the route $(B, 4, 1, 2)$. The same holds true for the link $(4, 3)$, though this is not so obvious. This rests on the observation that the solution is not unique. If messages are recalled along each route from B to 2, as is clearly possible, since the flow on each link is both increased and decreased by z , then for $2z = z$ sufficiently small, the messages which are then backlogged at A could be sent via the route $(A, 1, 4, 3)$ to their destination.

V. DISCUSSION

The problems mentioned earlier should be viewed merely as suggestive of a host of other essentially combinatorial problems which arise in the general field of communication. We shall now single out a few for further discussion; still others can be found by checking the list of references provided, not all of which are referred to in the text.

A routing problem which has been studied extensively is the one which requires the determination of the maximum steady state flow of a homogeneous commodity through a capacitated network from a source to a sink, Boldyreff's flooding technique is described in [6], and the minimum cut maximum flow theorem is proved in [15], where additional references can be found.

It is apparent that many combinatorial problems arise in the design and utilization of switching networks [22; 19]. In the absence of suitable

analytic techniques for handling such problems, one frequently resorts to the use of simulation devices known as "throwdown" machines [17]. With reference to blocking in networks [9; 25], it would be of interest to find general and efficient techniques for calculating the probability of finding at least one path available from a given point i to another point N in a network, under various assumptions concerning the probabilities of finding the individual links available. The functional equation technique of Part III does not appear to be immediately applicable, as in the determination of a path with highest probability of being available.

Interesting treatments of the effects of congestion in the networks, from still a different viewpoint, can be found in Wardrop [32], Prager [28], and Charnes and Cooper [8].

The solution of the optimal routing problem discussed in Part IV can be directly applied to the determination of optimal interoffice trunking arrangements as is indicated in [21]. Other methods of solution, based on the primal-dual algorithm, are discussed in W. Jewell's Massachusetts Institute of Technology doctoral dissertation, 1958.

Lastly, mention may be made of the problem of determining minimal cost augmentations to be made to a given system to provide a satisfactory grade of service in view of anticipated increases in future demands for service. These are given a linear programming formulation in [20]. Much work remains to be done, however, in order to find efficient methods of solution.

REFERENCES

1. R. Bellman, *Dynamic programming*, Princeton, Princeton University Press, 1957.
2. ———, *On a routing problem*, Quart. Appl. Math. vol. 16 (1958) pp. 87–90.
3. ———, *Notes on the theory of dynamic programming—transportation models*, Management Sci. vol. 4 (1958) pp. 191–195.
4. ———, *Combinatorial processes and dynamic programming*, see page 217 of this volume.
5. R. Bellman and R. Kalaba, *On k th best policies*, The RAND Corp., P-1417, 1958.
6. A. Boldyreff, *Determination of the maximal steady state flow of traffic through a railroad network*, Operations Res. vol. 3 (1955) pp. 443–466.
7. E. Brockmeyer, H. Halstrøm, and A. Jensen, *The life and works of A. K. Erlang*, Copenhagen, 1948.
8. A. Charnes and W. Cooper, *Extremal principles for simulating traffic flow in a network*, Proc. Nat. Acad. Sci. U.S.A. vol. 44 (1958) pp. 201–204.
9. C. Clos, *A study of non-blocking switching networks*, Bell System Tech. J. vol. 32 (1953) pp. 406–424.
10. G. B. Dantzig, *Discrete-variable extremum problems*, Operations Res. vol. 5 (1957) pp. 266–277.
11. G. B. Dantzig, A. Orden and P. Wolfe, *The generalized simplex method for minimizing a linear form under linear inequality restraints*, Pacific J. Math. vol. 5 (1955) pp. 183–195.
12. R. Dorfman, P. Samuelson and R. Solow, *Linear programming and economic analysis*, New York, McGraw-Hill Book Company, Inc., 1958.

13. P. Elias, A. Feinstein and C. Shannon, *A note on the maximum flow through a network*, IRE Transactions on Information Theory vol. IT-2 (1956) pp. 117-119.
14. L. R. Ford, Jr., *Network flow theory*, The RAND Corporation, P-923, 1956.
15. L. Ford, Jr. and D. Fulkerson, *A simple algorithm for finding maximal network flows and an application to the Hitchcock problem*, Canad. J. Math. vol. 9 (1957) pp. 210-218.
16. ———, *A suggested computation for maximal multi-commodity network flows*, The RAND Corporation, P-1114, 1958.
17. G. Frost, W. Keister and A. Ritchie, *A throwdown machine for telephone traffic studies*, Bell System Tech. J. vol. 32 (1953) pp. 292-359.
18. T. C. Fry, *Probability and its engineering uses*, New York, D. Van Nostrand Company, Inc., 1928.
19. F. Hohn, *Some mathematical aspects of switching*, Amer. Math. Monthly, vol. 62 (1955) pp. 75-90.
20. R. Kalaba and M. Juncosa, *Optimal design and utilization of communication networks*, Management Sci. vol. 3 (1956) pp. 33-44.
21. ———, *Optimal utilization and extension of interoffice trunking facilities*, Commun. and Electronics, January 1959, pp. 998-1003.
22. W. Keister, A. Ritchie and S. Washburn, *The design of switching networks*, New York, D. Van Nostrand Company, Inc., 1951.
23. D. König, *Theorie der endlichen und unendlichen Graphen*, New York, reprinted by Chelsea Publishing Company, 1950.
24. J. B. Kruskal, Jr., *On the shortest spanning subtree of a graph and the traveling salesman problem*, Proc. Amer. Math. Soc. vol. 7 (1956) pp. 48-50.
25. C. Y. Lee, *Analysis of switching networks*, Bell System Tech. J. vol. 34 (1955) pp. 1287-1315.
26. W. Miehle, *Link-length minimization in networks*, Operations Res. vol. 6 (1958) pp. 232-243.
27. E. C. Molina, *Application of the theory of probability to telephone trunking problems*, Bell System Tech. J. vol. 6 (1927) pp. 461-494.
28. W. Prager, *Problems of traffic and transportation*, Proceedings of Symposium on Operations Research in Business and Industry, Kansas City, 1954, pp. 105-113.
29. R. C. Prim, *Shortest connection networks and some generalizations*, Bell System Tech. J. vol. 36 (1957) pp. 1389-1401.
30. C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. vol. 27 (1948) pp. 379-423 and pp. 623-656.
31. C. Truitt, *Traffic engineering techniques for determining trunk requirements in alternate routing trunk networks*, Bell System Tech. J. vol. 33 (1954) pp. 277-302.
32. J. G. Wardrop, *Some theoretical aspects of road traffic research*, Proc. Inst. Civ. Engineers (London) vol. 1 (1952) pp. 325-378.
33. R. I. Wilkinson, *Theories for toll traffic engineering in the U.S.A.*, Bell System Tech. J. vol. 35 (1956) pp. 421-514.
34. G. J. Minty, *A comment on the shortest route problem*, Operations Res. vol. 5 (1957) p. 724.

THE RAND CORPORATION,
SANTA MONICA, CALIFORNIA

DIRECTED GRAPHS AND ASSEMBLY SCHEDULES

BY

J. D. FOULKES

1. Introduction. The problem considered here can be illustrated by a task which each of us faces every morning. In getting dressed you have to perform a sequence of operations in order to "assemble" a presentable man from the components of pants and shoes, toothpaste and shaving cream. Suppose a letter is assigned to each individual operation :

- A. Put on Shirt,
- B. Brush Teeth, etc.

A morning schedule can be represented by a permutation of the n letters ABC etc. Some of the operations must be performed in a mandatory sequence, e.g., the operation "Put on Socks" must precede the operation "Put on Shoes." Between others a preferential sequence can be guessed, e.g., "Brush Teeth" should precede "Put on Tie." Some pairs will not be related in any way (socks and shorts) and in a fourth case we could have the mandatory requirement that two operations should never occupy adjacent positions in a schedule. These conditions can be conveniently represented on a graph, Figure 1a. Each node represents an operation, and the branches indicate the pair-wise relationships between operations. The directed branch A to B indicates that A should precede B ; the nondirected branch AE indicates a "don't care" condition and finally the lack of a branch linking E and B shows that these two operations should not occupy adjacent positions in a schedule. Figure 1b shows a tabular method of organizing the same information. This is a symmetrical matrix in which a 1 indicates that a directional branch goes from the element in that row to the element in the column (see the indicated relations between A and B , and A and D). "Don't care," conditions are indicated by d , and the absence of a branch by the letter p .

It will be assumed that each schedule or permutation has associated with it a figure of merit, a numerical measure of the desirability of adopting that schedule. A further hypothesis is that a schedule which violates any one of the precedence relations is of necessity a poor one.

On graphs of the type shown in Figure 1a a schedule is represented by a non-branching tree which threads every node once and once only. A unidirectional tree of this type will be called a path, e.g., $DEABC$ in Figure 1a. The second hypothesis above assumes that the schedules which are worth examining in detail are represented by paths. Mathematically, the problem of interest is that of enumerating the number of paths in a directed graph.

2. Equivalence classes. The method of solution suggested here involves the repeated application of one basic concept, the partition of the nodes of a graph into a set of equivalence classes. Two nodes x and y will be said to belong to the same equivalence class if it is possible to trace a directed route from x to y and back again. In Figure 1a, A , B , and C belong to the same

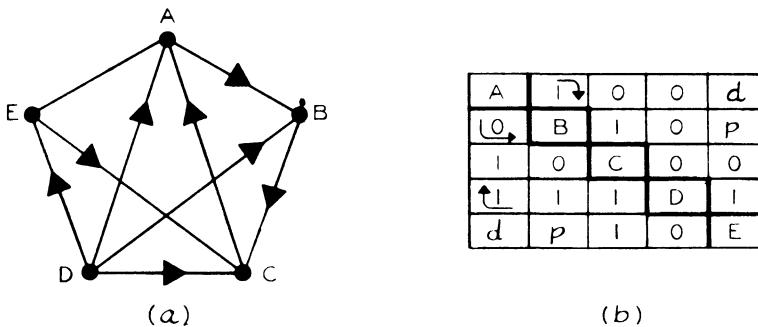


FIGURE 1

equivalence class. In order to give a simple explanation of the utility of this concept, it will be expedient to confine attention for the time being to graphs which are fully directed, i.e., graphs in which every pair of nodes are connected by a directed branch.¹ In these cases the graph can be broken down into a simply ordered set of equivalence classes, Figure 2a. A theorem in the appendix proves that all the members of a class are linked by at least one circuit. If the number of paths threading the nodes of a single equivalence Class 1 is p_1 , then the number of paths in the entire graph is $p_1 \times p_2 \times p_3 \dots$ etc. To find the number of paths threading a single class, consider the number of paths which start at a particular node x say. Remove x from the class and partition the remaining nodes into equivalence classes, Figure 2b. Suppose that the number of paths threading each of these sub classes is q_1, q_2 etc., and that the number of directed branches going from x to Class 1 is q_x , then the number of paths starting at x is $q_x \times q_1 \times q_2$ etc. A repeated application of this process breaks the graph into sets of ordered nodes.

3. Example I. This process can be illustrated by considering the example shown in Figure 3a. This table displays the interrelationships between the fourteen nodes numbered 1, 2, 3 . . . 14. A glance at the table indicates that element 10 precedes, and element 7 succeeds all other elements. Eliminating

¹ L. Redei has proved that at least one path through such a graph always exists: *Ein kombinatorischer Satz*, Acta Sci. Math. Szeged (1934–1935).

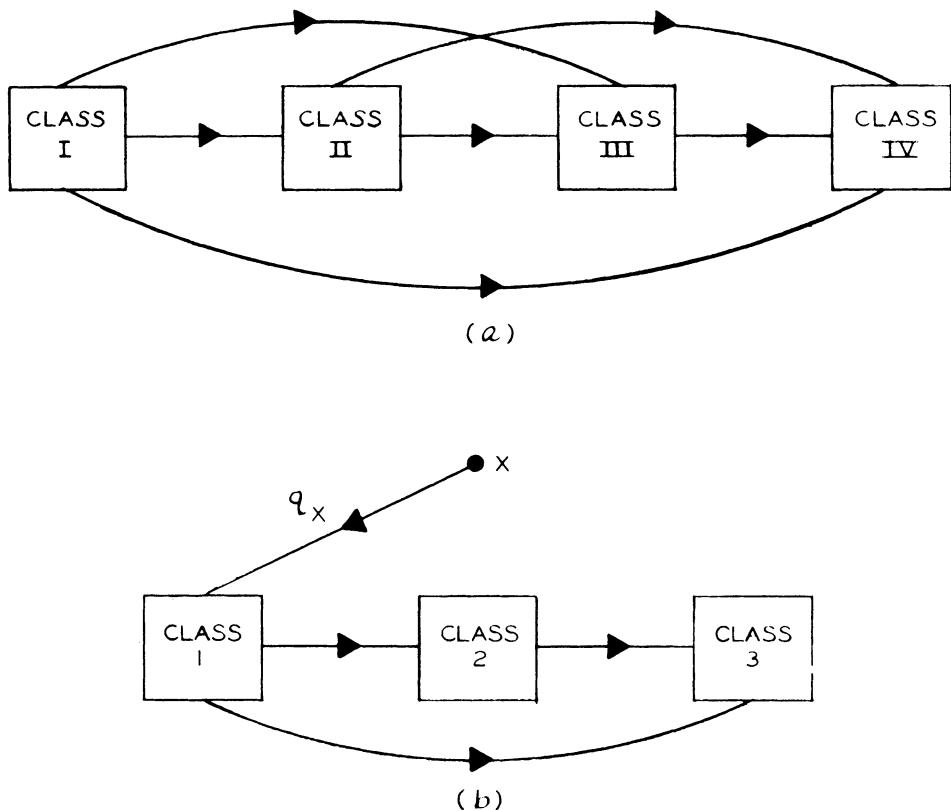


FIGURE 2

these two, the matrix of Figure 3b is obtained where the entries + should be read as 0. If this matrix is squared, the result will indicate which elements are interconnected by paths of length one or two.² For example, element 5 cannot be reached in one step from element 1, the entry in row 1, column 5 is 0. However, in multiplying row 1 by column 5, the ones in the row and column 6 coincide indicating that the path 1 to 6 to 5 exists. Thus the entry in row 1 column 5 of the squared matrix will be a 1. The matrix of Figure 3b is in fact the squared matrix where the 0's which have changed into 1's are indicated by +'s. Elements in the row and column of a + sign, therefore, belong to the same equivalence class. Collecting these together

² This is similar to the technique used to find impedance relations between nodes of a contact switching graph. See F. E. Hohn and L. R. Schissler, *Boolean matrices in the design of combinatorial switching circuits*, Bell System Tech. J. vol. 34 (1955).

1															
0	2														
0	0	3													
0	1	0	4												
1	1	1	1	5											
0	1	1	1	1	6										
0	0	0	0	0	0	7									
1	1	1	1	1	1	1	8								
1	1	1	1	1	1	1	1	9							
1	1	1	1	1	1	1	1	10							
1	1	1	1	1	1	1	0	1	0	11					
0	1	0	0	0	0	1	0	0	0	0	12				
1	1	1	1	1	1	1	0	1	0	1	13				
0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	14

(A)

1	1	1	1	+	1		0	0		0	1	0	1		
0	2	1	+	0	0		0	0		0	+	0	0		
0	+	3	1	0	0		0	0		0	1	0	+		
0	1	+	4	0	0		0	0		0	1	0	+		
1	1	1	1	5	+		0	0		0	1	0	1		
+	1	1	1	1	1	6	0	0		0	1	0	1		
						7									
1	1	1	1	1	1	1	8	+		1	1	1	1		
1	1	1	1	1	1	1	9			+	1	+	1		
							10								
1	1	1	1	1	1	1	+	1		11	1	0	1		
0	1	+	+	0	0		0	0		0	12	0	1		
1	1	1	1	1	1	1	+	1		1	1	13	1		
0	1	1	1	1	0	0	0	0		0	+	0	14		

(B)

FIGURE 3

the matrix of Figure 3c is obtained. Squaring this leads to no further reduction in the number of zeros and hence the partitioning process is at an end. Figure 4 displays the equivalence classes α , β , γ , etc., into which the graph has broken and the ordering which exists between these classes. The

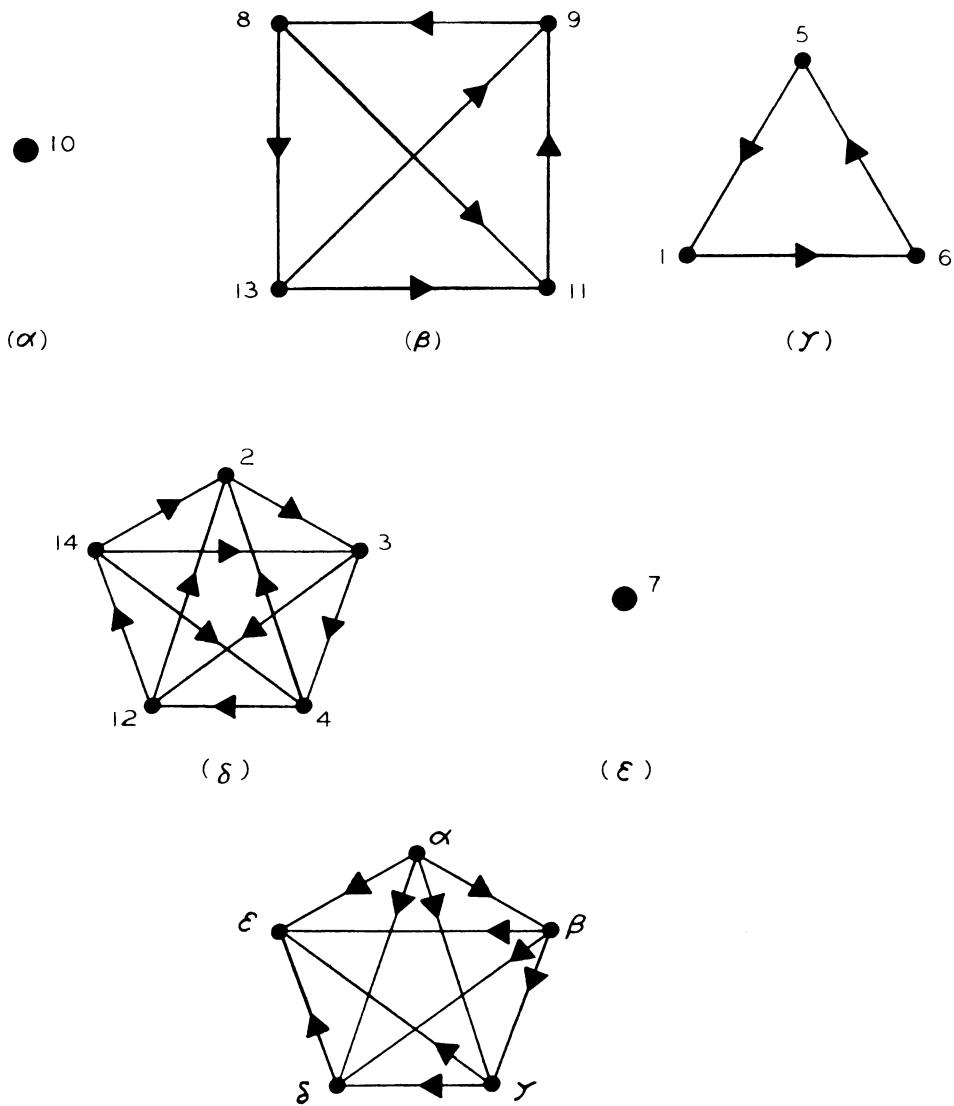


FIGURE 4

procedure for counting the paths threading each class is illustrated in Figure 5 which uses the class $(\delta) = [2, 3, 4, 12, 14]$ as an example. The top figure considers the paths which start at 3. If 3 is removed, the remaining nodes break into the ordered sub classes $[4, 12, 14]$ and $[2]$. Two directed branches

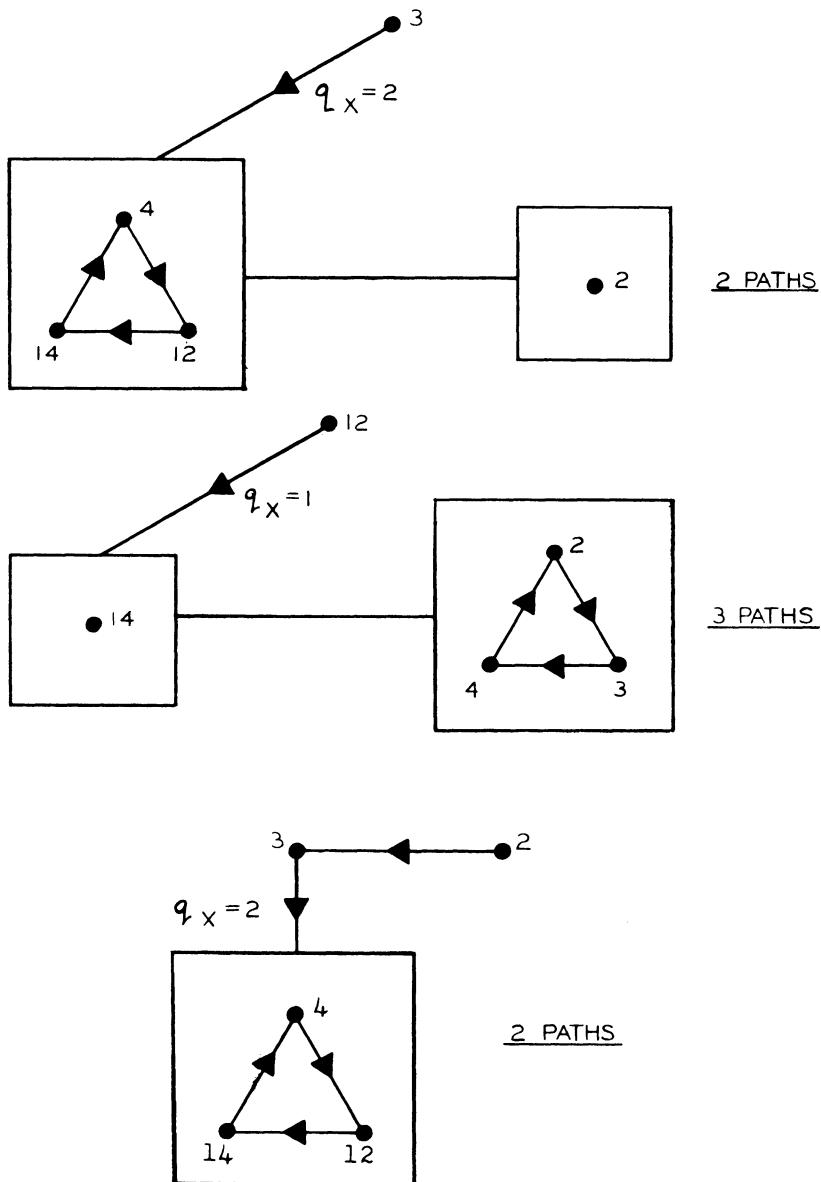


FIGURE 5

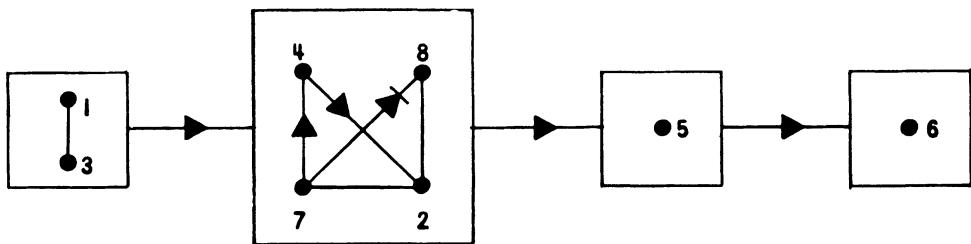
go from 3 to [4, 12, 14] so that $q_x = 2$ and hence two paths start from 3. The breakdown for paths starting from 12 and 2 are also given in the figure.

One finds that five paths thread the class [8, 9, 11, 13], three thread [1, 5, 6] and ten thread the class [2, 3, 4, 12, 14]. The number of paths threading the entire graph therefore is $1 \times 5 \times 3 \times 10 \times 1 = 150$. This number should be compared with $14!$ which is of the order 8×10^{10} .

4. Example II. The manner in which the above techniques have to be modified to cope with graphs involving mandatory sequences, non-directed

I	SHORTS							
0x	2	PANTS						
d	1	3	UNDERSHIRT					
0	1	0x	4	SHIRT				
0	0	0x	0x	5	TIE			
0	0	0x	0x	0	6	COAT		
0	d	0	1	1	1	7	SOCKS	
0	d	0	p	1	1	0x	8	SHOES

(A)



(B)

FIGURE 6

branches, and missing branches can be explained by returning to the example of the morning schedule. Figure 6a shows the relationships to be satisfied between eight operations. Mandatory sequences, e.g., "Socks" before "Shoes" have been starred. The other preferential sequences indicated,

e.g., "Tie" before "Coat" are spot judgments based on common sense. Three "don't care" conditions have been included as well as one arbitrary p condition. The d 's can be eliminated from the table by immediately placing those elements which lie in the same row and column as a d in the same equivalence class. Treating P as a 0 the matrix can be successively squared to yield the partitioning shown in Figure 6b. In counting the paths threading the class [2, 4, 7, 8] care must be taken not to violate the mandatory requirement that 7 should precede 8 (indicated by a crossed arrow in the figure). The schedule 8, 2, 7, 4 for example calls for "Shoes" to be donned before "Socks." Notice that two elements related in a mandatory manner never occur in distinct equivalence classes which place them in an incorrect sequence. The worst that can happen is for them to belong to the same class. In this event schedules which violate their relationship must be discarded. In Figure 6b there are two paths in the first class and two in the second. There are therefore $2 \times 2 \times 1 \times 1 = 4$ paths threading the entire graph.

5. Conclusions. The techniques described above do not lend themselves readily to hand computation. A computer program capable of implementing them would not be hard to draw up, but the memory capacity called for would probably be large. The real utility of these ideas might lie in the theoretical background which they provide for the scheduler. The theory has an obvious relevance to processes involving a rough and ready pair-wise ordering of operations. Some of these problems are:

- (i) Assembly line problems.
- (ii) The problem of placing examination and other candidates in an order which maximally agrees with an ordering of subsets drawn up by several judges.
- (iii) The problem of analyzing a group of consumer preference tests which have been conducted on an AB basis.

In all these cases the concept of the equivalence classes presented above should prove helpful to the analyst.

Appendix.

THEOREM. *In a fully directed graph at least one circuit links all the nodes belonging to any one equivalence class.*

An inductive proof will be given. Assume that the theorem is true for n nodes and arrange these around a circle with the directed circuit proceeding in a clockwise direction. Place the $(n+1)$ th node in the center. If it belongs to the same equivalence class, there must be at least one directed branch proceeding from a node on the circle to the center. Call this node 1 and number the nodes 1, 2, 3, etc., proceeding in a clockwise direction. Again

there must be at least one directed branch going from the center to a node on the circle. Let the first branch of this type encountered in going clockwise around the circle be j . Then the circuit $1, 2, 3 \dots (j - 1), (n + 1), j, j + 1 \dots (n - 1), n, 1$ exists. Hence the theorem is true for $(n + 1)$ nodes. It is true for three nodes, and thus it is true for any n .

BELL TELEPHONE LABORATORIES, INC.,
MURRAY HILL, NEW JERSEY

This page intentionally left blank

A PROBLEM IN BINARY ENCODING

BY

E. N. GILBERT

1. Introduction. R. W. Hamming [4] defines *error-correcting encoding* in the following way. Let n and E be two given positive integers with $2E < n$. Then an error-correcting encoding is a list of n -tuples of binary digits such that every pair of n -tuples in the list differ in at least $2E + 1$ of the n places. Error-correcting encodings are used when information must be encoded into binary digits and transmitted through a medium or device which changes an occasional 1 to a 0 or a 0 to a 1. By restricting the n -tuples which can be sent to be those of an error-correcting encoding, the correct n -tuple can be reconstructed from the received n -tuple even though as many as E received digits are incorrect. For, only the correct n -tuple differs from the received n -tuple in E or fewer places.

In constructing an error-correcting encoding one hopes to make the number of n -tuples in the list as large as possible. For, the amount of information which an n -tuple can convey is an increasing function (logarithm) of the number of n -tuples in the list. Longest possible lists with $E = 1$ were constructed by Hamming. The more general problem has been considered by several authors [1; 2; 3; 5; 7; 9] but remains largely unsolved.

In many situations the transmission errors tend to occur in isolated clusters or *bursts* of several errors close together. Errors produced by radio static, sticking relay contacts, or switching transients are of this kind. One might use an error-correcting encoding (in Hamming's sense) to combat bursts. To do so would be wasteful because such an encoding is designed to correct many other kinds of errors (in which the incorrect digits are not grouped together in bursts) which are rare. In this paper we construct some special *burst-correcting encodings* for use in such situations.

2. Burst noise. Let $(x_0, x_1, \dots, x_{n-1})$ be a transmitted n -tuple of binary digits and suppose that the corresponding received n -tuple is (y_0, \dots, y_{n-1}) . The binary n -tuple (z_0, \dots, z_{n-1}) with

$$y_i \equiv x_i + z_i \pmod{2}$$

will be called the *noise n -tuple*. We wish to consider only errors which group together in bursts extending over at most E digits and such that each pair of bursts is separated by at least D correct digits. Thus we define a *burst sequence* to be any one of the 2^{E-1} binary sequences 1, 11, 101, ..., 11...101, 11...111, which start with 1, end with 1, and contain 1, 2, ..., or E digits. The noise n -tuples of interest here are those which can be decomposed into burst sequences interspersed with runs of D or more consecutive

0's. Runs of 0's of lengths $< D$ may occur at the beginning and at the end of the n -tuple. Such an n -tuple will be called a *burst n-tuple*.

We always assume $D \geq E - 1$. Under this assumption a given burst n -tuple can be decomposed into burst sequences and runs of 0's in just one way. Otherwise (say when $D = 2$ and $E = 4$) a single long burst sequence (1001) may be decomposed into two short burst sequences separated by a run of 0's.

We will require the number $F(n)$ of distinct burst n -tuples. Any burst n -tuple containing k burst sequences is of the form

$$R_0 B_1 R_1 B_2 R_2 \cdots B_k R_k$$

in which R_0, \dots, R_k are runs of 0's and B_1, \dots, B_k are burst sequences (assuming $k \geq 1$). The number of possible burst sequences of length e digits is the coefficient of x^e in the generating series

$$\begin{aligned} b(x) &= x + x^2 + 2x^3 + \cdots + 2^{E-2}x^E \\ &= x + x^2\{1 - (2x)^{E-1}\}(1 - 2x)^{-1}. \end{aligned}$$

In the special case $E = 1$, set $b(x) = x$. Each of R_1, \dots, R_{k-1} has the generating series

$$r(x) = x^D + x^{D+1} + \cdots = x^D(1 - x)^{-1}.$$

R_0 and R_k may be runs of any length, and so have generating series $(1 - x)^{-1}$. Then the number $F(n, k)$ of burst n -tuples containing k burst sequences must be the coefficient of x^n in the generating function

$$\begin{aligned} f_k(x) &= (1 - x)^{-2} r^{k-1}(x) b^k(x), && \text{if } k \geq 1 \\ &= (1 - x)^{-1}, && \text{if } k = 0. \end{aligned}$$

The generating function for $F(n)$ is then

$$(1) \quad \sum_{n=0}^{\infty} F(n)x^n = (1 - x)^{-1} + b(x)(1 - x)^{-2}\{1 - r(x)b(x)\}^{-1}.$$

For small values of n , $F(n)$ has the simple formulas

$$\begin{aligned} F(n) &= 2^n, && \text{if } 0 \leq n \leq E, \\ F(n) &= (n + 2 - E)2^{E-1}, && \text{if } E \leq n \leq D + 1, \\ F(n) &= 1 + (n - D - 2)2^{n-D-1} + (n + 2 - E)2^{E-1}, && \text{if } D + 2 \leq n \leq D + E. \end{aligned}$$

For $n \geq D + E$ the recurrence equation

$$F(n) = F(n - 1) + F(n - D - 1) + \sum_{e=2}^E 2^{e-2}F(n - D - e)$$

holds. If $E = 1$ the summation over e is to be omitted. A second recurrence equation, valid for $n \geq D + E + 1$, is

$$\begin{aligned} F(n) = 3F(n-1) - 2F(n-2) + F(n-D-1) - F(n-D-2) \\ - 2^{E-1}F(n-D-E-1). \end{aligned}$$

An asymptotic formula for $F(n)$ for large n may be obtained by examining the poles of the generating function (1). These occur at $x = 1$ and at roots of $r(x)b(x) = 1$. Since $r(0)b(0) = 0$ and $r(1)b(1) = \infty$ there is a positive real root, say $x = T^{-1} < 1$. Since $r(x)b(x)$ has positive coefficients, it is monotone and T^{-1} is the only positive real root. Moreover T^{-1} is the pole of (1) which has the smallest modulus. Then $F(n) \sim \text{const. } T^n$ where the constant is $-T$ times the residue of $b(x)(1-x)^{-2}\{1-r(x)b(x)\}^{-1}$ at the pole T^{-1} . For our purposes it suffices to note that $\log F(n)$ is asymptotic to $n \log T$ for large n .

3. Encodings. If $a = (a_0, \dots, a_{n-1})$ and $b = (b_0, \dots, b_{n-1})$ are any two n -tuples, we write $a + b$ for the n -tuple $(a_0 + b_0, \dots, a_{n-1} + b_{n-1})$ in which the coordinates of a and b are added modulo 2. By a *burst-correcting encoding* we will mean a list of n -tuples a, b, \dots, w such that, for every binary n -tuple x , at most one of the n -tuples $a + x, b + x, \dots, w + x$ is a burst n -tuple. If transmitted n -tuples are always selected from such a list, then a received n -tuple x may be correctly interpreted as that n -tuple j of the list for which $j + x$ is a burst n -tuple.

THEOREM I. *A necessary and sufficient condition for a list a, b, \dots, w of n -tuples to be a burst-correcting encoding is*

(c) *no pair i, j of distinct n -tuples in the list have a sum $i + j$ which equals a sum of two burst n -tuples.*

To prove that (c) is necessary, suppose that (c) fails, i.e., suppose the list contains i and j such that $i + j = y + z$ where y and z are burst n -tuples. Then, for the n -tuple $x = i + y = j + z$, both $i + x$ and $j + x$ are burst n -tuples. Conversely, if the list is not a burst-correcting encoding, it contains n -tuples i, j such that, for some x , $i + x$ and $j + x$ are burst n -tuples. Then, since $i + j = (i + x) + (j + x)$, (c) fails.

Our problem will be to construct burst-correcting encodings which contain large numbers of n -tuples. First some simple estimates will be given for the largest number of n -tuples which a burst-correcting encoding can have.

THEOREM II. *No burst-correcting encoding can contain more than $[2^n/F(n)]$ n -tuples.*

Proof. Suppose that a burst-correcting encoding contains K n -tuples a, b, \dots, w . For each n -tuple i in the list, let $S(i)$ be the set of all $F(n)$ n -tuples of the form $i + z$ where z is a burst n -tuple. By hypothesis $S(a), S(b), \dots, S(w)$ are disjoint sets. Then the number of n -tuples in their union is $KF(n) \leq 2^n$.

A burst-correcting encoding will be called *full* if no additional n -tuple can be adjoined to it to make an enlarged list which is also a burst-correcting encoding. A full encoding may be constructed by selecting n -tuples one at a time. Each new n -tuple may be any one of the n -tuples which, together with the n -tuples already selected, forms a burst-correcting encoding. Ultimately there are no n -tuples left from which to choose; then the list is full.

THEOREM III. *Let $G(n)$ be the number of distinct n -tuples which can be expressed as sums of two burst n -tuples. Any full encoding contains at least $2^n/G(n)$ n -tuples (and hence at least $2^{n+1}/F(n)\{F(n) + 1\}$ n -tuples).*

Proof. Let an encoding contain K n -tuples a, b, \dots, w . For each i in the list construct a set $S^*(i)$ of all $G(n)$ n -tuples of the form $i + y + z$ where y and z are burst n -tuples. If there is an n -tuple x which is in none of $S^*(a), S^*(b), \dots, S^*(w)$ then, by Theorem I, x may be adjoined to the list. Hence if the original list is full, each n -tuple x must belong to at least one of $S^*(a), \dots, S^*(w)$ and $KG(n) \geq 2^n$. The parenthetical remark follows because $G(n)$ is at least as large as the number of combinations of burst n -tuples taken two at a time allowing repetitions.

The number K of n -tuples which can belong to a burst-correcting encoding is bounded only very roughly by Theorems II and III. Fortunately, our main interest is in $\log K$ rather than K itself. The redundancy (Hamming [4] and Shannon [8]) of the encoding is $R = 1 - n^{-1} \log_2 K$. R measures the amount by which the rate of transmitting information has been decreased in order to obtain error-free transmission. Letting $n \rightarrow \infty$, Theorems II and III give $\log_2 T$ and $2 \log_2 T$ as bounds on the redundancies of encodings with large n .

4. Explicit encodings. We now exhibit some encodings which have structures systematic enough to permit machines to perform the encoding and decoding computations without great difficulty. In every case we will have $n \leq D + E$. Then the burst n -tuples to be corrected are either of the form $R_0B_1R_1$ or $R_0B_1R_1B_2R_2$, where B_1, B_2 are burst sequences and R_0, R_1, R_2 are runs of 0's. However, in the double-burst case $R_0B_1R_1B_2R_2$, the inequality $n \leq D + E$ forces R_0, B_1, B_2, R_2 to contain a total number of digits $\leq E$.

For the isolated case $E = 2, D = 5, n = 7$ one may construct a list of 8 7-tuples (x_0, \dots, x_6) . Digits x_0, x_1, x_2 may form any of the 8 possible triples 000, 001, ..., 111. The other digits are computed from the recurrence $x_i \equiv x_{i-1} + x_{i-3} \pmod{2}$. This is a best-possible encoding in the sense that the bound given in Theorem II is achieved.

Hamming (unpublished) has suggested an encoding when n is a multiple of E , say $n = qE$. Given an n -tuple $x = (x_0, \dots, x_{n-1})$ let E q -tuples X^i , $i = 0, \dots, E - 1$ be defined; the coordinates of X^i will be those coordinates

of x which have subscripts $\equiv i \pmod{E}$. The encoding consists of all n -tuples x for which each of X^0, X^1, \dots, X^{E-1} belongs to an error-correcting encoding which can correct a single error in q digits. This encoding corrects many other kinds of errors besides bursts. As long as no two erroneous digits have subscripts congruent mod E , the errors may be corrected. Accordingly the number of n -tuples in the encoding is much less than $2^n/F(n)$.

Some slightly better encodings are obtainable by the following construction. Pick a set of integers $M_1 < M_2 < \dots < M_r$. M_i will be called the *moduli* of the encoding. Take n to be the least common multiple of the moduli. For the n -tuple (x_0, \dots, x_{n-1}) we define $M_1 + \dots + M_r$ linear forms $S(a, M_i)$, ($i = 1, \dots, r$, and $a = 0, 1, \dots, M_i - 1$). $S(a, M_i)$ is the sum of those coordinates x_j having subscripts $j \equiv a \pmod{M_i}$.

THEOREM IV. *The list of all n -tuples, for which all $M_1 + \dots + M_r$ of the linear forms $S(a, M_i)$ are $0 \pmod{2}$, is a burst-correcting encoding with $E = [(M_1 + 1)/2]$ and $D = n - E$.*

The encoding was suggested by a somewhat similar one (unpublished) invented by W. D. Lewis. Lewis' encoding has two moduli and corrects single bursts of length 2.

In discussing this encoding it is convenient to arrange the digits of an n -tuple cyclically with z_0 following z_{n-1} . Then (since $n = E + D$) each burst n -tuple, considered as a cyclic arrangement, contains only a single burst-sequence and a run of at least $n - E$ consecutive 0's. For, even in the double burst case, $B_2 R_2 R_0 B_1$ appears in the cyclic arrangement as a single burst. In these encodings $n - E > E$. Then a cyclic arrangement of a burst n -tuple can contain only one run of at least $n - E$ consecutive 0's. The 1 which immediately follows this unique run of 0's will be called the *first 1*. For n -tuples $R_0 B_1 R_1$, the first 1 appears at the beginning of B_1 . For $R_0 B_1 R_1 B_2 R_2$ it appears at the beginning of B_2 .

Suppose, in using this encoding, that an n -tuple $y = (y_0, \dots, y_{n-1})$ has been received. We must now show how to compute the correct transmitted n -tuple $x = (x_0, \dots, x_{n-1})$. For $i = 1, \dots, r$ and $a = 0, 1, \dots, M_i - 1$ let $R(a, M_i)$ be the sum of all y_j having $j \equiv a \pmod{M_i}$. If a burst of length $\leq M_1$ has occurred, at most one digit y_j in a sum $R(a, M_i)$ is incorrect. Then if $R(a, M_i) \equiv 1 \pmod{2}$ just one of the n/M_i digits y_j with $j \equiv a \pmod{M_i}$ is incorrect.

For a given i , suppose that $R(a, M_i) \equiv 1 \pmod{2}$ holds for the values a_1, \dots, a_h of a . Then there are h incorrect digits. If z_J is the first 1 in the noise n -tuple ("first" in the special sense defined earlier) then the number $A_i \equiv J \pmod{M_i}$ will be one of a_1, \dots, a_h . To identify A_i , let the numbers $R(0, M_i), R(1, M_i), \dots, R(M_i - 1, M_i)$ be arranged cyclically. None of digits $y_J, y_{J+1}, \dots, y_{J+E-1}$ (which are the only possible erroneous ones) appear in $R(A_i - 1, M_i), R(A_i - 2, M_i), \dots, R(A_i - M_i + E, M_i)$. Then

in the cyclic arrangement of $R(a, M_i)$, $R(A_i, M_i)$ will follow a run of at least $M_i - E$ 0's. There cannot be two such runs of 0's. For, two runs of $M_i - E$ 0's would require $2(M_i - E) + 2 = 2M_i + 2 - 2[(M_1 + 1)/2] > M_i$ terms in the $R(a, M_i)$ sequence. Thus A_i may be selected from a_1, \dots, a_n by finding the longest run of consecutive 0's in the cyclic arrangement of the $R(a, M_i)$.

Find A_i for $i = 1, 2, \dots, r$. Then the position J of the first error satisfies the simultaneous congruences

$$(2) \quad J \equiv A_i \pmod{M_i} \quad i = 1, \dots, r.$$

This congruence is in the form to which the Chinese remainder theorem T. Nagell [6, Theorem 40] applies. A solution of (2) is unique modulo n . Then J may be computed by solving (2). Once J is known all the other errors are easily found. For if $R(A_1 + b, M_1) \equiv 1 \pmod{2}$ then y_{J+b} is an erroneous digit.

The number of n -tuples in the encoding is 2^{n-s} where s is the rank of the system of $M_1 + \dots + M_r$ congruences $S(a, M_i) \equiv 0 \pmod{2}$. Then s/n is the redundancy of the encoding. Since, for $i = 1, \dots, r$,

$$\sum_a S(a, M_i) = x_1 + \dots + x_n$$

the rank s satisfies $s \leq M_1 + \dots + M_r - r + 1$. Depending on the choice of moduli, there may be other linear dependencies.

If M_1, \dots, M_r are relatively prime then one finds $s = M_1 + \dots + M_r - r + 1$.

If the moduli are relatively prime except for one pair M_i, M_j which has greatest common divisor D , then $s = M_1 + \dots + M_r - D - r + 1$.

If the moduli have a greatest common divisor D the encoding has the same redundancy as the one with moduli M_i/D .

Since $E = [(M_1 + 1)/2]$ it is desirable, for a given n , to pick moduli which are as nearly equal as possible.

An encoding for a typical case $D = 27$, $E = 3$, $n = 30$ is obtained by taking moduli 5, 6. The redundancy is 1/3. Theorem II provides a lower bound $n^{-1} \log_2 F(n) = .23$ on the redundancy for $n = 30$ and a limiting lower bound $\log_2 T = .18$ for encodings with the same D and E and with large n . Following Hamming's suggestion one obtains an encoding with $D = 27$, $E = 3$, $n = 30$, and redundancy 2/5. Since $2/5 > 2 \log_2 T$, all full encodings with sufficiently large n will be less redundant than Hamming's. As a rule our encodings are less redundant than comparable Hamming encodings when the moduli are chosen nearly equal. On the other hand, take moduli 8, 9, 10 and hence $E = 4$, $D = 356$, $n = 360$. Our encoding then has redundancy greater than $2n^{-1} \log_2 F(n)$, and so is not full. The Hamming encoding with these parameters is still more redundant.

In addition to the practical combinatorial problem of constructing simple

explicit burst-correcting encodings which have low redundancy, there are some interesting theoretical questions. What is the maximum number of n -tuples possible in a burst-correcting encoding? Do there exist burst-correcting encodings which have redundancies arbitrarily close to $\log_2 T$? For the latter problem one can show that $1 - \log_2 T$ is the smallest capacity (in the sense of Shannon [8]) attained by any binary channel which has burst noise. This result is suggestive but gives no new information about the redundancies which are achievable when all errors must be corrected.

REFERENCES

1. P. Elias, *Error-free coding*, I.R.E. Transactions of the Professional Group on Information Theory, PGIT-4 (1954) pp. 29-36.
2. E. N. Gilbert, *A comparison of signalling alphabets*, Bell System Tech. J. vol. 31 (1952) pp. 504-522.
3. M. J. E. Golay, *Binary coding*, I.R.E. Transactions of the Professional Group on Information Theory, PGIT-4 (1954) pp. 23-28.
4. R. W. Hamming, *Error detecting and error correcting codes*, Bell System Tech. J. vol. 29 (1950) pp. 147-160.
5. S. P. Lloyd, *Binary block coding*, Bell System Tech. J. vol. 36 (1957) pp. 517-535.
6. T. Nagell, *Introduction to number theory*, New York, Wiley, 1951.
7. M. Plotkin, Research Division Report 51-20, Moore School of E.E., University of Pennsylvania, 1951.
8. C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. (1948) pp. 379-423; 623-656.
9. D. Slepian, *A class of binary signalling alphabets*, Bell System Tech. J. vol. 35 (1956) pp. 203-234.

BELL TELEPHONE LABORATORIES, INC.,
MURRAY HILL, NEW JERSEY

This page intentionally left blank

AN ALTERNATIVE PROOF OF A THEOREM OF KÖNIG AS AN ALGORITHM FOR THE HITCHCOCK DISTRIBUTION PROBLEM¹

BY

MERRILL M. FLOOD

Introduction. Several forms of the Hitchcock distribution problem [1; 2] were discussed in a previous paper by the present author [3], including one often called the "assignment problem" and another often called the "transportation problem". The essential equivalence of these various algebraic statements of the Hitchcock distribution problem was noted, and the relationship between the Hitchcock problem and the one now commonly known as the "traveling salesman problem" was discussed. It was also shown how a method due to H. W. Kuhn, which was in turn based upon a graph-theoretic theorem of König, could be used to provide computational algorithms adequate for the Hitchcock distribution problem and helpful in attacking the traveling salesman problem.

The Hitchcock distribution problem, in assignment form, requires an algorithm to determine a permutation of the rows of a square matrix that minimizes its trace. The present paper provides such an algorithm, and includes a simple numerical example to illustrate its use.

In presenting the algorithm, a constructive proof is first given for König's Theorem [4]. The steps of this proof, as in papers by Kuhn [5] and others [6], provide the mathematical description and justification for the corresponding steps of the algorithm. For numerical analysis purposes, contrary to the usual case in pure mathematics, it is often the proof of the theorem that is important rather than the theorem itself. An elegant constructive proof of a theorem may yield only a very inefficient computational process as compared with a mathematically less elegant proof. Each such proof, as in the present case, usually leads to a wide variety of detailed computational procedures. No attempt is made here to set forth specific and efficient computational routines, but this will be done elsewhere on the basis of the proof presented here.

¹ *Acknowledgments.* This work was done largely while the author was at Columbia University. It was supported in part by the University of Michigan, under Signal Corps Contract #DA-36-039SC64627, and by Princeton University, under Office of Naval Research Contract #N60NR-27011. Grateful acknowledgment is made to A. W. Tucker, and to his associates at Princeton University, for several stimulating discussions of the Hungarian Method.

Proof of König's Theorem.

KÖNIG'S THEOREM. *If A is a square array of two kinds of marks, say zeros and plusses, and if*

- (a) *x is the maximum number of zeros that can be found in the array such that no two of them are in the same line, and*
- (b) *y is the minimum number of lines that can be found such that every zero of the array is contained in one of them,*

then $x = y$. (A line is either a row or a column.)

Proof. It will be convenient to introduce some special notation for certain types of possible subarrays, as follows :

- denotes a subarray each of whose elements may be zero or plus,
- 0 denotes a square subarray all of whose main diagonal elements are zeros,
- r denotes a subarray with at least one zero in each row,
- c denotes a subarray with at least one zero in each column,
- + denotes a subarray with no zeros,
- rc denotes a subarray with at least one zero in each line.

It will be shown how to make a sequence of permutations on the rows and columns of A that will bring it into "standard form", and in such a way that the first z elements on the main diagonal of the permuted array are all zeros while a set of z designated lines contains all of the zeros of the array. Since the validity of König's Theorem is obviously unaffected by permuting rows and columns of the array A , the reduction to a standard form of this kind constitutes the required constructive proof.

STEP 1. Permute so that A takes the form:

$$A \longrightarrow \begin{array}{|c|c|} \hline 0 & \\ \hline & \\ \hline + & \\ \hline \end{array}$$

This is easily done by first permuting so that any zero moves to the first diagonal position, next permuting so that any remaining zero not in the new first row or column moves to the second diagonal position, next permuting so that any remaining zero not in the new first two rows and columns moves to the third diagonal position, and so forth until there are no zeros remaining. This terminates the process unless there is at least one zero in each of A_{12} and A_{21} , for otherwise either the rows or the columns through A_{11} contain

all the zeros of A . The process is also terminated if every diagonal element is zero.

STEP 2.

$$A = \begin{array}{|c|c|} \hline 0 & A_{12} \\ \hline \hline A_{21} & + \\ \hline \end{array} \longrightarrow \begin{array}{|c|c|c|} \hline 0 & c & + \\ \hline r & + & + \\ \hline + & + & + \\ \hline \end{array} = B.$$

The columns of A_{12} are permuted, and the rows of A_{21} are permuted, thus leaving A_{11} unchanged. The columns of A_{12} containing zeros become the columns of B_{12} , and the columns of A_{12} having no zeros become the columns of B_{13} ; similarly for A_{21} .

STEP 3.

$$A = \begin{array}{|c|c|c|} \hline 0 & c & + \\ \hline r & + & + \\ \hline + & + & + \\ \hline \end{array} \longrightarrow \begin{array}{|c|c|c|c|} \hline 0 & & + & + \\ \hline & 0 & rc & + \\ \hline r & & + & + \\ \hline + & + & + & + \\ \hline \end{array} = B.$$

If each row of A_{12} contains a zero then proceed in a manner exactly similar to that for Step 4, following, to replace some zero in A_{11} by two drawn from A_{12} and A_{21} , then returning to Step 2. Otherwise, the rows of A_{12} are permuted so that no row of B_{13} contains a zero but each row of B_{23} , and also each column of B_{23} , of course, contains at least one zero. In order to retain the same zeros on the diagonals of B_{11} and B_{22} as were in the diagonal of A_{11} , the columns of A_{11} are subjected to exactly the same permutation as were the rows of A_{11} . As a result of this permutation, there is at least one zero in each row of the subarray $[B_{31} | B_{32}]$, since this is simply A_{21} with its columns permuted.

STEP 4.

$$A = \begin{array}{|c|c|c|c|} \hline j & k & & \\ \hline 0 & & + & + \\ \hline & 0 & rc & + \\ \hline r & & + & + \\ \hline + & + & + & + \\ \hline \end{array} \xrightarrow{j \quad i} \begin{array}{|c|c|c|c|} \hline 0 & & + & + \\ \hline & 0 & rc & + \\ \hline r & + & + & + \\ \hline + & + & + & + \\ \hline \end{array} = B.$$

An actual permutation need be made in this step only if A_{32} contains at least one zero, say in position (ij) . Now, since A_{23} has at least one zero in each

row, it has a zero in the j th row and, say, k th column. If columns j and k are now interchanged, obtaining a new array A^* partitioned the same as A , the zeros are retained in all the diagonal positions of A_{11}^* and A_{22}^* , but at least one new zero is introduced into A_{33}^* at the position (ik) . It is possible, therefore, to further permute the rows and columns of A_{33}^* so as to move this zero from the (ik) position of A^* to take the first diagonal position of A_{33}^* . Consequently, there is now an array with one more zero on the diagonal than at the end of Step 1, and Steps 2 and 3 must be repeated until finally the form shown for Step 4 is reached in which B_{32} has no zeros.

STEP 5.

$$A = \begin{array}{|c|c|c|c|} \hline 0 & & + & + \\ \hline & 0 & rc & + \\ \hline r & + & + & + \\ \hline + & + & + & + \\ \hline \end{array} \rightarrow \begin{array}{|c|c|c|c|} \hline 0 & & + & + \\ \hline & 0 & + & + \\ \hline & & 0 & rc & + \\ \hline rc & + & + & + & + \\ \hline + & + & + & + & + \\ \hline \end{array} = B.$$

The process terminates if A_{12} contains no zero, for then the columns through A_{11} , together with the rows through A_{22} , contain all the zeros of A . If A_{12} does contain at least one zero, and if each column of A_{31} contains a zero, then proceed in a manner exactly similar to that for Step 6, following, to replace some pair of zeros in A_{11} and A_{22} by three drawn from A_{31} , A_{12} , and A_{23} ; then return to Step 2. Otherwise the columns of A_{31} are permuted so as to move all its zeros into B_{41} . Then the rows of A_{11} are permuted in the same way as were its columns, so as to hold the original zeros on the diagonals of A_{11} rearranged for the diagonals of B_{11} and B_{22} .

STEP 6.

$$A = \begin{array}{|c|c|c|c|c|} \hline i & j & k & & \\ \hline 0 & & & + & + \\ \hline & 0 & & + & + \\ \hline & & 0 & rc & + \\ \hline rc & + & + & + & + \\ \hline + & + & + & + & + \\ \hline \end{array} \xrightarrow{j \rightarrow l} \begin{array}{|c|c|c|c|c|} \hline i & & + & + & + \\ \hline & 0 & & + & + \\ \hline & & 0 & rc & + \\ \hline rc & + & + & + & + \\ \hline + & + & + & + & + \\ \hline \end{array} = B.$$

An actual permutation need be made in this step only if A_{13} contains at least one zero, say in position (ij) . If so, then in a manner exactly similar to Step 4, it is possible to replace the two zeros in positions (ii) and (jj) on the original diagonal by three new zeros: the one at (ij) , one in row j of A_{34} ,

and one in column i of A_{41} . As the number of zeros on the diagonal is increased by at least one by this permutation, over the number there at the end of Step 4, Steps 2 through 5 are repeated until finally the form shown for Step 6 is reached in which B_{13} has no zeros.

STEP 7.

$$A = \begin{array}{|c|c|c|c|c|} \hline 0 & & + & + & + \\ \hline & 0 & & + & + \\ \hline & & 0 & rc & + \\ \hline rc & + & + & + & + \\ \hline + & + & + & + & + \\ \hline \end{array} \longrightarrow \begin{array}{|c|c|c|c|c|c|} \hline 0 & c & + & + & + & + \\ \hline & 0 & & + & + & + \\ \hline & & 0 & & + & + \\ \hline & & & 0 & rc & + \\ \hline rc & + & + & + & + & + \\ \hline + & + & + & + & + & + \\ \hline \end{array} = B.$$

The process terminates if A_{12} contains no zero, for then the columns through A_{11} , together with the rows through A_{22} and A_{33} , contain all the zeros of A . If each column of A_{12} contains at least one zero, then proceed in a manner exactly similar to that for Step 8, following, to replace some triple of zeros in A_{11} , A_{22} , and A_{33} by four drawn from A_{41} , A_{12} , A_{23} , and A_{34} ; then return to Step 2. Otherwise the columns of A_{12} are permuted so as to move all those containing no zeros into B_{13} . Exactly the same permutation is made on the rows of A_{22} , so that the diagonal zeros of A_{22} are retained for B_{22} and B_{33} .

STEP 8.

$$A = \begin{array}{|c|c|c|c|c|c|} \hline h & i & j & k & & \\ \hline 0 & c & + & + & + & + \\ \hline & 0 & & & + & + \\ \hline & & 0 & & + & + \\ \hline & & & 0 & rc & + \\ \hline rc & + & + & + & + & + \\ \hline + & + & + & + & + & + \\ \hline \end{array} \xrightarrow{j} \begin{array}{|c|c|c|c|c|c|} \hline h & c & + & + & + & + \\ \hline i & 0 & & + & + & + \\ \hline & & 0 & & + & + \\ \hline & & & 0 & rc & + \\ \hline rc & + & + & + & + & + \\ \hline + & + & + & + & + & + \\ \hline \end{array} = B.$$

An actual permutation need be made in this step only if A_{24} contains at least one zero, say in position (ij) . If so, then in a manner exactly similar to Steps 4 and 6, the total number of zeros on the diagonal is increased by replacing zeros in the three positions (ii) , (jj) and (hh) by four at positions : (ij) , (hi) , in row j of A_{45} , and in column h of A_{51} , where h denotes a row in A_{12} that has a zero in column i . Again, Steps 2 through 7 are repeated as

necessary finally to reach the form shown for Step 8 in which B_{24} has no zeros.

STEP 9.

$$A = \begin{array}{|c|c||c|c|c|c|} \hline 0 & c & + & + & + & + \\ \hline \text{---} & 0 & \text{---} & \text{---} & \text{---} & \text{---} \\ \hline \text{---} & \text{---} & 0 & \text{---} & \text{---} & \text{---} \\ \hline \text{---} & \text{---} & \text{---} & 0 & rc & + \\ \hline rc & + & + & + & + & + \\ \hline + & + & + & + & + & + \\ \hline \end{array} \rightarrow \begin{array}{|c|c||c|c|c|c|} \hline 0 & c & + & + & + & + \\ \hline \text{---} & 0 & c & + & + & + \\ \hline \text{---} & \text{---} & 0 & \text{---} & \text{---} & \text{---} \\ \hline \text{---} & \text{---} & \text{---} & 0 & \text{---} & \text{---} \\ \hline rc & + & + & + & + & + \\ \hline + & + & + & + & + & + \\ \hline \end{array} = B.$$

The process terminates if A_{23} contains no zero, for then the columns through A_{11} and A_{22} , together with the rows through A_{33} and A_{44} , contain all the zeros of A . If each column of A_{23} contains at least one zero, then proceed as for Step 10, following, to increase the number of diagonal zeros prior to a return to Step 2. Otherwise the columns of A_{23} are permuted so as to move all those containing no zeros into B_{24} . Exactly the same permutation is made on the rows of A_{33} , so that the zeros of A_{33} are retained in B_{33} and B_{44} .

STEPS 10 and on. If, after Step 9, there is a zero anywhere in B_{35} then it can easily be used to increase the total number of zeros on the diagonal—in a manner exactly similar to that for Steps 4, 6 and 8. This would be done by using with it appropriate zeros from B_{61}, B_{12}, B_{23} , and B_{56} . This process is continued until, at some stage, the process terminates in the manner illustrated by the discussion of Step 9.

Upon such termination, the array will have been permuted to have one of the “standard forms” shown next.

Process terminates at Step 1

$$\boxed{0} \quad \text{or} \quad \begin{array}{|c|c|} \hline 0 & \text{---} \\ \hline \text{---} & + \\ \hline \end{array} \quad \text{or} \quad \begin{array}{|c|c|} \hline 0 & + \\ \hline \text{---} & + \\ \hline \end{array} = A$$

Process terminates at Step 4

$$\begin{array}{|c|c|c|c|} \hline 0 & + & + & + \\ \hline \text{---} & 0 & rc & + \\ \hline r & + & + & + \\ \hline + & + & + & + \\ \hline \end{array}$$

Process terminates at Step 6

$$\begin{array}{|c|c|c|c|c|} \hline 0 & + & + & + & + \\ \hline \text{---} & 0 & \text{---} & + & + \\ \hline \text{---} & \text{---} & 0 & rc & + \\ \hline \text{---} & \text{---} & \text{---} & + & + \\ \hline rc & + & + & + & + \\ \hline + & + & + & + & + \\ \hline \end{array}$$

It is easily seen that these forms are all special cases of the following general form of \tilde{A} , special in the sense that they may be represented by selected rows and columns of \tilde{A} .

Process terminates at Step 8, or later

					2 	1 		1 + n	2 + n	3 + n	4 + n
1	2	3	4	-----	\tilde{n}	\tilde{n}	\tilde{n}	\tilde{n}	\tilde{n}	\tilde{n}	
1	0	c	+	+	- - - -	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	
2	0	c	+	- - - -	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	
3		0	c		+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	
4			0		+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	
	---	---	---					---	---	---	
$m - 2$					0	c	+ + + +	+ + + +	+ + + +	+ + + +	
$m - 1$				- - - -		0	c	+ + + +	+ + + +	+ + + +	
m				- - - -			0	+ + + +	+ + + +	+ + + +	
$m + 1$				- - - -				0	+ +	+ +	
$m + 2$				- - - -					0	rc	+
$m + 3$	rc	+	+	+	- - - -	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	
$m + 4$	+	+	+	+	- - - -	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	

$= \tilde{A}.$

In this form, the zeros are all contained within the columns passing through $\tilde{A}_{11}, \tilde{A}_{22}, \dots$, and \tilde{A}_{mm} and the rows passing through $\tilde{A}_{m+1\ m+1}$ and $\tilde{A}_{m+2\ m+2}$. Since there must obviously be at least as many lines needed to contain all the zeros as there are zeros on the diagonal, this establishes the theorem.

Basic algorithm. The “basic algorithm” consists in alternate applications of the process of permuting the matrix to standard form \tilde{A} and then using the Hungarian Method of Kuhn [3; 5].

The Hungarian Method reduces the sum of the elements of the matrix, by successive additions and subtractions of constants from the elements in lines of the matrix, until the diagonal elements are all zero. In general, a constant f is added to all of the rows R passing through $\tilde{A}_{m+1\ m+1}$ and $\tilde{A}_{m+2\ m+2}$, and the same constant f is subtracted from all of the columns C passing through $\tilde{A}_{m+1\ m+1}, \tilde{A}_{m+2\ m+2}, \tilde{A}_{m+3\ m+3}$ and $\tilde{A}_{m+4\ m+4}$. This constant

f , called a "fuse", is chosen as the value of the smallest element in columns C but not in rows R .

Since the solutions of assignment and transportation problems are left unchanged by adding a constant to each element of any line of the cost matrix, the Hungarian Method may be used in the basic algorithm in this way.

An Example (Kuhn [5])

				1	2	3	4		1	2	3	4			
1	8	7	9	9	-7	1	1	0	2	2	1	1	0	2	0✓
2	5	2	7	8	-2	2	3	0	5	6	2	3	0	✓	5
3	6	1	4	9	-1	3	5	0	3	8	3	5	0	3	6
4	2	3	2	6	-2	4	0	1	0	4	4	0✓	1	0	2

				1	2	3	4		1	2	3	4		
4	0✓	1	2	0	1	4	2	1	3	2	4	1	3	
2	3	0✓	4	5	2	0✓	0	1	2	2	0✓	4	3	5
1	1	0	0✓	2	3	4	2	1	0✓	1	2	0✓	0	2
3	5	0	6	3	6	0	5	3		3	0	6	5	3

				2	4	1	3		2	4	1	3		
2	0✓	4	2	4	1	4	2	0	2	4	2	0	2	3
1	0	0✓	0	1	2	0✓	0	✓	1	2	0✓	0	✓	2
4	2	3	0✓	0	3	4	2	0✓	0	4	3	0	✓	0
3	0	6	4	2	3	5	0	4	2	0	1	2	0✓	0

				2	4	1	3		1	2	3	4		
2	0✓	2	0	2	2	0✓	0	1	4	2	0	✓	3	
1	2	0✓	0	1	1	0✓	0	✓	2	0✓	0	✓	2	
4	2	3	0✓	0	3	4	2	0✓	0	4	3	0	✓	0
3	0	6	4	2	3	5	0	4	2	0	1	2	0✓	0

The required permutation on rows is therefore (14), corresponding to the zeros marked with checks. The other solution is the permutation (1 2 3 4), corresponding to the zeros marked with crosses.

REFERENCES

- Frank L. Hitchcock, *The distribution of a product from several sources to numerous localities*, J. Math. Phys. vol. 20 (1941) pp. 224–230.
- Merrill M. Flood, *On the Hitchcock distribution problem*, Pacific J. Math. vol. 3 no. 2 (1953) pp. 369–386.

3. Merrill M. Flood, *The traveling salesman problem*, Operations Res. vol. 4 no. 1 (1956) pp. 61–75. Also in: Operations Research for Management, vol. 2, McCloskey and Coppinger, editors, Johns Hopkins Press, 1956, pp. 340–357.

4. a. D. König, *Über Graphen und ihre Anwendung auf Determinantentheorie und Mengenlehre*, Math. Ann. vol. 77 (1916) pp. 453–465.

b. J. Egerváry, *Matrixok kombinatorius tulajdonságairól*, Mat. Fiz. Lapok (1931) pp. 16–28. (Translated as *Combinatorial properties of matrices* by H. W. Kuhn, ONR Logistics Project, Princeton, 1953, mimeographed.)

5. a. H. W. Kuhn, *The Hungarian method for the assignment problem*, Naval Res. Logist. Quart. vol. 2 nos. 1 and 2 (1955) pp. 83–97.

b. ———, *Variants of the Hungarian method for assignment problems*, Naval Res. Logist. Quart. vol. 3 no. 4 (1956) pp. 253–258.

6. a. L. R. Ford, Jr. and D. R. Fulkerson, *Solving the transportation problem*, Management Sci. vol. 3 no. 1 (1956) pp. 24–32.

b. Bernard A. Galler and Paul S. Dwyer, *Translating the method of reduced matrices to machines*, Naval Res. Logist. Quart. vol. 4 no. 1 (1957) pp. 55–71.

UNIVERSITY OF MICHIGAN,
ANN ARBOR, MICHIGAN

This page intentionally left blank

INDEX

- Adjugate, 150
Algorithm, 299, 305
Allocation problem, 218
Alternating graph, 124
Approximation
 Cebyshev, 231
 functional, 229
 in policy space, 233, 235
 monotonicity of, 232
 successive, 232, 236
Arcs, 117, 123
Assembly schedules, 281
Assignment problem, 299
Autotopism, 68
- Back-track, 92
Baker-Campbell-Hausdorff coefficients, 203
Bernoulli number, 206
Block designs, 2–5
Blocking, 269
Boldyreff, A., 235
Book-binding problem, 237
Bottleneck problems, 238
Burst noise, 291
- Caterer problem, 124, 237
Cebyshev approximation, 231
Central collineation, 50
Chinese remainder theorem, 296
Circulation theorem, 117, 118, 119
Classes of quadratic forms, 206
Collineation, 15, 49, 62
 central, 50
 elementary, 64, 65
 group, 9–13, 62
Combinations, 184
Combinatorial equivalence, 129ff.
Communication network problems, 261ff.
Commutators, 203
Complete graph, 162
Compound, 150
Configuration, v, k, λ , 165
Congruence, 201
Constraints, 221, 239
- Continuous version of the simplex technique, 239
Convex
 hull, 142
 set, 113, 124
 spaces, 6
Coordinatized, 62, 63
Covering theorems, 204
Curse of dimensionality, 227
Cyclic, 15
Cyclotomic numbers, 95
Cyclotomy, theory of, 95
- Decomposable, 169
Desarguesian plane, 10, 11, 63
Design of experiments, 246
Dickson-Hurwitz sums, 97
Difference set, 16, 109
Dimensionality
 curse of, 227
 reduction in, 241
Directed graphs, 281
Discreteness, 222
Distribution problem, Hitchcock, 299
Distributive lattice, 85
Division algebra, 54, 61
Doubly stochastic matrix, 166, 169ff.
Dreyfus, S., 217
Dual, 253
 representation, 93
Duality theorem of linear programming, 115, 117, 256
Dynamic programming, 217ff.
- Elementary
 collineations, 64, 65
 operations, 139
Encoding, 291
 full, 294
Endpoints, 162
Equivalence relation on matrices, 129
Error-correcting, 291
Exhaustive searches, 195ff.
Extreme
 point, 142
 ray, 145

- Fermat's last theorem, 205
 Finite
 Fourier series, 97
 graph, 162
 Parseval relation, 97
 planes, 53ff.
 projective plane, 2, 3, 62
 homomorphisms, 49
 Flood, M. M., 257
 Flooding technique, 235
 Flow problems, multi-commodity, 270
 Flow theorems, 118
 Flows in networks, 117
 Football pool, 204
 Ford, L. R., Jr., 224, 234, 257
 Forecast, 204
 Fourier series, finite, 97
 Free plane, 45, 52
 Fulkerson, D. R., 224, 235, 257
 Functional
 approximations, 226, 229
 equations, 221, 225
 Gauss forms, 207
 Graphs, 117, 118, 123, 162
 alternating, 124
 directed, 281
 Harris, T. E., 234
 Harris transportation problem, 233
 Hitchcock distribution problem, 299
 Hitchcock-Koopmans transportation problem, 224, 232, 251
 Homomorphisms
 of loops, 48
 of projective planes, 45ff.
 Hungarian method of Kuhn, 305
 Incidence matrix, 2, 113, 117, 123, 124, 164
 of G , 163
 of the v, k, λ configuration, 165
 Induced matrix, 151
 Invariants, 196, 197
 Inverse of a matrix, 140
 Isomorph rejection, 195ff.
 Isomorphic planes, 68
 Isotope, 54, 57, 60
 Isotopic algebras, 54, 55, 66, 68
 Jacobi sum, 96
 Join, 162
 Kantorovitch, L. V., 120, 251
 König's theorem, 86, 299, 300
 Königsberg bridge problem, 261
 Kronecker forms, 207
 Kruskal tree, 263
 Kuhn, H. W., 257, 299, 305
 Lagrange
 multipliers, 218, 226, 227, 244
 resolvent, 98
 Latin squares, 5–8, 71ff.
 Left vector space, 61
 Legendre-Sophie St. Germain criterion, 205
 Line at infinity, 63
 Linear inequalities, 113ff., 141
 solvability of, 144
 Linear programming, 129, 270
 duality theorem of, 115, 117
 integer solutions in, 211ff.
 Lines, 62, 162
 Loop, 62
 Majorized, 152
 Marcus, M., 156, 166
 Matrices with non-negative elements, 170
 Matrix of relations, 208
 Min-cut max-flow theorem, 115, 118, 119
 Minimal cost connecting networks, 261
 Monotonicity of approximation, 232
 Multi-commodity flow problems, 270
 Multiplier, 16, 17
 Mutually exclusive activities, 222
 n th shortest paths, 267
 n -tuple decomposition, 201
 Network flow, 122
 Network problems, 234
 communication, 261
 Newman, M., 166
 Nodes, 117, 123, 124
 Nonlinear transportation problem, 225, 233
 Norm, 155
 Operations, elementary, 139
 Optimal
 paths through networks, 264
 routing of messages, 262
 routing problems, 270
 trajectory, 236
 Optimality, principle of, 220, 235, 240
 Order, 45
 of a pivotal transformation, 136

- Ordinary point, 62
 Oriented graph, 162
 Orthogonal latin squares, 5–9, 91, 93, 195
 10×10 , 71ff., 197
 Orthogonality relation, 99
- Parseval relation, 97
 Partially ordered sets, 85
 Partitioned, 20
 Paths, 117, 123, 124
 Pauli exclusion principle, 141
 Permanent, 166, 169ff.
 Permutation matrix, 113, 124
 Permutations, 184ff.
 Pivot, 133
 Pivotal transform, 135
 Planar ternary rings, 45, 62
 Poincaré, H., 163
 Point at infinity, 62
 Points, 62, 162
 Polar cone, 146
 Polyhedral cone, 146
 Polyhedron problem, 142, 147
 Power matrix, 151
 Primitive roots, 206
 Principle of optimality, 220, 235, 240
 Problem of the queens, 91, 93
 Programming
 dynamic, 217
 linear, 211
 Projective plane, 45, 197
 finite, 2, 3, 62
 homomorphisms, 45ff.
 of order eight, 94, 198
 order, 15
- Quadratic field units, 206
 Quadratic forms, classes of, 206
 Quasigroup, 54, 62
 Queens problem, 91, 93
- Rado-Hall theorem, 87
 Recurrence relations, 221
 Reduction in dimensionality, 241
 Redundancy, 294
 Reliability, 243
 Remainder theorem, Chinese, 296
 Representatives, systems of, 114, 122
 distinct, 113, 114, 124, 166
 restricted, 115, 116,
 Residue difference sets, 109
 Resolvent of Lagrange, 98
- Right
 characteristic function, 58
 powers, 58
 Routing problem, 235
- SEAC—the National Bureau Standards Eastern Automatic Computer, 201
 SWAC, 91, 93
 Scheduling problems, 237
 Semigroups of order five, 199
 Sequential
 search, 245
 testing, 246
 Shadow prices, 272
 Shears, 65
 Sieve, 182
 Simplex method, 129
 continuous version of, 239
 Slightly intertwined
 matrices, 239
 symmetric matrices, 241
 Solvability problem, 144
 Steiner triple systems, 1, 2, 9, 197
 of order 15, 199
 Stochastic matrix, doubly, 166, 169ff.
 Successive approximations, 226, 232, 236
 Systems of
 distinct representatives, 113, 114, 124, 166
 representatives, 114, 122
 restricted representatives, 115, 116
- Ternary ring, 45, 46, 62
 Threshold methods, 252
 Tinsley, M. F., 166
 Translate, 17
 Translation group, 64
 Transportation problem, 124, 251, 299
 Harris, 233
 Hitchecock-Koopmans, 224, 232, 251
 nonlinear, 225, 233
 Traveling salesman problem, 235, 299
 Twisted fields, 69
- Unimodular property, 114, 115, 117, 118, 123, 124, 125
 Urn problem, 141
- v, k, λ configuration, 165
 van der Waerden, B. L., 166
 Vandermonde matrix, 160
 Veblen, O., 163
 Veblen-Wedderburn systems of order 16, 198
 Warehousing problems, 124, 237

