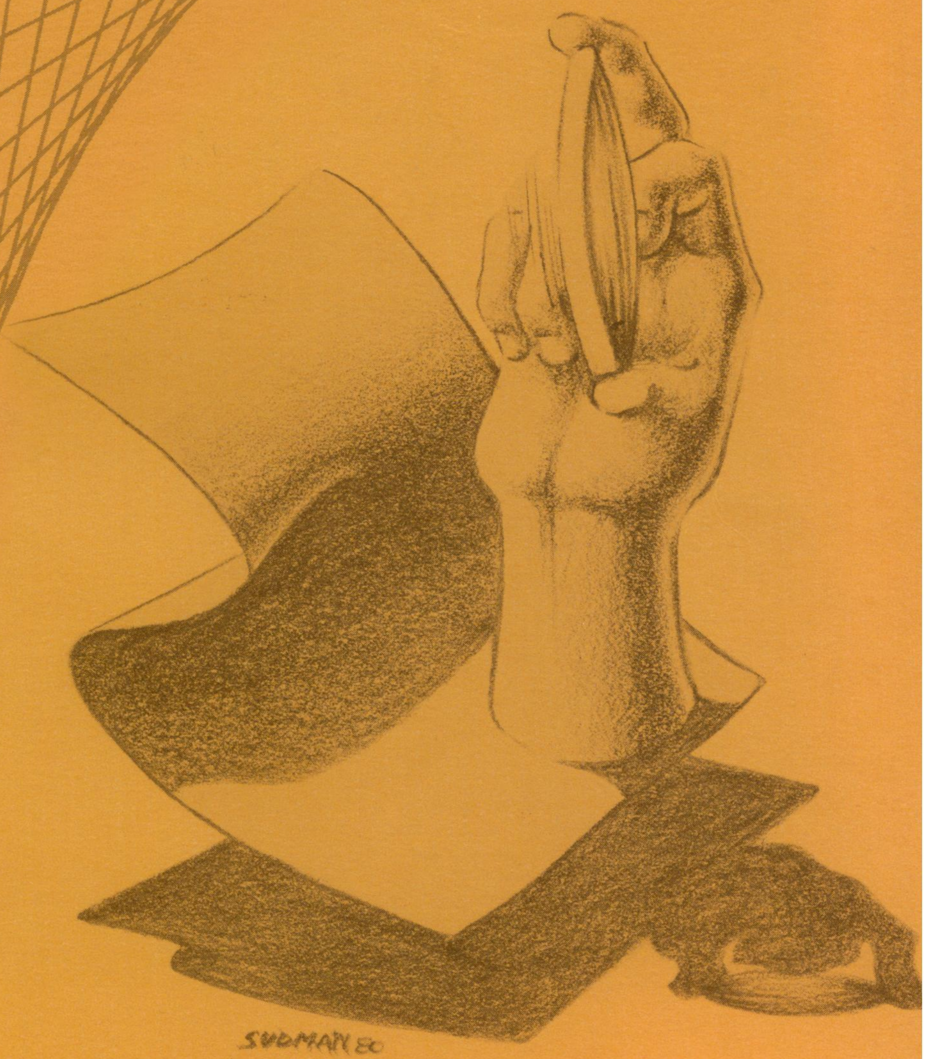
MATHEMATICS MAGAZINE



Vol. 53, No. 1
January 1980

CATASTROPHES • INTERVAL GRAPHS
ARC LENGTH • VOTING PROCEDURES

MATHEMATICS MAGAZINE

Volume 53, Number 1, January 1980
(ISSN—0025-570X)

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

COVER: The cusp catastrophe is a model for a popsicle stick snapping back and forth; see p. 13.

EDITORIAL POLICY

Mathematics Magazine is a journal of collegiate mathematics designed to enrich undergraduate study of the mathematical sciences. The *Magazine* should be an inviting, informal journal emphasizing good mathematical exposition of interest to undergraduate students. Manuscripts accepted for publication in the *Magazine* should be written in a clear and lively expository style. The *Magazine* is not a research journal, so papers written in the terse "theorem-proof-corollary-remark" style will ordinarily be unsuitable for publication. Articles printed in the *Magazine* should be of a quality and level that makes it realistic for teachers to use them to supplement their regular courses. The editors especially invite manuscripts that provide insight into applications and history of mathematics. We welcome other informal contributions, for example, brief notes, mathematical games, graphics and humor.

Editorial correspondence should be sent to: Mathematics Magazine, Department of Mathematics, St. Olaf College, Northfield, Minnesota 55057. Manuscripts should be prepared in a style consistent with the format of *Mathematics Magazine*. They should be typewritten and double spaced on $8\frac{1}{2}$ by 11 paper. Authors should submit the original and one copy and keep one copy as protection against possible loss. Illustrations should be carefully prepared on separate sheets of paper in black ink, the original without lettering and two copies with lettering added; the printers will insert printed letters on the illustration in the appropriate locations.

Authors planning to submit manuscripts may find it helpful to obtain the more detailed statement of guidelines available from the editorial office.

ABOUT OUR AUTHORS

Donald O. Koehler ("The Cost Accounting Problem") received his undergraduate degree from Hanover College and his Ph.D. in functional analysis from Indiana University. He has taught at Hanover College, DePauw University, the University of New Castle in Australia (Summer 1972), and is currently Associate Professor of Mathematics at Miami University in Oxford, Ohio, and Chairman of the Ohio Section of the MAA. This article began as an answer to a student question in a finite mathematics class and has evolved to its present form as the result of many discussions with accountants (practicing and teaching) and numerous visits to the library.

Jan W. Auer ("Mathematical Preliminaries to Elementary Catastrophe Theory") was born in Utrecht, Germany, and studied engineering before turning to mathematics. He holds a Ph.D. from McGill University where he specialized in differential geometry and algebraic topology. He is currently working with catastrophe theory and applications of fiber bundle theory to general relativity. He moved to his current position at Brock University after teaching at McGill University and the University of Toronto.

BUSINESS INFORMATION. Mathematics Magazine is published by the Mathematical Association of America at Washington, D.C., five times a year in January, March, May, September, and November. Ordinary subscriptions are $18 per year. Members of the Mathematical Association of America or of Mu Alpha Theta may subscribe at special reduced rates. Colleges and university mathematics departments may purchase bulk subscriptions (5 or more copies to a single address) for distribution to undergraduate students.

Subscription correspondence and notice of change of address should be sent to A. B. Willcox, Executive Director, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Back issues may be purchased, when in print, from P. and H. Bliss Co., Middletown, Connecticut 06457.

Advertising correspondence should be addressed to Raoul Hailpern, Mathematical Association of America, SUNY at Buffalo, Buffalo, New York 14214.

Second class postage paid at Washington, D.C., and additional mailing offices.

The Cost Accounting Problem

A survey of methods of cost allocation using linear algebra and Markov chains.

DON O. KOEHLER
Miami University
Oxford, OH 45056

Any organization selling goods or services needs to determine its costs for the purpose of establishing prices, adding personnel or eliminating departments, and for external reporting. The problem is complicated because there are usually costs not easily identifiable with any particular good or service. For example, to what good or service should be assigned the cost of research and development for products never making it into production? Or what product (if any) should bear the cost of accounting services performed for that portion of the maintenance department providing services for the accounting department? And which group of students should be charged for time a faculty member spends on research or in a committee meeting deciding whether to charge colleagues for overdue books? The cost accountant is the person who must answer questions like these, and the techniques he or she uses in solving these problems have interesting relationships with several different branches of mathematics.

To illustrate these ideas we shall consider a simplified firm having four departments. Two of these departments, Maintenance and Accounting, are service departments producing no goods or services for sale, and hence no income to cover their expenses. The remaining two departments, Machining and Finishing, are production departments producing products and services to be sold. Our problem will be to assign the relevant costs of our company to the appropriate production departments.

Initially the accountant assigns these costs as direct costs of the appropriate departments. A general rule is to assign each cost to that department with which it is most clearly identifiable, and for which there is no intervening basis for assignment. Let us assume that in our example the assignments have been made and are as displayed in the third line of TABLE 1. Next the accountant must decide how the work of each service department is distributed among the departments it serves. This may be decided on a variety of bases, such as machine or labor hours, materials used, number of workers, and so forth. Traditionally, the work a service department provides itself is divided proportionally among the remaining departments. In any case let us assume that these decisions have been made, the services have been measured, and the results are those displayed in the first and second lines of TABLE 1. Once the accountant has this data he or she can proceed to assign all the direct costs of the service departments to the appropriate production departments.

			User of Service		
Source of Service	Accounting	Maintenance	Machining	Finishing	Totals
Accounting	0	20%	20%	60%	100%
Maintenance	10%	0	80%	10%	100%
Direct Costs	\$39,200	\$9,800	\$30,000	\$15,000	\$94,000

Accounting Data

TABLE 1

System of Equations Method

One method for assigning direct costs involves constructing and solving a system of linear equations. This method, which frequently appears in books on finite mathematics and linear algebra, is easy to describe. The total cost of a department is taken to be the sum of its direct cost plus the indirect cost of services provided by other departments. Let x_i be the total cost of the ith department (ordering the departments in the order in which they appear in TABLE 1). According to the data in TABLE 1, Accounting has direct costs of \$39,200, and uses ten percent of the services provided by Maintenance. As a result its total costs will be

$$x_1 = 39{,}200 + .1x_2.$$

In a similar manner we have

$$x_2 = 9{,}800 + .2x_1,$$
$$x_3 = 30{,}000 + .2x_1 + .8x_2,$$
$$x_4 = 15{,}000 + .6x_1 + .1x_2.$$

This system can be solved by row reduction to obtain total costs for each of the four departments,

$$x_1 = 41{,}000, \qquad x_2 = 18{,}000, \qquad x_3 = 52{,}600, \qquad x_4 = 41{,}400.$$

These total costs add up to 153,000, considerably more than the original 94,000, because during the redistribution of indirect costs the same dollar may be counted several times, once each time it is transferred. The accountant uses these total costs together with the data in TABLE 1 to construct TABLE 2 which summarizes the cost transfers.

Matrix Algebra Solution

A second method for assigning cost, closely related to the first, involves matrix algebra. The system of linear equations (from the first method) is rewritten as a system of matrix equations, the first involving the service departments and the second involving the production departments:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 39200 \\ 9800 \end{bmatrix} + \begin{bmatrix} 0 & .1 \\ .2 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 30000 \\ 15000 \end{bmatrix} + \begin{bmatrix} .2 & .8 \\ .6 & .1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Costs (with transfers added)

From/To	Accounting	Maintenance	Machining	Finishing	Totals
Direct Costs	39,200	9,800	30,000	15,000	94,000
Accounting	0	8,200	8,200	24,600	41,000
Maintenance	1,800	0	14,400	1,800	18,000
Total Costs Including Transfers	41,000	18,000	52,600	41,400	153,000

Cost Transfers

To/From	Accounting	Maintenance	Machining	Finishing	Totals
Accounting	0	1,800	0	0	1,800
Maintenance	8,200	0	0	0	8,200
Machining	8,200	14,400	0	0	22,600
Finishing	24,600	1,800	0	0	26,400
Total Transfers	41,000	18,000	0	0	59,000
Total Costs After Transfers	–0–	–0–	52,600	41,400	94,000

Summary of Total Costs and Transfers

TABLE 2

The first equation is of the form $\vec{x}=\vec{d}+B\vec{x}$, whose solution is $\vec{x}=(I-B)^{-1}\vec{d}$. The second matrix equation is of the form $\vec{y}=\vec{f}+C\vec{x}$, which now becomes $\vec{y}=\vec{f}+C\cdot(I-B)^{-1}\vec{d}$. Since

$$(I-B)^{-1}=\left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}-\begin{bmatrix}0 & .1\\ .2 & 0\end{bmatrix}\right)^{-1}=\begin{bmatrix}1 & -.1\\ -.2 & 1\end{bmatrix}^{-1}=\begin{bmatrix}\frac{100}{98} & \frac{10}{98}\\ \frac{20}{98} & \frac{100}{98}\end{bmatrix},$$

we can compute

$$\vec{x}=(I-B)^{-1}\vec{d}=\begin{bmatrix}\frac{100}{98} & \frac{10}{98}\\ \frac{20}{98} & \frac{100}{98}\end{bmatrix}\begin{bmatrix}39200\\ 9800\end{bmatrix}=\begin{bmatrix}41000\\ 18000\end{bmatrix}$$

and

$$\vec{y}=\begin{bmatrix}30000\\ 15000\end{bmatrix}+\begin{bmatrix}.2 & .8\\ .6 & .1\end{bmatrix}\begin{bmatrix}41000\\ 18000\end{bmatrix}=\begin{bmatrix}52600\\ 41400\end{bmatrix}.$$

While this solution is the same as that obtained by the first method, some accountants prefer the matrix approach because it provides a good deal of economy and simplification. However, others have theoretical objections to this method in either form since the redistributed costs of the service departments displayed in TABLE 2 total $59,000 instead of equaling $49,000, the sum of the direct costs of the service departments. There are several readable articles in the accounting literature exploring these ideas, and the reader may find it interesting to read them: [3], [14], [16], [19], and [21]. See also [17] and [18].

Other Methods Used by Accountants

While some accountants use linear algebra methods to solve the cost transfer problem, it has traditionally been solved by other far simpler methods. We now consider three methods discussed in Horngren [9].

Direct Redistribution. In this widely used method the direct costs of the service departments are redistributed directly to production departments on the basis of services rendered. To apply this method we note that Accounting does eighty percent of its work for production departments, twenty percent of it being done for Machining and sixty percent for Finishing (see TABLE 1.) Thus in the direct redistribution method Accounting would charge two-eighths of its cost of $39,200 to Machining, the remaining six-eighths to Finishing, while ignoring all work done for Maintenance. The details of the transactions are displayed in TABLE 3.

	Accounting	Maintenance	Machining	Finishing	Totals
Direct Costs	39,200	9,800	30,000	15,000	94,000
Redistributions:					
Accounting					
$(0,0,\frac{2}{8},\frac{6}{8})$	(39,200)		9,800	29,400	0
Maintenance					
$(0,0,\frac{8}{9},\frac{1}{9})$		(9,800)	8,711.11	1,088.89	0
Total Costs	−0−	−0−	48,511.11	45,488.89	94,000

The Direct Redistribution of Costs

TABLE 3

In our example the total costs of the production departments obtained by this direct redistribution method differ from those obtained by the linear algebra methods by eight or nine percent. Because differences of this magnitude could affect the outcome of the decision making process, an accountant using the direct redistribution method may periodically compare the results with those of the linear algebra methods. However, differences of as much as five percent may be acceptable provided they don't significantly change the financial picture (turn a loss into a profit or vice versa).

Sequential redistribution. This method consists of a sequence of redistributions of service department costs, where at each step the costs of some service department are redistributed, on the basis of services rendered, to all those departments not previously considered, not just to production departments. Suppose in our example we first redistribute Accounting's costs to the remaining three departments in a 2:2:6 ratio. Then we would next redistribute Maintenance's costs (including those charges from Accounting) to the remaining two departments in an 8:1 ratio. The results of this method are displayed in TABLE 4, and are in close agreement with those obtained by the linear algebra methods.

	Accounting	Maintenance	Machining	Finishing	Totals
Direct Costs	39,200	9,800	30,000	15,000	94,000
Redistributions:					
Accounting					
$(0,\frac{2}{10},\frac{2}{10},\frac{6}{10})$	(39,200)	7,840	7,840	23,520	0
Maintenance					
$(0,0,\frac{8}{9},\frac{1}{9})$		(17,640)	15,680	1,960	0
Total Costs	−0−	−0−	53,520	40,480	94,000

The Step Method of Redistribution of Costs

TABLE 4

The sequential redistribution method generally provides results closer to the linear algebra solution than does the direct redistribution method. However, the results are affected by the order in which the redistributions take place. In our example, redistributing Maintenance first would have yielded a total cost of $47,885 for Machining and $46,115 for Finishing. In fact, by changing the data in TABLE 1 the reader can obtain an example in which both orders of redistributing costs yield solutions significantly different from the linear algebra solutions.

Successive Simultaneous Redistribution. This method, which takes into account all the reciprocal services provided by the service departments, consists of a succession of simultaneous redistributions of service department costs to all departments on the basis of services rendered. In our example, Accounting and Maintenance first redistribute their respective costs on the basis of services rendered. In so doing, each service department is billed by the other for services rendered (see TABLE 5). In step two, both service departments redistribute these new costs, with the process continuing for as many steps as necessary. Interestingly enough, the successive

	Accounting	Maintenance	Machining	Finishing	Totals
Direct Costs	39,200	9,800	30,000	15,000	94,000
First Redistribution:					
Accounting $(0,\frac{2}{10},\frac{2}{10},\frac{6}{10})$	(39,200)	7,840	7,840	23,520	0
Maintenance $(\frac{1}{10},0,\frac{8}{10},\frac{1}{10})$	980	(9,800)	7,840	980	0
Costs After First Redistribution	980	7,840	45,680	39,500	94,000
Second Redistribution:					
Accounting $(0,\frac{2}{10},\frac{2}{10},\frac{6}{10})$	(980)	196	196	588	0
Maintenance $(\frac{1}{10},0,\frac{8}{10},\frac{1}{10})$	784	(7,840)	6,272	784	0
Costs After Second Redistribution	784	196	52,148	40,872	94,000
Costs After Third Redistribution	19.6	156.8	52,461.6	41,362	94,000
Costs After Tenth Redistribution	.00012	.00003	52,599.99+	41,399.99+	94,000

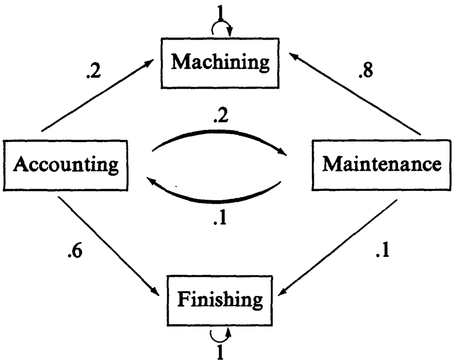
Results of Ten Successive Simultaneous Redistributions

TABLE 5

redistributions seem to lead to the same total costs for the production departments as do the linear algebra methods. It is also of interest to compute the total amounts redistributed by the service departments in the sequence of redistributions. For example, for Accounting the amounts are: 39,200 after one redistribution, 40,180 after two, 40,961 after three, ...40,999.99+ after 10, and so forth. Apparently the redistributed amounts converge to the total costs for the service departments obtained in the linear algebra methods, and the total transfers (e.g., for Accounting, $980+784+19.6+\cdots+.00012=1800$) equals the amounts "counted twice" in the linear algebra method.

The Markov Chain Solution

It is no accident that in our example the succession of simultaneous redistributions led to the same solution as the linear algebra methods. This will always happen and can be verified by interpreting the method as an absorbing Markov chain. We shall illustrate the idea with our example. To be consistent with the accountant's use of left hand matrix notation, we shall use left hand notation for Markov chains. Thus the i-jth entry of the transition matrix will be the probability an object located in the jth state will move to the ith state during a given time period. For an account of this approach see Pearl [20]. The states of the Markov chain are taken to be the various departments of the organization. As the direct cost dollars are repeatedly redistributed they can be thought of as moving among the states according to the transition probabilities contained in the accounting data displayed in TABLE 1. That is, the chance that a direct cost dollar located in (charged to) department j will move (be redistributed) to department i during the next time period (redistribution) is taken to be the proportion of the work of department j that is done for department i. (A production department is assumed to do all of its work for itself.) The data in TABLE 1 yields the following transition graph and transition matrix T:



	Acct.	Maint.	Mac.	Fin.
Accounting	0	.1	0	0
Maintenance	.2	0	0	0
Machining	.2	.8	1	0
Finishing	.6	.1	0	1

$= T.$

Note that the production departments play the role of absorbing states, and since it is possible to get to an absorbing state from any other state, our Markov chain is absorbing.

Let $\vec{d}_0 = (39200, 9800, 30000, 15000)'$ be the initial distribution of direct cost dollars among the four departments. Then after the first redistribution the costs will be distributed as

$$\vec{d}_1 = T\vec{d}_0 = \begin{bmatrix} 0 & .1 & 0 & 0 \\ .2 & 0 & 0 & 0 \\ .2 & .8 & 1 & 0 \\ .6 & .1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 39200 \\ 9800 \\ 30000 \\ 15000 \end{bmatrix} = \begin{bmatrix} 980 \\ 7840 \\ 45680 \\ 39500 \end{bmatrix}.$$

After the second redistribution the costs will be $\vec{d}_2 = T\vec{d}_1 = (784, 196, 52148, 40872)'$, and so forth. To determine the limiting distribution we express the transition matrix in terms of the notation used in the matrix algebra method:

$$T = \left[\begin{array}{cc|cc} 0 & .1 & 0 & 0 \\ .2 & 0 & 0 & 0 \\ \hline .2 & .8 & 1 & 0 \\ .6 & .1 & 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} B & 0 \\ \hline C & I \end{array}\right].$$

The $i-j$th entry of $C(I-B)^{-1}$ is the probability that a direct cost dollar initially in the jth nonabsorbing state will be absorbed in the ith absorbing state. Thus if $\vec{d} = (39200, 9800)'$ is the initial distribution of direct cost dollars for the service departments, $C(I-B)^{-1}\vec{d}$ is the eventual distribution of the dollars among the production departments. As a result, the total cost vector $\vec{y}$ for the production departments can be obtained as the sum of the vector of redistributed costs and the vector $\vec{f}$ of production department's direct costs. The result, $\vec{y} = \vec{f} + C(I-B)^{-1}\vec{d}$, is precisely the matrix algebra solution obtained earlier.

We can also verify that the total costs of the service departments will equal the total amounts redistributed by the service departments in the sequence of redistributions. In the matrix algebra discussion we showed that the total costs of the service departments equaled $(I-B)^{-1}\vec{d}$. According to the Markov chain theory the $i-j$th entry of $(I\text{-}B)^{-1}$ is the average number of times the ith nonabsorbing state is visited by an object starting in the jth nonabsorbing state. Thus if $\vec{d}$ is the initial distribution vector for the nonabsorbing states, the ith entry of $(I-B)^{-1}\vec{d}$

is the expected number of visits to the ith nonabsorbing state during the sequence of time periods. In terms of our example, this is the number of direct cost dollars charged to the ith production department during the sequence of redistributions. Therefore, it must equal the total amount redistributed by the ith service department.

These facts provide some additional dividends. For example, in the successive redistributions method it is clear the service departments costs are all borne by the production departments. As a result this will always be the case when the linear algebra solutions are used. In addition, the $i-j$th entry of $C(I-B)^{-1}$, namely

$$\begin{bmatrix} \frac{36}{98} & \frac{82}{98} \\ \frac{62}{98} & \frac{16}{98} \end{bmatrix} = \begin{array}{r} \\ \text{Machining} \\ \text{Finishing} \end{array} \begin{array}{cc} \text{Accounting} & \text{Maintenance} \\ \left[\begin{array}{c} .367 \\ .633 \end{array}\right. & \left.\begin{array}{c} .837 \\ .163 \end{array}\right] \end{array}$$

describes how likely it is for a direct cost dollar of the jth service department to be eventually redistributed to the ith production department. This means that Machining pays 36.7% of the costs of Accounting and 83.7% of the costs of Maintenance while Finishing pays 63.7% and 16.3% respectively. These percentages completely detail the way in which service department costs are paid by production departments. They take into account all direct costs of a service department that are eventually redistributed to a production department, and in this way give an accurate record of the work done, directly and indirectly, by each service department for each production department. (For other ways in which Markov chains are associated with accounting see **[2]**, **[5]**, and **[11]**.)

The Numerical Analysis Solution

There is another mathematical techique associated with the fact that the successive redistributions converge to the same solution as the linear algebra methods. The successive redistribution method can be thought of as one in which the accountant initially uses \$39,200, \$9,800, \$30,000, and \$15,000 as estimates of the total costs of the four departments, and then performs a series of adjustments to arrive at the correct answer. This same technique can be used to solve the system of equations obtained in the first solution method.

Consider the first two equations

$$x_1 = 39{,}200 + .1x_2$$
$$x_2 = 9{,}800 + .2x_1.$$

A solution of this system is a pair (x_1, x_2), which when substituted in both equations results in equality. If a pair (a,b) "close" to (x_1, x_2) is substituted in the right hand side of the equations, we should expect both sides to be close.

$$x_1 \approx 39{,}200 + .1b$$
$$x_2 \approx 9{,}800 + .2a.$$

This suggests using the right hand side of the system, together with an approximate solution, to generate another approximate solution that may be better. Using our accounting experience as a guide we use $a = 39{,}200$ and $b = 9{,}800$ as our initial estimate. Doing this yields new estimates

$$x_1 \approx 39{,}200 + .1(9{,}800) = 40{,}180$$
$$x_2 \approx 9{,}800 + .2(39{,}200) = 17{,}640.$$

Repeating the process using these new estimates yields $x_1 \approx 40{,}964$, $x_2 \approx 17{,}836$. The third iteration is $x_1 \approx 40{,}983.60$ and $x_2 \approx 17{,}992.80$, and so forth. We see that the estimates produced by this method are converging to the solution. They are in fact the total redistributions arrived at by the method of successive simultaneous redistribution. This technique is well known: it is called the Jacobi Simultaneous Method for solving a system of linear equations, and a discussion of it can be found in most numerical analysis books. The successive redistribution method the accountant uses can be thought of as a simple variation of it. Although the method fails for some general systems of equations, it will always solve the system associated with the cost accounting problem (see [4]).

The Input-Output Model

We have explored relationships between the cost accounting problem and several branches of mathematics. We conclude our remarks by showing how the cost accounting problem provides insight into another common problem that arises when interrelated industries determine production levels in order to meet both internal and external demands.

To illustrate the ideas we use two interrelated companies, United Coal and Consolidated Electricity. Each company uses output from the other as input to its own production. We shall assume that to produce one dollar's worth of coal requires ten cents' worth of electricity (to run the drilling equipment), and to produce one dollar's worth of electricity requires twenty cents' worth of coal (to power the generators). The production of these goods also requires "factor payments" including wages (returns to labor), rents (returns to land and natural resources), interest (returns to loan funds), and profit (the residual). All the pertinent economic information is displayed in TABLE 6.

	Consolidated Electricity	United Coal
Consolidated Electricity	0	.10
United Coal	.20	0
Electric Factor Payments	.80	0
Coal Factor Payments	0	.90

Costs to Produce One Dollar's Worth of Coal and Electricity

TABLE 6

We shall assume each company has an adequate inventory of their product available for sale. That is, United Coal has a supply of previously mined coal, and Consolidated Electricity has a supply of electricity as represented by stockpiled coal. (Without such inventories, the companies could not operate. In fact they would face an interesting problem in attempting to restart

	Income		Factor Payments		Inventory Replaced	
	Electricity	Coal	Electricity	Coal	Electricity	Coal
From Consumers	100,000	25,000				
First Redistribution:						
Electricity (0,.2,.8,0)	(100,000)	20,000	80,000		100,000	
Coal (.1,0,0.9)	2,500	(25,000)		22,500		25,000
Totals	2,500	20,000	80,000	22,500	100,000	25,000
Second Redistribution:						
Electricity (0,.2,.8,0)	(2,500)	500	2,000		2,500	
Coal (.1,0,0.9)	2,000	(20,000)		18,000		20,000
Totals	2,000	500	82,000	40,500	102,500	45,000

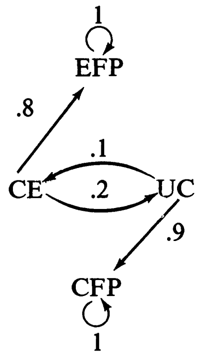
The First Two Redistributions of Income

TABLE 7

production if both inventories were ever depleted by a strike.) We further assume that each day consumers use \$100,000 worth of electricity and \$25,000 worth of coal. Our problem is to determine production levels that will maintain both inventories at current levels.

The process of replacing amounts used by consumers requires additional coal and electricity, thus necessitating further production. This is similar to the accounting problem where the simultaneous redistribution of costs resulted in additional service department costs, thus requiring further redistributions. In fact, the accountant's successive redistributions can be used to solve this problem, where instead of redistributing costs, we redistribute income to pay for replacement of inventories. The first two steps of this process are illustrated in TABLE 7.

The Markov chains theory can be used to arrive at a final solution. We begin with the transition graph and transition matrix:



	CE	UC	EFP	CFP
Consolidated Electric	0	.1	0	0
United Coal	.2	0	0	0
Electric Factor Payments	.8	0	1	0
Coal Factor Payments	0	.9	0	1

$$= \left[\begin{array}{c|c} B & 0 \\ \hline C & I \end{array}\right]$$

In this example companies play the role of the service departments and the factor payments are like the production departments. The total inventory replaced by a company, like the total costs of the service department in the accounting problem, equals the total income acquired during the sequence of redistributions. Thus the answer to the problem of production levels is the same as the answer to the problem of cost allocation: $\vec{x}=(I-B)^{-1}\vec{d}$, where $\vec{d}=(100000, 25000)'$ is the vector of direct incomes. Hence

$$\vec{x}=\begin{bmatrix} \frac{100}{98} & \frac{10}{98} \\ \frac{20}{98} & \frac{100}{98} \end{bmatrix}\begin{bmatrix} 100{,}000 \\ 25{,}000 \end{bmatrix}=\begin{bmatrix} 104{,}591.84 \\ 45{,}918.36 \end{bmatrix}.$$

We conclude that replacing the amounts used by consumers requires \$104,592 worth of electricity, and \$45,919 worth of coal. This may seem confusing since the values of the replacement inventories exceed the values of the goods sold to consumers. However, the accounting analysis displayed in TABLE 8 shows that there is no difficulty. The companies' incomes equal their expenses, the additional income being generated by intercompany sales. The factor payments are like the total costs of the production departments. Therefore $\vec{y}=\vec{f}+C(I-B)^{-1}\vec{d}$, where $\vec{f}=(0,0)'$ is the direct income of the factor payment. Hence

$$\vec{y}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}+\begin{bmatrix} \frac{8}{10} & 0 \\ 0 & \frac{9}{10} \end{bmatrix}\begin{bmatrix} \frac{100}{98} & \frac{10}{98} \\ \frac{20}{98} & \frac{100}{98} \end{bmatrix}\begin{bmatrix} 10{,}000 \\ 25{,}000 \end{bmatrix}=\begin{bmatrix} 83{,}673.47 \\ 41{,}326.53 \end{bmatrix}.$$

Of course this is not the only way to solve the input-output problem. Any of the accountant's techniques can be used. It can also be solved without any reference to the accounting problem, but the analogy is a natural one that provides insight into the problem and its solution. Actually the relationship between the two problems is a two-way street, and accountants have explored ways to use the input-output model in solving the accounting problem. (See **[1]**, **[7]**, **[8]**, **[10]**, **[12]**, and **[13]**.)

Income		Consolidated Electricity	United Coal
From:	Consumers	100,000	25,000
	Consolidated Electricity		20,918.37
	United Coal	4,591.84	
	Totals	104,591.84	45,918.37
Expenses			
To:	Consolidated Electricity		4,591.84
	United Coal	20,918.37	
	Electric Factor Payments	83,673.47	
	Coal Factor Payments		41,326.53
	Totals	104,591.84	45,918.37

Income and Expenses of Consolidated Electric and United Coal

TABLE 8

References

[1] William F. Bentz, Input-output analysis for cost accounting, planning and control: a proof, The Accounting Rev., 48 (1973) 377–380.

[2] Neil Churchill, Accounting for affirmative action programs, a stochastic flow approach, The Accounting Rev., 15 (1975) 643–56.

[3] ______, Linear algebra and cost allocations: some examples, The Accounting Rev., 39 (1964) 894–904.

[4] S. T. Conte and Carl de Boor, Elementary Numerical Analysis, 2nd ed., McGraw-Hill, New York, 1972.

[5] A. Wayne Corcoran and Wayne E. Leininger, Stochastic process costing models, The Accounting Rev., 48 (1973) 105–113.

[6] Dartmouth College Writing Group, Modern Mathematical Methods and Models, vol. 2, Cushing-Malloy, Ann Arbor, Michigan, 1959.

[7] Gerald A. Feltham, Some quantitative approaches to planning for multiproduct production systems, The Accounting Rev. 45 (1970) 11–26.

[8] Trevor E. Gambling, A technological model for use in input-output analysis and cost accounting, Management Accounting, 50 (1968) 33–38.

[9] Charles T. Horngren, Cost Accounting: A Managerial Emphasis, 3rd ed., Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

[10] Yuji Ijiri, An application of input-output analysis to some problems in cost accounting, Management Accounting, 49 (1968) 49–61.

[11] John G. Kemeny, J. Laurie Snell, and Gerald L. Thompson, Introduction to Finite Mathematics, 3rd ed., Prentice-Hall, Englewood Cliffs, New Jersey, 1974, pp. 462–468.

[12] John Leslie Livingstone, Input-output analysis for cost accounting, planning and control, The Accounting Rev., 44 (1969) 48–69.

[13] ______, Input-output analysis for cost accounting, planning and control: a reply, The Accounting Rev., 48 (1973) 381–82.

[14] ______, Matrix algebra and cost allocation, The Accounting Rev., 43 (1968) 503–508.

[15] Wassily W. Leontief, The Structure of American Economy 1919–1939, 2nd ed., Oxford Univ. Press, New York, 1951.

[16] Rene P. Manes, Comment on matrix theory and cost allocation, The Accounting Rev., 40 (1965) 640–643.

[17] Liao Mawsen, A matrix approach to the depreciation lapse schedule preparation, The Accounting Rev., 51 (1976) 364–69.

[18] George R. McGrail, Accounting and matrix theory, Woman CPA, 38 (1976) 8–9, 28.

[19] Roland Minch and Enrico Petri, Matrix models of reciprocal service cost allocation, The Accounting Rev., 47 (1972) 576–580.

[20] Martin Pearl, Matrix Theory and Finite Mathematics, McGraw-Hill, New York, 1973.

[21] Thomas H. Williams and Charles H. Griffin, Matrix theory and cost allocation, The Accounting Rev., 39 (1964) 671–678.

Mathematical Preliminaries to Elementary Catastrophe Theory

Snapping a popsicle stick back and forth depicts certain degenerate singularities.

JAN W. AUER
Brock University
St. Catharines, Ontario
Canada L2S 3A1

The purpose of this article is to introduce the nonspecialist to some of the mathematical concepts of Elementary Catastrope Theory, concepts which often occur in the form of jargon, or allusions, in the literature. This situation has helped fuel the controversy over the uses and misuses of catastrophe theory, and therefore it may be prudent to acknowledge immediately that, for this article, Elementary Catastrophe Theory is the study of singularities of parametrized families of C^{∞} real-valued functions. We shall henceforth refer to these ideas as simply "Catastrophe Theory," although other interpretations are sometimes accorded the latter phrase, depending on the user's affiliations in the current controversy.

Fortunately, the author of an introductory article need not take sides (there being little at stake) and can extol with impunity Christopher Zeeman's and René Thom's many interesting examples (if not "models") and speculations (see **[1]**, **[8]**, **[9]**). At the same time, the recent remarkable book by Tim Poston and Ian Stewart **[6]** need not compromise the most scrupulous mathematician, and contains a wealth of hard applications. The paper **[5]** by Martin Golubitsky is an excellent introduction to the subject at a more advanced level.

Definitions

Let R^n denote the space of n-tuples of real numbers. Then a real-valued function F of n variables, denoted F: $R^n \to R$, assigns to each $x \equiv (x_1,\dots,x_n) \in R^n$ a unique real number $F(x) \equiv F(x_1,\dots,x_n)$. To say that F is C^{∞} means that all partial derivatives of F of all orders exist and are continuous. For example, with $n=2$, let $F(x_1,x_2)=x_1^3+x_1x_2$; because this function is important in catastrophe theory, we shall denote the variables x_1, x_2 by x, u respectively in order to conform with the convention established below for families of functions. Thus, $F(x,u)=x^3+xu$ (see FIGURE 1). The graph of this function is a surface in R^3; in fact it is the fold catastrophe, the so-called universal unfolding of the function $f(x)=x^3$ (we will say more about this later). The first order partial derivatives of F are

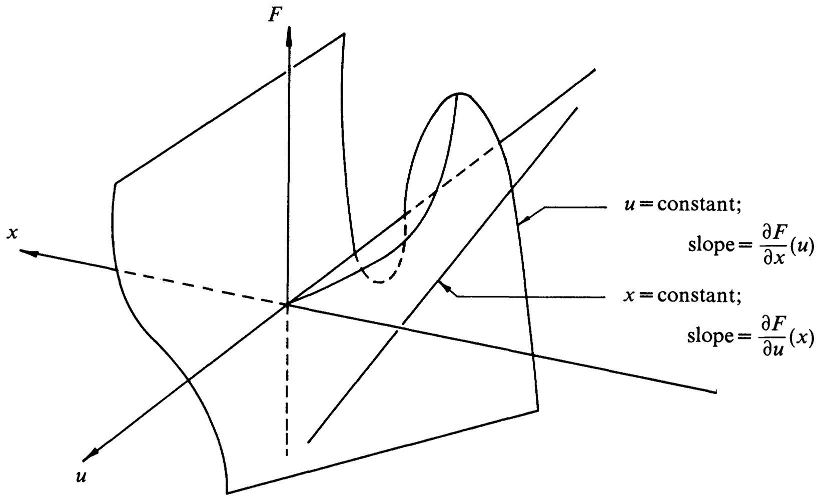
$$\frac{\partial F}{\partial x}=3x^2+u, \text{ and } \frac{\partial F}{\partial u}=x,$$

while the second order partial derivates of F are

$$\frac{\partial^2 F}{\partial x^2}=6x, \quad \frac{\partial^2 F}{\partial u^2}=0, \quad \frac{\partial^2 F}{\partial x \partial u}=\frac{\partial^2 F}{\partial u \partial x}=1.$$

Clearly, F is C^{∞}.

We say that F has a **singularity** (or **critical point**) at $x_0 \in R^n$ if its derivative at x_0, the vector $DF(x_0)=\left(\frac{\partial F}{\partial x_1}(x_0),\dots,\frac{\partial F}{\partial x_n}(x_0)\right)$, is the 0 vector. In other words, F has a singularity at x_0 whenever *all* first partial derivatives vanish at x_0; otherwise x_0 is a **regular point** of F. Thus, for example, $0=(0,0)$ is a singularity of $F(x,u)=x^3+xu$, the example of FIGURE 1.



The fold catastrophe $F(x,u)=x^3+xu$, the universal unfolding of $f(x)=x^3$.

FIGURE 1

Finally, let's consider a C^∞ function F: $R^n \times R^r \to R$. Denote points of R^r by $u \equiv (u_1,\ldots,u_r)$ and those of $R^n \times R^r$ by $(x,u) \equiv (x_1,\ldots,x_n,u_1,\ldots,u_r)$. Then, for each fixed u, one obtains a C^∞ function F_u: $R^n \to R$ by defining $F_u(x)=F(x,u)$ for $x \in R^n$. In this way F can be interpreted as a **C^∞ r-parameter family** of C^∞ functions from R^n to R; the variables $u_1,\ldots,u_r$ are called the parameters of the family. For example, $F(x,u)=x^3+xu$ is a 1-parameter family yielding $F_0(x)=x^3$ when $u=0$. The singularities of F_u are an important tool in analyzing the idea of stability of functions, to which we return later. First, however, we will see how C^∞ families arise in applications.

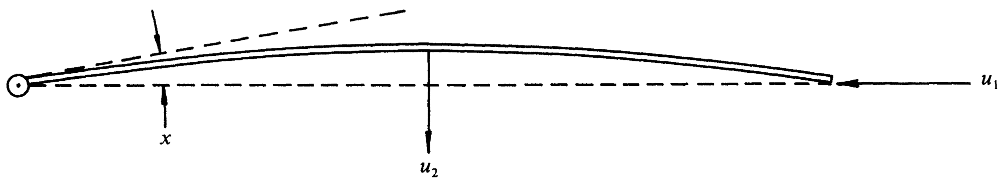
Applications

The most successful applications of catastrophe theory to date are to the analysis of discontinuous changes in systems governed by a potential energy function F. The behavior of such a system is determined by those states for which the potential function F is a minimum. An irreproachable example of such a process in the context of catastrophe theory is that of the buckling beam. To motivate the mathematics which follows, we consider briefly an elementary idealization of this example.

As almost everyone knows from experience, when a beam or column is compressed longitudinally with a smoothly increasing force, the beam will not bend until the magnitude of the applied force attains a sufficient value, say u_1: at that moment the beam "buckles." Consider, for example, a popsicle stick squeezed between the thumb and forefinger; since a traditional popsicle stick is much wider than it is thick, we can assume that it will bend only in a direction perpendicular to its width (of course the beam may break when buckled or be irreversibly deformed, but these consequences need not concern us).

Alternatively, if while applying the longitudinal buckling force u_1, we apply as well a transverse force of magnitude u_2, then the beam (popsicle stick) may suddenly "jump" to a new position: the system displays a discontinuous change in a "behavior variable" x for a smooth change in the "parameters" or "control variables" u_1 and u_2. In the terminology of catastrophe theory, the system has undergone a "catastrophic change."

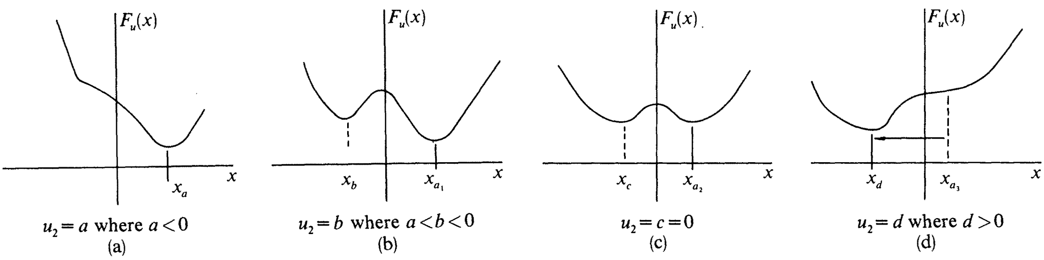
Under suitable simplifying assumptions we can analyze this behavior heuristically as follows; explicit calculations may be found in [6, p. 291]. For fixed values of the forces u_1 and u_2, the actual position assumed by the beam (see FIGURE 2), as measured by its angle of deflection x, is that which minimizes the potential energy due to the elastic properties of the beam and the



The buckling beam: whenever the longitudinal force u_1 is sufficiently large to cause the beam to "buckle," or to deflect by an angle x, a transverse force u_2 of sufficient magnitude can cause the beam to "jump" to a new (buckled) position.

FIGURE 2

applied forces. Denote the potential energy by $F(x,u_1,u_2)$, or $F_u(x)$, for short, where $u=(u_1,u_2)$. Then it is not hard to see using popsicle stick intuition, that, depending on u_1 and u_2, $F_u(x)$ may have a minimum for either one value of x or for two values, one negative and one positive. For example, when $u_2=0$ and u_1 is sufficiently small, there is one minimum at $x=0$: the beam does not bend. Now suppose that for some values of u_1 and u_2 the beam is initially displaced with angle x positive; if u_2 is large enough and negative, say $u_2=a<0$ (this corresponds to pushing up on the beam), the potential energy will have a single minimum, at $x=x_a>0$ as shown in FIGURE 3(a). As u_2 is made more positive, say $u_2=b$, a second minimum (at negative $x=x_b$) for the potential energy develops competing with the minimum at positive x, now at $x=x_{a_1}$, say, as in FIGURE 3(b). This new minimum ultimately rivals the minimum for positive x when $u_2=0$ (FIGURE 3(c)). However, not until u_2 reaches some distinct positive value d does the beam "jump" to the minimum at negative $x=x_d$. This corresponds to the disappearance of the minimum for $F_u(x)$ at positive x, represented in FIGURE 3(d). In terms of the popsicle stick, it requires some considerable force to cause the stick to jump from one side to another when $u_1>0$.



Potential functions for the buckling beam, with u_2 increasing.

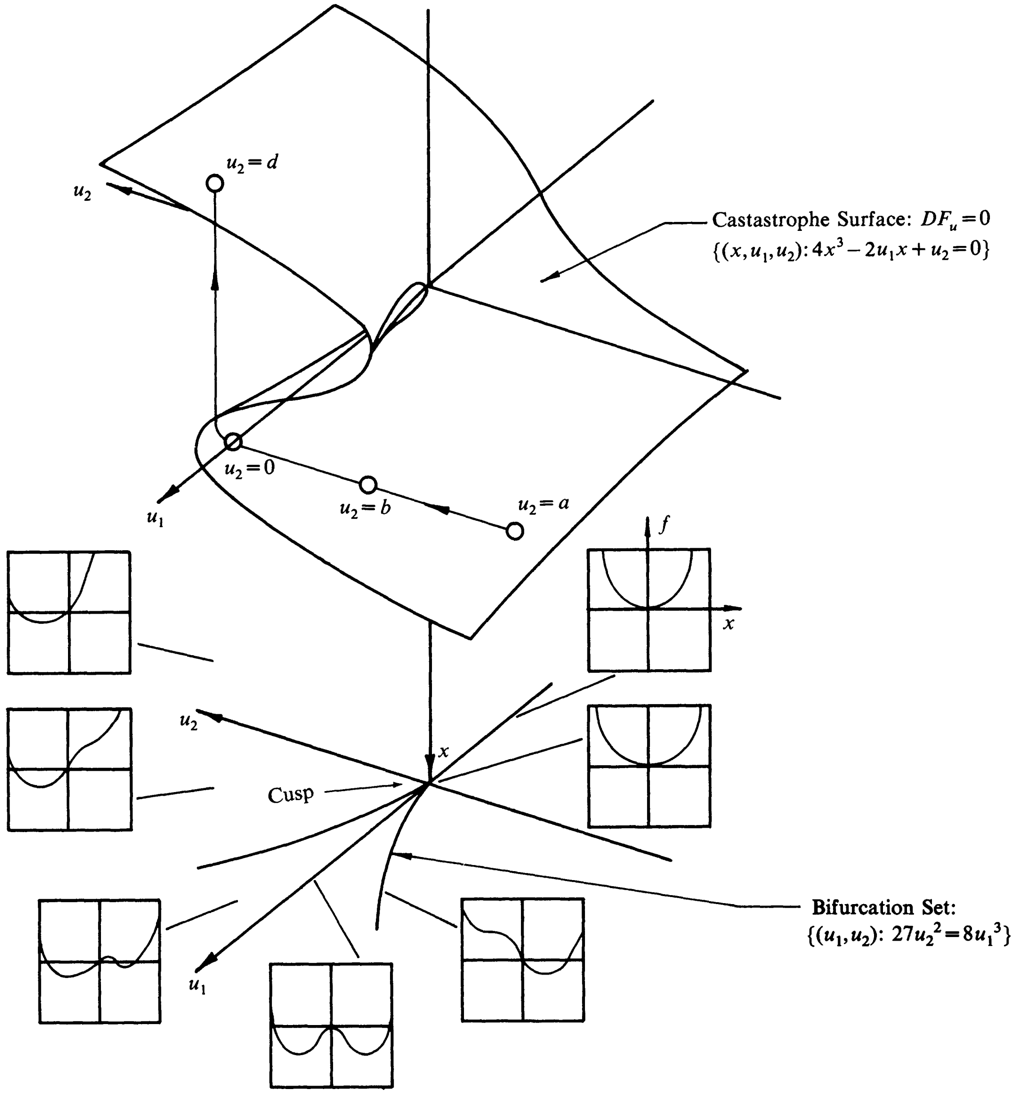
FIGURE 3

Thus, in this example, the parametrized family of C^∞ functions is $F(x,u_1,u_2)\equiv F(x,u)$, where $x\in R$ and $u\in R^2$. The set of singularities of $F_u(x)$, namely $\{(x,u)|DF_u(x)=0\}$, yields the surface in $R^3=R\times R^2$ known as the cusp catastrophe (FIGURE 4). In the popsicle stick example, the cusp is not necessarily at the origin, however. The bending process discussed above is illustrated by the path from $u_2=a$ to $u_2=d$ shown on that surface.

Morse Theory

Catastrophe theory, in some sense, generalizes certain classical results about singularities due to Whitney and Morse. We begin our formal treatment of catastrophe theory with this classical theory.

In the study of singularities there are three basic ideas: stability, classification, and genericity. To make these ideas precise, recall that in R^n, we can speak of the **distance** between two points



The cusp catastrophe $F(x, u_1, u_2) = x^4 - u_1x^2 + u_2x$.

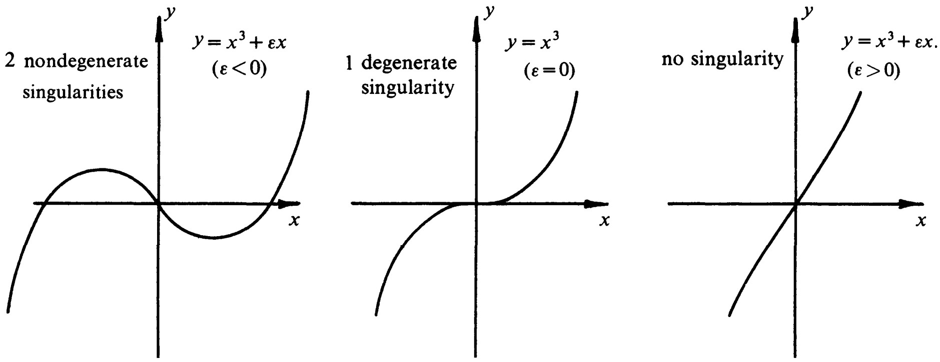
FIGURE 4

x_0, x_0', by using the usual Pythagorean formula. The **open ball** $B_\delta(x_0)$ with center x_0 and radius δ is then the set of points in R_n whose distance from x_0 is less than δ. The open balls define a **topology** on R^n, and we say that a set $U \subset R^n$ with $x_0 \in U$ is a **neighborhood** of x_0 if there is some ball $B_\delta(x_0) \subset U$. A set U is called **open** if it is a neighborhood of each of its points. To say that a property $P(x)$ holds **near** x_0 then means that there is some neighborhood of x_0 such that $P(x)$ is true when $x \in U$.

The first important classical result concerns the behavior of a function f: $R^n \to R$ near a regular point x_0. It says that *if $Df(x_0) \neq 0$ (i.e., if x_0 is a regular point of f), then on some neighborhood U of x_0 one can "change coordinates" so that f on U is just the coordinate (or projection) function π_1: $(x_1, \ldots, x_n) \to x_1$*. Specifically, this means that there is a **diffeomorphism** ϕ, that is, an invertible C^∞ function $\phi: U \to V \subset R^n$ with a C^∞ inverse, so that $f \circ \phi(x_1, \ldots, x_n) = \pi_1(x_1, \ldots, x_n) = x_1$. For example, consider $f(x_1, x_2) = x_1 + x_2^2$ at $x_0 = (0,0)$. Then $Df(x_0) = (1,0) \neq 0$.

Define $\phi: R^2 \to R^2$ by $\phi(x_1,x_2)=(x_1-x_2^2,x_2)$. Then $f\circ\phi(x_1,x_2)=f(x_1-x_2^2,x_2)=(x_1-x_2^2)+x_2^2=x_1$. Moreover, ϕ is invertible near 0 because $D\phi$, the matrix of partial derivatives $\partial\phi_i/\partial x_j$, does not have a zero determinant at x_0. In fact, $D\phi(x_0)=D\phi(0)=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$.

Since changing coordinates does not change a function qualitatively (that is, it does not change its singularities), this theorem on changing coordinates amounts to a classification of functions near their regular points. More generally, two C^∞ functions f,g: $R^n \to R^m$ are called (C^∞) **equivalent** if there are diffeomorphisms ϕ: $R^n \to R^n$ and ψ: $R^m \to R^m$ such that $\psi\circ f=g\circ\phi$. In other words, f, g are the same in suitable coordinates; *qualitatively* f, g look the same. (This notion of equivalence is the nonlinear analogue of changing bases in linear algebra: two linear transformations are equivalent if they are the same in suitable bases.) Two functions that are qualitatively different, hence inequivalent, are $f(x)=x^3$ and $g(x)=x^3+\varepsilon x$ where $\varepsilon\neq 0$: because g has either 0 or 2 critical points while f has just 1 (see FIGURE 5), there are no diffeomorphisms ψ and ϕ defined on a neighborhood of the origin so that $\psi\circ f=g\circ\phi$.



$f(x)=x^3$ is unstable.

FIGURE 5

Consider now a function f: $R^n \to R$ with singularity at $x_0 \in R^n$. The function f is said to have a **nondegenerate singularity** at x_0 if the Hessian matrix $\left[\dfrac{\partial^2 f}{\partial x_i \partial x_j}(x_0)\right]$ is nonsingular; i.e., if its determinant is not zero. Otherwise, if both the derivative vector and the determinant of the Hessian matrix vanish at x_0, then f is said to have a **degenerate singularity** there. For example, at $x_0=0$, $f(x)=x^2$ has a nondegenerate singularity, while $f(x)=x^3$ has a degenerate singularity.

We can now state the second major result of classical Morse theory: *If x_0 is a nondegenerate singularity of f, then for some $k \leqslant n$, f is equivalent in a neighborhood of x_0 to the function* $g(x_1,\dots,x_n)=-(x_1^2+\cdots+x_k^2)+(x_{k+1}^2+\cdots+x_n^2)$. In fact, k is the index of the Hessian, the number of negative elements in a diagonal representation of it; k is frequently called the **index** of the singularity. This theorem gives a classification of types of functions near a nondegenerate singularity: there is just one basic type for each $k=0,1,\dots,n$.

In fact, one can say more. Just as one can speak of "nearness" of points in R^n, so also we can define nearness of functions f,g: $R^n \to R$ (or, more generally, f,g: $R^n \to R^m$). Roughly speaking, f and g are close if for a certain nonnegative integer k the function f and its partial derivatives of order less than or equal to k are close in value to g and its corresponding partial derivatives at each $x \in R^n$. This idea can then be used to define neighborhoods, and hence a topology (the Whitney C^∞ topology: **[10]**, p. 42) on the set C^∞ (R^n) of all C^∞ functions $f: R^n \to R$. It can also be used locally (rather than globally) to identify functions that are near each other on some neighborhood, even if they are far apart elsewhere in R^n. Intuitively, g is near f in this context if it is a small perturbation of f. For example, locally (that is, near 0) $g(x)=x^3+\varepsilon x$ where $\varepsilon\neq 0$ is sufficiently small, is near $f(x)=x^3$, although this is not the case globally.

The two concepts of equivalence and nearness, both measures of similarity of functions, are related through the concept of stability. A function f is called **stable** if whenever a function g is near enough to f, then f and g are equivalent. In other words, f is stable if there is a neighborhood U_f of f such that all functions g in U_f are equivalent to f. The standard example of an unstable function is $f(x)=x^3$, because for any small $\varepsilon \neq 0$, $g(x)=x^3+\varepsilon x$ is near f, but not equivalent to f (FIGURE 5).

A function whose singularities are distinct and nondegenerate is called a **Morse** function. The main result of Morse theory is that *Morse functions are locally stable.* Specifically, if f has a nondegenerate singularity (of index k) at $x_0 \in R^n$, and if g is sufficiently close to f, then g looks like f *near* x_0: it also has a nondegenerate singularity of index k at some point x_0' near x_0.

The importance of this result depends, of course, on how many Morse functions there are. The answer is simple, and fortuitous: "almost every" function is a Morse function. The formal translation of the phrase "almost every" involves the topological notions of open and dense sets. Recall that a subset $\mathcal{G}$ of a topological space $\mathcal{F}$ (we will be interested in the case $\mathcal{F}=C^\infty(R^n)$) is called **open** if whenever f is in $\mathcal{G}$ and g is sufficiently near f, then g is in $\mathcal{G}$, and **dense** if every f in $\mathcal{F}$ is either in $\mathcal{G}$ or in its closure, i.e., has the property that there are g in $\mathcal{G}$ arbitrarily close to f. For example, "almost every" point of a closed disc in the plane is an interior point of the disc.

More generally, if in a class of functions $\mathcal{F}$ (with a topology) the set of functions $\mathcal{F}_P$ having a property P is a countable intersection of open dense sets, we call the property P **generic**. This holds, in particular, when $\mathcal{F}_P$ is itself open and dense in $\mathcal{F}$. A classical result on density is the Weierstrass Approximation Theorem: polynomials are dense in the set of continuous real-valued functions on an interval $[a,b]$ since every such function either is a polynomial or is arbitrarily close to one. However, the polynomials are not open in the space of continuous real-valued functions on $[a,b]$, so they are not generic.

We can now state precisely the last of our classical results: *The class M of Morse functions is open and dense in* $C^\infty(R^n)$. Thus, almost every f: $R^n \to R$ is a Morse function. Moreover, every stable map f: $R^n \to R$ is a Morse function: any neighborhood U of f contains some Morse function g, and if U is small enough, g will be equivalent to f, so that f too must be a Morse function. Thus, in this case, stable mappings are generic. In general, however, this is not true.

Mather's Theorem

In contrast to Morse theory, the focus of catastrophe theory is on the *degenerate* singularities. For example, we saw that $f(x)=x^3$ has a degenerate singularity at the origin, and that f is not stable. Nevertheless, we can "imbed" f in the "stable" C^∞ one-parameter family $F(x,u)=x^3+ux$ (see FIGURE 1. The concept of stability for parametrized families of functions is not quite the same as that defined above for ordinary functions, but the differences are not essential to our purposes). This imbedding of unstable functions into stable families is central to Thom's classification of functions with degenerate singularities.

Suppose f has a singularity at x_0; in other words, $Df(x_0)=0$. We assume for simplicity that $x_0=0$, the origin in R^n, and that $f(0)=0$, for we can, if necessary, always change coordinates by translation to arrange this. Also, since we are interested in the behavior of a function f near the singularity 0, we lump together all functions which coincide with f near 0 and call this set of functions the **germ** of f (at 0), denoted $\tilde{f}$. Denote by G_n the set of all germs of C^∞ functions f: $R^n \to R$ where $f(0)=0$.

We come, finally, to the idea of an unfolding suggested in FIGURE 1. An r-parameter **unfolding** of a germ $\tilde{f}$ is a germ $\tilde{F} \varepsilon G_{n+r}$ represented by an r-parameter family of functions $F: R^n \times R^r \to R$ such that $F(x,0)=f(x)$. For each (x_0,u_0) near 0, the family $\tilde{F}$ defines a germ $\tilde{F}_{u_0}$ at x_0 near $\tilde{f}$. Thus, F unfolds or deforms f into a family of germs. Henceforth, for brevity, we shall denote the germ of f by f itself. For example, $F(x,u)=x^3+ux$ is a one parameter unfolding of $f(x)=x^3$, while $F(x,u)=x^3+\sin(u_1)x+u_2x^2$ is a *2* parameter unfolding of x^3. Thus a given singularity has in general infinitely many unfoldings.

Because there are so many possible unfoldings, it is important to identify certain unfoldings that play a distinguished role. This inquiry was answered in an important theorem of J. Mather that singles out a so-called "universal" unfolding that (a) has a minimum number of parameters, (b) is stable, and (c) is qualitatively the same as any other unfolding. To describe this result we need the notion of "codimension" of a germ.

First recall that if W is an $n-r$ dimensional subspace of a (real) vector space V of dimension n, the **codimension** of W is $n-(n-r)=r$. Alternatively, it is the dimension of the quotient space V/W. Codimension is especially important in infinite dimensional spaces because, even when V and W are infinite dimensional, it can happen that dim V/W, the codimension of W, is still finite.

Because all germs in G_n satisfy $f(0)=0$, the set G_n forms a vector space that will usually be infinite dimensional. Consider, as above, a germ $f \in G_n$ that has a singularity at the origin, so $Df(0)=0$. Let $\langle \frac{\partial f}{\partial x} \rangle$ denote the subspace of G_n consisting of those germs of the form $b(x) = \sum_{i=1}^{n} a_i(x) \frac{\partial f}{\partial x_i}(x)$, where $a_i(x)$ are germs at 0. Each $b(x)$ consists of linear combinations (whose coefficients are germs) of the partial derivatives of f. Since $Df(0)=0$, each $\frac{\partial f}{\partial x_i}$ vanishes at 0; hence $b(0)=0$, so $b(x)$ really is in G_n (the space $\langle \frac{\partial f}{\partial x} \rangle$ is the ideal in the R-algebra of all germs at 0 generated by these partial derivatives). Then define **codim** $\boldsymbol{f}$, the **codimension** of the germ f, to be the dimension of the quotient space $G_n / \langle \frac{\partial f}{\partial x} \rangle$. One can show that codim f is the number of germs at 0 inequivalent to f. In many cases, computation of codim f can be accomplished using Taylor's theorem in a suitable form; we will illustrate this shortly. Now, however, we can state Mather's theorem.

THEOREM. *A germ f has a stable universal unfolding if and only if* codim f *is finite.*

In fact, if codim $f=r$ and $b_1(x), \ldots, b_r(x)$ are germs at 0 representing a basis for the quotient space $G_n / \langle \frac{\partial f}{\partial x} \rangle$, then $F(x,u)=f(x)+\sum_{i=1}^{r} b_i(x)u_i$ is a (stable) universal unfolding of f.

Consider, for example, the germ $f(x)=x^3$: $R \to R$. First, using Taylor's theorem, it is not hard to see that any C^∞ function g such that $g(0)=0$ can be expressed near 0 as $g(x)=\frac{dg}{dx}(0)x+p(x)x^2$ where $p(x)$ is a C^∞ function at 0. Secondly, $\frac{\partial f}{\partial x}=\frac{df}{dx}=3x^2$, so g represents the element $[g]$ of $G_1 / \langle \frac{\partial f}{\partial x} \rangle$ obtained by forgetting about terms involving x^2: i.e., $[g]=\frac{dg}{dx}(0)x$, a real multiple of x. Thus the vector space $G_1 / \langle \frac{\partial f}{\partial x} \rangle$ has basis x, and thus dimension 1. By Mather's theorem $f(x,u)=x^3+xu$ is a stable universal unfolding of $f(x)=x^3$.

It is not hard to show that f is stable (near 0) if and only if codim $f=0$, so in a sense codim f measures the instability of f. In particular, any Morse function has codimension 0.

Thom's Theorem

We come now to Thom's celebrated list of the seven elementary catastrophes (see TABLE 1). This theorem classifies germs of codimension at most 4 with degenerate singularities. The seven types are called the elementary catastrophes. By Mather's theorem, such germs have stable unfoldings with at most four parameters. These singularities are potentially important because, as we have seen, the parameters in an unfolding correspond to control variables, and if these are space and time coordinates, no more than four are needed. However, in many examples the controls are not space and time, and generally more than four are needed (see, for example, [6] for further details).

Germ f	Codim f	Universal Unfolding $F(x,u)$	Name
x^3	1	x^3+u_1x	Fold
x^4	2	$x^4-u_1x^2+u_2x$	Cusp
x^5	3	$x^5+u_1x^3+u_2x^2+u_3x$	Swallowtail
x^3+y^3	3	$x^3+y^3+u_1xy-u_2x-u_3y$	Hyperbolic Umbilic
x^3-xy^2	3	$x^3-xy^2+u_1(x^2+y^2)-u_2x-u_3y$	Ellipitic Umbilic
x^6	4	$x^6+u_1x^4+u_2x^3+u_3x^2+u_4x$	Butterfly
x^2y+y^4	4	$x^2y+y^4+u_1x^2+u_2y^2-u_3x-u_4y$	Parabolic Umbilic

Thom's Seven Elementary Catastrophes

TABLE 1

Thom's theorem says, specifically, that any singular germ of codimension at most 4 is qualitatively the same as one of the germs (with corresponding stable unfoldings) listed in TABLE 1. The unfoldings are unique up to the addition of a nondegenerate quadratic in other variables, and multiplication by $+1$ or -1. Alternatively, Thom's theorem is an exhaustive list (up to equivalence) of all stable unfoldings with four or fewer parameters: any other unfolding must be equivalent to one in the list.

The germs which appear in the list have at most two "behavior" variables (with (x_1,x_2) denoted there by (x,y)). This is because any degenerate singular germ f: $R^n \to R$ with $n \geqslant 3$ necessarily has codimension of at least six, a result that can be proved by considering the rank of the Hessian matrix of f at 0.

Finally, it can be proved that the set of C^∞ r-parameter families F which produce stable unfoldings F_u at each $u \in R^r$ is open and dense in the set of all r-parameter families, provided $r \leqslant 5$. This is no longer true when $r \geqslant 6$. In particular, it follows that the property of being a "Thom unfolding" is generic: typically, an unfolding must be one in the list of seven when $r \leqslant 4$.

For Thom the significance of the stability of the unfoldings for physical applications is that for events in nature to be observable, they must be capable of repetition in experiments, in spite of the perturbations due to the fact that conditions can never be reproduced exactly. Thus, the family of functions which generate the events must be stable. However, in assessing Thom's work, it is important to realize that there are other notions of stability, and models of processes that lack stability which agree with experimental observation [**6**].

The author wishes to thank the referees and editors for their valuable suggestions.

References

Here is a sample of introductory sources on Elementary Catastrophe Theory, listed in approximate order of increasing mathematical difficulty. For further references, see [**6**]: it contains a virtually complete guide to the literature of catastrophe theory.

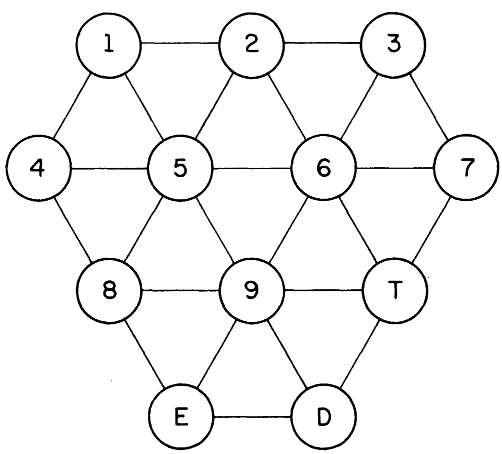
[1] E. C. Zeeman, Catastrophe theory, Sci. Amer. (April 1976), 65–83.
[2] A. E. R. Woodcock and M. Davis, Catastrophe Theory, Dutton, 1978.
[3] Tim Poston and I. N. Stewart, Taylor Expansions and Catastrophes, Pitman, 1976.
[4] J. Callahan, Singularities and plane maps, I, II, Amer. Math. Monthly, vol. 81, no. 3 (1974) and vol. 84, no. 10 (1977).
[5] M. Golubitsky, An introduction to catastrophe theory and its applications, SIAM Rev., vol. 20, no. 2 (1978).
[6] Tim Poston and I. N. Stewart, Catastrophe Theory and its Applications, Pitman, 1978.
[7] Th. Bröcker and L. Lander, Differentiable Germs and Catastrophes, Cambridge Univ. Press, New York, 1975.
[8] E. C. Zeeman, Catastrophe Theory: Selected Papers (1972–1977) Addison-Wesley, 1977.
[9] René Thom, Structural Stability and Morphogenesis (transl. by D. H. Fowler), Benjamin-Addison-Wesley, 1975.
[10] M. Golubitsky and V. Guilleman, Stable Mappings and Their Singularities, Springer-Verlag, 1973.

Never Rush to Be First in Playing Nimbi

AVIEZRI S. FRAENKEL
The Weizmann Institute of Science
Rehovot, Israel

HANS HERDA
Boston State College
Boston, MA 02115

Nimbi is a two-player board game invented by Piet Hein. Twelve tokens, placed on the twelve vertices of a hexagonal board numbered $1, 2, \ldots, 9, T, E, D$, constitute the initial position of Nimbi (see FIGURE 1). The board contains twelve **rows**: four horizontal rows, four diagonal rows in northeast to southwest direction and four diagonal rows in northwest to southeast direction. The two players play alternately. Each player at his turn can remove any contiguous portion from a single row. (Thus 159D may be removed in one move. But if 5 had been removed previously, then neither 19 nor 1D nor 19D can be removed in a single move, though 1 or 9 or D or even 9D can.) In Piet Hein's version, the player making the last move is the loser, his opponent the winner. We call this version LPL-Nimbi. The other version, in which the player making the last move wins and his opponent loses, will be called LPW-Nimbi.



The Nimbi Board

FIGURE 1

Both versions of Nimbi are **combinatorial games**, which, for our purposes, are defined to comprise all finite two-player 0-1 games (finite: finite number of positions; 0-1: outcomes are lose and win only) with perfect information (unlike some card games where information is hidden) and without chance moves (no dice), in which the players play alternately. A combinatorial game is **last player losing** (LPL) if the player first unable to move wins, and it is **last player winning** (LPW) if the player first unable to move loses. In every combinatorial game, either the first or the second player has a winning strategy. A simple proof of this fact, due to Steinhaus, is given by Kac [7]. Since it is very short, we reproduce it here. Denote by $a_1, a_2, \cdots$ and by $b_1, b_2, \cdots$ the moves of Al and Beth, respectively. Suppose that Al makes the first move. The fact that Beth has a winning strategy can be expressed symbolically as follows:

$$(\forall a_1)(\exists b_1)(\forall a_2)(\exists b_2)\cdots(\forall a_n)(\exists b_n) \quad \text{Beth wins.}$$

The negation of this statement is obtained by the familiar De Morgan's rule, and it reads:

$$(\exists a_1)(\forall b_1)(\exists a_2)(\forall b_2)\cdots(\exists a_n)(\forall b_n) \quad \text{Beth does not win.}$$

This, however, is clearly the statement that Al has a winning strategy, and the proof is complete.

Nimbi

Nimbi has a long and dramatic history. It is the last replique in a dialogue down the ages.

Probably the oldest game in the world is Nim, which orginated in the Orient thousands of years ago. It was played with the most simple material: 12 stones usually placed in heaps of 3, 4, and 5. Two players took turns making a move. A move consisted in removing from any one of the heaps as many and as few stones as one wished, i.e., at least one stone and at most the whole heap. The aim was to force the opponent to take the last stone.

Simple in principle but difficult to master.

It has entertained people all over the world for thousands of years and kept them groping for a general principle to reveal the right moves in each situation. In 1901, the French-American mathematician Charles Leonard Bouton, succeeded—by means of a subtle analysis—to find a very simple principle, applicable by anyone, telling you whether a situation was lost or won and in the latter case which move or moves would ensure you the final victory. So the ancient game was turned into a beautiful mathematical solution but was destroyed as a game. This destruction was taken up as a challenge by the Danish author, scientist and inventor, Piet Hein, who set himself the task to revive Nim and give it back its old dignity as an unconquered game. And this by means of a change that should not make it less simple in principle but should bring it outside the reach of the analysis of Charles Leonard Bouton.

Half a century after Bouton's assassination of Nim, Piet Hein succeeded in this strange and difficult task, the greatest difficulty being to save the simplicity in principle, at the same time making it unconquerable by analysis. Piet Hein's new principle was made the topic of an article by Martin Gardner in "Scientific American" and in one of his books about Mathematical Puzzles and Diversions. During a couple of decades mathematicians have tried to destroy even this new game, attempting to find a general principle that would cover all versions of it with varying numbers of stones, as Bouton's analysis did in respect of Nim. Their efforts hitherto have been in vain and there are considered to be fair chances that Piet Hein has succeeded so thoroughly that his game will never be destroyed. For the principle, if ever one is found, is likely to be so complicated that the actual experimenting in the single situations proves to be simpler; meaning, that the playing of it is not beaten by theory, thus forever reassuring its position as a game.

—*from the game brochure*

Since both LPL-Nimbi and LPW-Nimbi are combinatorial games, either the first or the second player must have a winning strategy in each.

The purpose of this note is to prove the following more specific (and perhaps surprising) result.

THEOREM. *In both LPL-Nimbi and LPW-Nimbi, the second player can force a win.*

This theorem is a "rare event" because almost all combinatorial games (whether LPL or LPW) are games in which the first player can force a win (Singmaster [**8**]). More precisely, if $F(n)$ is the number of combinatorial games of length not exceeding n for which the first player has a winning strategy and $N(n)$ is the total number of combinatorial games of length not exceeding n (where the length of a game is the number of moves from beginning to end), then $\lim_{n\to\infty} F(n)/N(n)=1$.

Before proving the theorem we shall summarize briefly some of the basic concepts and vocabulary of combinatorial games. Let Γ be a combinatorial game, S its set of positions, and a, b two positions in S. Then b is a **follower** of a if there is a move from a to b. A position with no follower is **terminal**, while an initial position (a position without predecessor) is called a **source**. A position q is called an ***N*-position** if the Next player can force a win from q irrespective of the moves of his opponent, or a ***P*-position** if the Previous player can force a win from q irrespective of the moves of his opponent. This dichotomy partitions the set S of all positions into the subset N of N-positions and the subset P of P-positions, where the partition depends on whether Γ is LPL or LPW. It follows that a position is an N-position if and only if it has a follower in P, and it is a P-position if and only if all its followers are in N. It thus appears that P-positions are relatively rare. Singmaster's result is a precise statement of this fact. Our analysis of Nimbi rests on the determination of a certain subset of P-positions.

First, however, we must define a so-called **Sprague-Grundy function** g which maps the set of positions S into the set of nonnegative integers. For each position $a\in S$, the Sprague-Grundy number $g(a)$ is defined to be the smallest nonnegative integer not appearing in the set $\{g(b)\}$ of all followers b of a. Thus, in particular, $g(a)=0$ if a is terminal. The importance of g for combinatorial games stems from the following two facts:

I. $P=\{a\in S: g(a)=0\}$ if Γ is an LPW-game.
II. The g-value of a position in a "disjunctive sum" game is the "nim-sum" of the g-values of the individual positions.

The second fact needs some explanation. Suppose that two players play a game Γ consisting of a finite collection of disjoint combinatorial games $\Gamma_1,\Gamma_2,\dots,\Gamma_m$, also called components, where each player at his turn selects some component Γ_i and makes a move in it. Then the game Γ is called the **disjunctive sum** of the games $\Gamma_1,\Gamma_2,\dots,\Gamma_m$. The g-value of a position in Γ is, according to II, the nim-sum of the g-values of the positions in the components Γ_i. To find the nim-sum, write each g-value to the base 2 as $\Sigma a_r 2^r$, then add the a_r's modulo 2 for each value of r without carrying to obtain a binary sum. For example, the nim-sum of 1, 2 and 3 is 0 (since $1_2\oplus 10_2\oplus 11_2=00_2$, where $\oplus$ denotes nim-sum) and the nim-sum of 3 and 6 is 5. The nim-sum of two numbers is 0 if and only if the numbers are the same. (For these and related facts about combinatorial games, see Conway [**1**], Smith [**9**] and Fraenkel [**3**].)

We are now ready to analyze LPL-Nimbi, using a position catalogue (TABLE 1a, b) containing a set of P-positions large enough to prove that the second player in LPL-Nimbi can always force a win. In TABLE 1 a position such as 127T is really just a sample of position number 1, because 23ED and other instances represent the same position. The names in TABLE 1 are designed to help in recognizing the shape of different samples representing the same position.

We have to verify that each position in TABLE 1a, b is a P-position. This is illustrated for position number 19 of TABLE 1b: We summarize in TABLE 2 the 32 possible moves the Next

Position Number	Sample Position	Name
1	127T	two 2's
2	2478TD	1, 2, straight 3
3	1378TE	1, 2, crooked 3
4	2578TD	two straight 3's
5	13468T	two crooked 3's
6	13479ED	two 2's and triangle
7	4578TED	small kite
8	134578TED	big kite
9	123456789TED	full board

(a)

Position Number	Sample Position	Name
10	1	one dot
11	1TE	three dots
12	19ED	dot and triangle
13	2689	small bucket
14	349TED	two dots and rhombus
15	178TED	dot and sled
16	24679D	dot and horse
17	124679	wrench
18	1234567	span
19	1245679TD	span with two triangles

(b)

Position Number	Sample Position	Name
20	(empty)	void
21	1T	two dots
22	9ED	triangle
23	278D	four dots
24	349ED	two dots and triangle
25	29TED	dot and rhombus
26	1568T	horse
27	12359	letter *A*
28	137TED	dot and big bucket
29	125679TD	span with triangle

(c)

A catalogue of *P*-positions (previous player winning positions) in the game of Nimbi. Those listed in (a) are *P*-positions for both LPL (last player losing) and for LPW (last player winning) versions of the game; those listed in (b) are *P*-positions only for LPL-Nimbi, while those in (c) are *P*-positions only for LPW-Nimbi.

TABLE 1

player can make by listing the locations he vacates (grouped by type), followed by a *P*-position number from TABLE 1a, b which the Previous player can attain by his countermove. Since the Previous player can reach a *P*-position in each of the 32 cases, position number 19 is indeed in *P*. To complete the proof, the reader should make a similar verification for each of the remaining *P*-positions of TABLE 1a, b. The full board is position number 9. To see that it is a *P*-position in LPL-Nimbi, verify that any move from position number 9 can be countered by a move to one of the *P*-positions numbered 2, 7, 8, 18 or 19. This completes the proof that LPL-Nimbi is a second player winning game.

In FIGURE 2 we present a strategy graph whose vertices are *P*-positions of LPL-Nimbi. Two vertices *a* and *b* are joined by a downward directed edge (*a*, *b*) if for some move of the Next player from *a*, the Previous player can respond by moving to *b*. Since all possible moves from

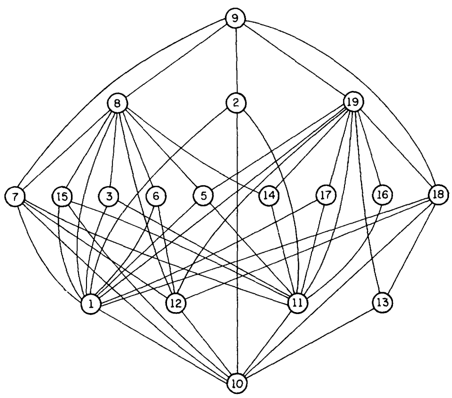
Type	Next Move	Attainable *P*-position	Type	Next Move	Attainable *P*-position	Type	Next Move	Attainable *P*-position
→4	4567	12	↗2's	TD	17	singles	D	1
↖4	159D	12		7T	13		T	16
→3's	567	12		69	14		9	1
	456	1		25	1		7	18
↗3	7TD	1		14	12		6	14
↖3's	59D	12					5	17
	159	11	↖2's	9D	12		4	18
	26T	12		59	13		2	12
→2's	9T	14		15	16		1	12
	67	11		6T	11			
	56	5		26	1			
	45	13						
	12	12						

The 32 possible moves the Next player can make from position number 19 and their rebuttals.

TABLE 2

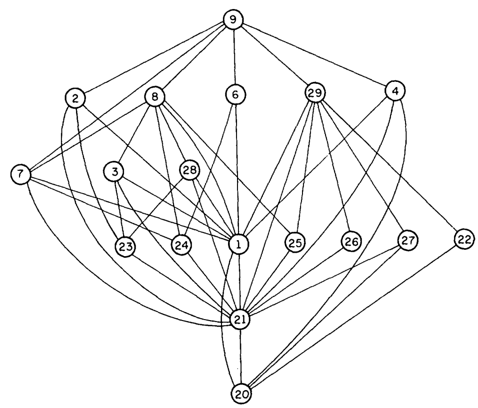
each P-position are taken into account, the graph gives enough information to enable the second player to win in all cases. Position 9 on top is the source, while position 10 at the bottom is the terminal position of the graph. The vertices of the strategy graph are precisely the P-positions of TABLE 1a, b, except for position 4 which is not used. If the second player adheres to the strategy indicated by this graph, player 1 can delay losing for at most eight moves. A possible sequence of such moves, omitting intervening N-positions, is $9 \to 8 \to 3 \to 1 \to 10$.

We turn now to LPW-Nimbi. In addition to a direct analysis via P-positions, as we just did for LPL-Nimbi, LPW games can be analyzed by using property I of the g-function. Also, property II can be used for those positions which can be decomposed into a disjunctive sum of smaller positions. It is easy to verify that the P-positions numbered 1 through 8 (but not 10 through 19) of LPL-Nimbi in TABLE 1 have g-value 0 and are, therefore, also P-positions in LPW-Nimbi. A list of ten additional P-positions of LPW-Nimbi (which are not P-positions of



A Strategy Graph for LPL-Nimbi

FIGURE 2



A Strategy Graph for LPW-Nimbi

FIGURE 3

LPL-Nimbi) is given in TABLE 1c. To verify that these are indeed P-positions, we could either check that the g-value of each position in TABLE 1c is 0, or use a process analogous to that used to prove that position number 19 is in P for LPL-Nimbi. To see that position number 9 is in P for LPW-Nimbi, verify that any move from it can be countered by a move to one of the positions 2, 4, 6, 7, 8, or 29. This completes (an outline of) the proof that LPW-Nimbi is a second-player winning game.

FIGURE 3 depicts a strategy graph for LPW-Nimbi, which is constructed analogously to the strategy graph of FIGURE 2. Position 9 is the source and position 20 is the terminal position. (Position 5 does not appear.) If the second player sticks to the strategy indicated by the graph, the first player can delay losing for at most ten moves, a possible sequence of moves being indicated by the P-positions $9 \to 8 \to 3 \to 23 \to 21 \to 20$.

For the interested reader we mention finally that LPL games are in general less tractable than LPW games. See, for example, Conway [1, Ch. 12] and Grundy and Smith [5]. Ferguson [2] found a subclass of tractable LPL games for which there is a winning strategy which is only a slight modification of the winning strategy of their LPW versions. It turns out that LPL-Nimbi is not in this class: Ferguson's condition A3 (which is also necessary) requires that if x is a component with g-value 1 and if y is a follower of x with g-value 0, then every component of y has g-value 0 or 1. The empty board (terminal position) has g-value 0. An isolated token has, therefore, g-value 1, a connected pair (like 37 or ED) has g-value 2. Also $g(137TED) = 0$, since position 28 in TABLE 1c has g-value 0 by property I. Since $g(1) = 1$, the nim-sum yields $g(37TED) = 1$. Since the follower 37ED, whose components are 37 and ED, satisfies $g(37ED) =$

$g(37) \oplus g(\mathrm{ED}) = 2 \oplus 2 = 0$, condition A3 does not hold. We remark that even LPW-Nimbi, played on a board of arbitrary size, appears to be much harder than the classical LPW games, like those of Guy and Smith [6], because LPW-Nimbi is not a disjunctive sum of disjoint combinatorial games.

Acknowledgments

The *P*-positions of the LPW-version of a game similar to Nimbi, but played on a rectangular board and without the row contiguousness condition, were computed by R. B. Eggleton, A. S. Fraenkel and B. Rothschild in 1973 for all $2 \times n$ rectangles and $3 \times m$ $(m \leqslant 5)$ rectangles. They called the game 2-dimensional Nim, did not publish the results, and were unaware that D. Fremlin [4] had examined this game, which he called Nim-squared, at about the same time and, using a computer, computed all *P*-positions which fit into a 4×4 square for both the LPL and the LPW version. S. -Y. R. Li notified us that 2-dimensional Nim is also being played on a triangular board.

We thank the editors and the referees for their helpful reorganizing and editing work. Hans Herda wishes to thank the Weizmann Institute of Science where this work was done. Finally, A. S. Fraenkel wishes to thank his son Abraham, age 12, for his help in checking out TABLE 1.

References

[1] J. H. Conway, On Numbers and Games, Academic Press, London, 1976.

[2] T. S. Ferguson, On sums of graph games with last player losing, Internat. J. Game Theory, 3 (1974) 159–167.

[3] A. S. Fraenkel, From Nim to Go, Proc. Symposium on Combinatorial Math. and Optimal Design, J. N. Srivastava, editor (Colorado State Univ. 1978), North-Holland, The Netherlands, to appear.

[4] D. Fremlin, Well-founded games, Eureka, 36 (1973) 33–37.

[5] P. M. Grundy and C. A. B. Smith, Disjunctive games with last player losing, Proc. Cambridge Philos. Soc., 52 (1956) 527–533.

[6] R. K. Guy and C. A. B. Smith, The *g*-values of various games, Proc. Cambridge Philos. Soc., 52 (1956) 514—526.

[7] M. Kac, Hugo Steinhaus—a reminiscence and a tribute, Amer. Math. Monthly, 81 (1974) 572–581.

[8] D. Singmaster, Almost all games are first-person games, Private communication.

[9] C. A. B. Smith, Compound games with counters, J. Recr. Math., 1 (1968) 67–77.

Solving an Exponential Equation

KENNETH ROGERS
University of Hawaii
Honolulu, HI 96822

Gerald A. Heuer in [2] studied the equation

$$a^x = \left(\frac{a+x}{2}\right)^{(a+x)/2}, \tag{1}$$

showing that for each a greater than e there is a unique solution with $x > a$. He gave numerical results, upper and lower estimates both asymptotic to $2a^2 - 2a\ln a - a$, and noted the integer solution $4^{12} = 8^8$. It turns out that this equation can be changed to the form

$$\frac{\ln u}{u} = \frac{\ln a}{a} \tag{2}$$

$g(37) \oplus g(\mathrm{ED}) = 2 \oplus 2 = 0$, condition A3 does not hold. We remark that even LPW-Nimbi, played on a board of arbitrary size, appears to be much harder than the classical LPW games, like those of Guy and Smith [6], because LPW-Nimbi is not a disjunctive sum of disjoint combinatorial games.

Acknowledgments

The *P*-positions of the LPW-version of a game similar to Nimbi, but played on a rectangular board and without the row contiguousness condition, were computed by R. B. Eggleton, A. S. Fraenkel and B. Rothschild in 1973 for all $2 \times n$ rectangles and $3 \times m$ ($m \leqslant 5$) rectangles. They called the game 2-dimensional Nim, did not publish the results, and were unaware that D. Fremlin [4] had examined this game, which he called Nim-squared, at about the same time and, using a computer, computed all *P*-positions which fit into a 4×4 square for both the LPL and the LPW version. S. -Y. R. Li notified us that 2-dimensional Nim is also being played on a triangular board.

We thank the editors and the referees for their helpful reorganizing and editing work. Hans Herda wishes to thank the Weizmann Institute of Science where this work was done. Finally, A. S. Fraenkel wishes to thank his son Abraham, age 12, for his help in checking out TABLE 1.

References

[1] J. H. Conway, On Numbers and Games, Academic Press, London, 1976.
[2] T. S. Ferguson, On sums of graph games with last player losing, Internat. J. Game Theory, 3 (1974) 159–167.
[3] A. S. Fraenkel, From Nim to Go, Proc. Symposium on Combinatorial Math. and Optimal Design, J. N. Srivastava, editor (Colorado State Univ. 1978), North-Holland, The Netherlands, to appear.
[4] D. Fremlin, Well-founded games, Eureka, 36 (1973) 33–37.
[5] P. M. Grundy and C. A. B. Smith, Disjunctive games with last player losing, Proc. Cambridge Philos. Soc., 52 (1956) 527–533.
[6] R. K. Guy and C. A. B. Smith, The *g*-values of various games, Proc. Cambridge Philos. Soc., 52 (1956) 514—526.
[7] M. Kac, Hugo Steinhaus—a reminiscence and a tribute, Amer. Math. Monthly, 81 (1974) 572–581.
[8] D. Singmaster, Almost all games are first-person games, Private communication.
[9] C. A. B. Smith, Compound games with counters, J. Recr. Math., 1 (1968) 67–77.

Solving an Exponential Equation

KENNETH ROGERS
University of Hawaii
Honolulu, HI 96822

Gerald A. Heuer in [2] studied the equation

$$a^x = \left(\frac{a+x}{2}\right)^{(a+x)/2}, \tag{1}$$

showing that for each a greater than e there is a unique solution with $x > a$. He gave numerical results, upper and lower estimates both asymptotic to $2a^2 - 2a\ln a - a$, and noted the integer solution $4^{12} = 8^8$. It turns out that this equation can be changed to the form

$$\frac{\ln u}{u} = \frac{\ln a}{a} \tag{2}$$

which is easier to handle and leads to a series solution by a method of Lagrange and Jacobi that deserves greater popularity. As we will show, this series gives better estimates and explains why Heuer's lower bound is so close.

Setting $(a+x)/2=y$ in (1) leads to $(2y-a)\ln a=y\ln y$, collecting terms gives $(y/a)\ln(a^2/y)=\ln a$, and putting $u=a^2/y$ yields (2). So equation (1) is equivalent to (2), where $x=(2a^2/u)-a$. Now $(\ln u)/u$ increases from 0 to $1/e$ on the range $1\leqslant u\leqslant e$, then it decreases and tends to zero as u goes to infinity. So each value t in the range $0<t<1/e$ is assumed exactly twice by $(\ln u)/u$: once with $u>e$ and once with $1<u<e$. This proves, since $x=(2a^2/u)-a$, that there is a unique x satisfying equation (1) with $a<x<2a^2-a$. Solving (2) is easy with a calculator: I took $a=60$, used my SR-50 and quickly got $u=1.076203, x=6630.188$, compared to Heuer's 6630.187. Similarly $a=7000$ led to $u=1.001267$; Newton's method yields greater accuracy: $u=1.001267214$ and $x=97868970.2$.

The method of Lagrange and Jacobi solves (2) for u as a power series in $t=(\ln u)/u$, but it also gives the series for $1/u$ needed to find x. Given a power series $t=\Sigma_1^\infty a_n w^n$, where $a_1\neq 0$, then any (convergent) power series $g(w)=\Sigma_0^\infty c_n w^n$ can be expanded as $g(w)=\Sigma_0^\infty p_n t^n$, where $p_0=c_0$ and np_n is the coefficient of w^{-1} in the expansion of $g'(w)/t^n$. (This expansion is very clearly presented in Bromwich [1].) To apply this method to (2), set $u=e^w$ (so $t=we^{-w}$) and let $g(w)=e^{cw}$. Thus $p_0=1$ and np_n is the coefficient of w^{-1} in the expansion of $cw^{-n}e^{(c+n)w}$; hence

$$e^{cw}=\sum_{n=0}^{\infty}\frac{c(c+n)^{n-1}}{n!}t^n,\qquad \text{where}\quad t=we^{-w}. \tag{3}$$

This series converges if $|t|<1/e$, the bound noted above for $\ln u/u$. (This result is one case of Ex. 4, p. 160 of [1].) Applied with $c=1$ and $c=-1$, equation (3) gives series for u and $1/u$ respectively:

$$u=\sum_{n=0}^{\infty}\frac{(n+1)^{n-1}t^n}{n!}=1+t+\frac{3t^2}{2!}+\frac{4^2t^3}{3!}+\frac{5^3t^4}{4!}+\cdots \tag{4}$$

$$\frac{1}{u}=\sum_{n=0}^{\infty}\frac{-(n-1)^{n-1}t^n}{n!}=1-t-\frac{t^2}{2!}-\frac{2^2t^3}{3!}-\frac{3^3t^4}{4!}-\cdots. \tag{5}$$

Now we can use (5) to find x:

$$\begin{aligned}x=(2a^2/u)-a&=2a^2\sum_{n=0}^{\infty}\frac{-(n-1)^{n-1}}{n!}\left(\frac{\ln a}{a}\right)^n-a\\&=2a^2-2a\ln a-a-(\ln a)^2-4(\ln a)^3/3a-\cdots.\end{aligned} \tag{6}$$

To estimate the series of logarithmic terms, set $c_n=(n-1)^{n-1}/n!$ and work with the ratio c_n/c_{n+1}:

$$c_n/c_{n+1}=\left(1-\frac{1}{n}\right)^{n-1}\left(1+\frac{1}{n}\right)>\left(1-\frac{1}{n}\right)^{n-1}>e^{-1}\qquad \text{for}\quad n\geqslant 2,$$

because

$$\ln\left(1-\frac{1}{n}\right)=-\Sigma 1/kn^k>-\Sigma 1/n^k=-1/(n-1).$$

Thus $c_{n+1}<ec_n$ if $n\geqslant 2$, and so $c_n<c_k e^{n-k}$ for all $n>k$. Hence

$$c_k t^k<\sum_{n=k}^{\infty}c_n t^n<\sum_{n=k}^{\infty}e^{n-k}c_k t^n=\frac{c_k t^k}{1-te}, \tag{7}$$

where $c_n t^n=[(n-1)^{n-1}/n!][(\ln a)/a]^n$ is the nth term in the series expansion (6) of x. Hence we can use (7) to approximate x arbitrarily closely. Heuer's upper estimate is $2a^2-2a\ln a-a$, so its distance from x exceeds $(\ln a)^2$. His lower estimate is $b_2=2a^2-a-(2+\varepsilon)a\ln a$, where $\varepsilon=2/(d+\sqrt{d^2-3})$ and $d=(1-2t)/t$. It isn't hard to show that $t+2t^2<1/d<\varepsilon=t+2t^2+$

$O(t^3)$, and so $x-b_2=O((\ln a)^3/a)$. This explains why b_2 is such a good estimate. In fact $b_2<2a^2-2a\ln a-a-(\ln a)^2-2(\ln a)^3/a$, and this is a better estimate as it is still less than x for $a\geqslant 18$. Using (6) and (7) with $a=40$ I had to go up to $k=7$ in order to obtain Heuer's tabulated value $x=2849.292$.

Finally we show that the only solution of (1) in integers a,x with $x>a>0$ is $a=4, x=12$. First note that $x>(a+x)/2>a$, so (1) implies that $a>1$. Then if p is a prime factor of a, it follows from the exponents of p in equation (1) that if p^s divides a, then p^{s+1} divides $(a+x)/2$. Hence if p^s is the highest power of p dividing a, p^{s+1} will divide $a+x$, so p^s must also be the highest power dividing x. Hence $x=ra$ where r and a have no common factors. Since a and x are of the same parity, it follows that r is odd. If $r=2k-1$, then $x=(2k-1)a$, so $(a+x)/2=ak$. So equation (1) becomes $a^{2ka-a}=(ak)^{ak}$, yielding $a^{k-1}=k^k$. The same exponent argument shows that k divides a, so $a=hk$. Thus $(hk)^{k-1}=k^k$ so $h^{k-1}=k\leqslant 2^{k-1}$, with strict inequality for $k>2$. Hence, since $k>1$, we must have $h=2=k$, $a=4$ and $x=ra=3a=12$.

References

[1] T. J. I. Bromwich, An Introduction to the Theory of Infinite Series, Macmillan, London, 1947.

[2] G. A. Heuer, Midpoint solutions of $x^x=a^b$, this MAGAZINE, 51 (1978) 181–183.

A Weighted Voting Model

ALEX KLEINER
Drake University
Des Moines, IA 50311

In a weighted voting body the power an individual member controls is rarely proportional to the number of votes that the member casts. Historically, two different methods for measuring power in a weighted voting body, the Shapley-Shubik index and the Banzhaf index, have found acceptance among mathematicians and in the courts. The survey paper [7] by W. F. Lucas provides an introduction (with extensive references) to both indices. More recently Straffin ([12], [13]) and others have proposed a generalized index which includes both the Shapley-Shubik and Banzhaf indices as special cases. In this note we apply Straffin's ideas to a special type of weighted voting body. Specifically, we are interested in the situation where one player has some number of votes greater than one and all others have one vote each. Voting bodies of this type can arise in many settings. Lucas cites instances from national and international governing bodies including the United Nations General Assembly, a 138-player game in which the Soviet Union has 3 votes [7]. In [6] Kleiner analyzes a board of this type which supplies services for a metropolitan region consisting of one large city, several suburban communities, and a rural area dotted with small towns. Committees with an even number of members in which each member has one vote and the chair has the additional power to break ties form a third class of examples. (Recent papers on power indices are listed in the references; see especially the papers in [9] and the references in [5].)

We will show that if any player observes that the other players behave in a homogeneous fashion, then that player is better off adopting such behavior rather than remaining independent. Although this result is what one might expect, we will obtain a nonintuitive result concerning the share of power of the player with "extra" votes.

$O(t^3)$, and so $x-b_2=O((\ln a)^3/a)$. This explains why b_2 is such a good estimate. In fact $b_2<2a^2-2a\ln a-a-(\ln a)^2-2(\ln a)^3/a$, and this is a better estimate as it is still less than x for $a\geqslant 18$. Using (6) and (7) with $a=40$ I had to go up to $k=7$ in order to obtain Heuer's tabulated value $x=2849.292$.

Finally we show that the only solution of (1) in integers a,x with $x>a>0$ is $a=4, x=12$. First note that $x>(a+x)/2>a$, so (1) implies that $a>1$. Then if p is a prime factor of a, it follows from the exponents of p in equation (1) that if p^s divides a, then p^{s+1} divides $(a+x)/2$. Hence if p^s is the highest power of p dividing a, p^{s+1} will divide $a+x$, so p^s must also be the highest power dividing x. Hence $x=ra$ where r and a have no common factors. Since a and x are of the same parity, it follows that r is odd. If $r=2k-1$, then $x=(2k-1)a$, so $(a+x)/2=ak$. So equation (1) becomes $a^{2ka-a}=(ak)^{ak}$, yielding $a^{k-1}=k^k$. The same exponent argument shows that k divides a, so $a=hk$. Thus $(hk)^{k-1}=k^k$ so $h^{k-1}=k\leqslant 2^{k-1}$, with strict inequality for $k>2$. Hence, since $k>1$, we must have $h=2=k$, $a=4$ and $x=ra=3a=12$.

References

[1] T. J. I. Bromwich, An Introduction to the Theory of Infinite Series, Macmillan, London, 1947.

[2] G. A. Heuer, Midpoint solutions of $x^x=a^b$, this MAGAZINE, 51 (1978) 181–183.

A Weighted Voting Model

ALEX KLEINER
Drake University
Des Moines, IA 50311

In a weighted voting body the power an individual member controls is rarely proportional to the number of votes that the member casts. Historically, two different methods for measuring power in a weighted voting body, the Shapley-Shubik index and the Banzhaf index, have found acceptance among mathematicians and in the courts. The survey paper [7] by W. F. Lucas provides an introduction (with extensive references) to both indices. More recently Straffin ([12], [13]) and others have proposed a generalized index which includes both the Shapley-Shubik and Banzhaf indices as special cases. In this note we apply Straffin's ideas to a special type of weighted voting body. Specifically, we are interested in the situation where one player has some number of votes greater than one and all others have one vote each. Voting bodies of this type can arise in many settings. Lucas cites instances from national and international governing bodies including the United Nations General Assembly, a 138-player game in which the Soviet Union has 3 votes [7]. In [6] Kleiner analyzes a board of this type which supplies services for a metropolitan region consisting of one large city, several suburban communities, and a rural area dotted with small towns. Committees with an even number of members in which each member has one vote and the chair has the additional power to break ties form a third class of examples. (Recent papers on power indices are listed in the references; see especially the papers in [9] and the references in [5].)

We will show that if any player observes that the other players behave in a homogeneous fashion, then that player is better off adopting such behavior rather than remaining independent. Although this result is what one might expect, we will obtain a nonintuitive result concerning the share of power of the player with "extra" votes.

We denote the n players by $N=\{1,2,3,\dots,n\}$ and suppose that player one is the distinguished player with x votes $(x>1)$ and furthermore that r votes are required for approval of an issue. As usual, we suppose at most one side can win any decision so that r is greater than half the total number of votes. We further suppose the x, r and n are all integers and that r is large enough so that player one acting alone can neither win nor block an opposing coalition. In formal terms we are considering the weighted majority game $[r;x,1,1,\dots,1]$ with $r>(n-1+x)/2$ and $1<x<r<n$.

To compute the Shapley-Shubik index we suppose that the players add their votes to a coalition in order of their degree of support on some issue. The unique player whose contribution to the total vote turns the coalition from losing to winning is called **pivotial** (or **marginal**) for that (ordered) coalition. Player i's index of power ϕ_i is the number of times that player i is marginal divided by $n!$, the total number of ordered coalitions. The Banzhaf index is computed by considering all 2^n possible combinations of yes and no votes. In any such combination, a player is marginal if a change of vote by that player will change the outcome. Player i's Banzhaf index β_i is the number of times player i is marginal divided by 2^n. (The Banzhaf index is frequently normalized by $\beta_i'=\beta_i/(\beta_1+\cdots+\beta_n)$.

Straffin's approach to weighted majority games (in **[13]**) is to assume that the n players form subsets called homogeneity classes. Within classes players tend to behave uniformly (in a probabilistic sense) while players in different classes act independently of one another. Straffin's model then uses as the power index the probability that a player is marginal given a particular homogeneity structure.

Formally we define a **partial homogeneity structure** to be a partition, $P=\{S_1,\dots,S_m\}$ of N and refer to the individual sets $S_1,\dots,S_m$ as **homogeneity classes**. If player i is a member of class j, then the probability that he votes yes on some issue, k, is assumed to be a random quantity, denoted by q_j (or $q_j(k)$). Over the set of all issues the $q_j(k)$'s are treated as independent random variables selected from the uniform distribution on $[0,1]$.

We denote the probability that an issue passes when player i votes for it and fails when he votes no by $E_i(q_1,\dots,q_m)$. The power index for a game with associated homogeneity structure, P, is defined by $\kappa'(P)=\kappa'=(\kappa_1',\dots,\kappa_n')$ where $\kappa_i'=\int_0^1\dots\int_0^1 E_i(q_1,\dots,q_m)dq_1\dots dq_m$. Note that $\kappa'(P)$ is an n dimensional vector; the ith component of $\kappa'(P)$ is the power index for player i under homogeneity structure P. Straffin has shown that if κ is the normalized version of κ' then $\kappa_i=\phi_i$, the Shapley-Shubik index, in the case of complete homogeneity (i.e. $m=1$) and that $\kappa_i=\beta_i'$, the normalized Banzhaf index, in the case of complete independence (i.e., $m=n$).

In this note we shall compare three homogeneity structures on the game $[r;x,1,\dots,1]$. These three are (1) complete homogeneity, (2) two homogeneous classes, one with only player one and the other with all $n-1$ other players, and (3) two homogeneous classes, one with player one and all but one of the one vote players and the other with a single one vote player. We shall denote by a a representative one-vote player from the same class as player one and by b a representative one-vote player from the class not containing player one. Notice that there is no player b in case 1 and no player a in case 2.

Case 1 is the easiest to handle since, as noted above, the index reduces to the Shapley-Shubik index. This particular game was first considered by Shapley and Shapiro **[11]** and the results are well known:

$$\kappa_1=\frac{x}{n},\qquad \kappa_a=\frac{n-x}{n(n-1)}.$$

In case two, where player one leaves the grand class, we have $m=2$ with $S_1=\{1\}$ and $S_2=\{2,3,\dots,n\}$. Player one makes a difference if and only if the number of one-vote players who vote for an issue is at least $r-x$ but no more than $r-1$. If we denote the binomial term $\binom{n}{k}x^k(1-x)^{n-k}$ by $f(n,k;x)$, then

$$E_1(q_1,q_2)=\sum_{k=r-x}^{r-1}\binom{n-1}{k}q_2^k(1-q_2)^{n-1-k}=\sum_{k=r-x}^{r-1}f(n-1,k;q_2).$$

Player b can make a difference if either player one and exactly $r-x-1$ of the other $n-2$ one-vote players vote for the issue or if player one votes against the issue and exactly $r-1$ of the others vote for the issue. Hence

$$E_b(q_1,q_2)=q_1f(n-2,r-x-1;q_2)+(1-q_1)f(n-2,r-1;q_2).$$

Thus, since $\int_0^1 f(n,k;x)dx=1/(n+1)$, it follows from the independence of q_1 and q_2 that

$$\kappa_1'=\int_0^1\int_0^1 E_1(q_1,q_2)dq_1dq_2=\frac{x}{n},$$

and

$$\kappa_b'=\int_0^1\int_0^1 E_b(q_1,q_2)dq_1dq_2=\frac{1}{n-1}.$$

We normalize these by dividing by $\Sigma_i\kappa_i'$ which is $(n+x)/n$ and get

$$\kappa_1=\frac{x}{n+x}\qquad\text{and}\quad \kappa_b=\frac{n}{(n-1)(n+x)}.$$

In case three, where a one vote player leaves the grand class, we have $m=2$ with $S_1=\{1,2,\ldots,n-1\}$, $S_2=\{n\}$. In this case we must compute indices for player one, for the one vote players in S_1 (denoted by player a) and for the lone player in S_2 (denoted by player b). As in case two, player one makes a difference if the number of one vote players who vote for the issue is between $r-x$ and $r-1$. This can happen either if player b votes yes and the number of one vote players from S_1 who vote yes is between $r-x-1$ and $r-2$ or if player b votes no and the number of one vote players from S_1 who vote yes is between $r-x$ and $r-1$. Thus

$$E_1=\sum_{k=r-x-1}^{r-2}f(n-2,k;q_1)q_2+\sum_{k=r-x}^{r-1}f(n-2,k;q_1)(1-q_2).$$

Player a makes a difference if either player one votes yes and exactly $r-x-1$ other one-vote players vote yes or player one votes no and exactly $r-1$ other one-vote players vote yes. However, if $x=r-1$, then it is impossible for player a to make a difference if players one and b both vote for the issue since there are r yes votes no matter what player a does. Thus we must consider the cases $x<r-1$ and $x=r-1$ separately. Likewise if $r=n-1$ player a cannot affect the outcome if both player one and player b vote no since there will be at most $n-2=r-1$ yes votes. Finally, if we recall that $r>(n-1+x)/2$ then $2r>n-1+x$ and if $x=r-1$ then $2r>n-1+r-1$ or $r>n-2$. Since $r<n$, we must have $r=n-1$ and the case $x=r-1$, $r<n-1$ is impossible. Thus to compute E_a we must consider three alternatives.

1. $x<r-1, r<n-1$. Here all four combinations of yes and no votes by players one and b are possible. If the sum of the yes votes from these two players is y, then there must be exactly $r-y-1$ yes votes cast by the $n-3$ other one vote players in S_1. Thus

$$E_a(q_1,q_2)=q_1f(n-3,r-x-2;q_1)q_2+q_1f(n-3,r-x-1;q_1)(1-q_2)$$
$$+(1-q_1)f(n-3,r-2;q_1)q_2+(1-q_1)f(n-3,r-1;q_1)(1-q_2).$$

2. $x<r-1$, $r=n-1$. This is similar to the first alternative except that players one and b cannot both vote no, so the last term in the above expression does not appear in this case.
3. $x=r-1$, $r=n-1$. Again this is similar to the first alternative except that players one and b must vote differently, so only the two middle terms of the result in 1 appear.

Player b makes a difference if exactly $r-1$ yes votes are cast by members of S_1. This can happen if player one and exactly $r-x-1$ other one-vote players vote for the issue or if player one votes against the issue and exactly $r-1$ other one-vote players vote for the issue. Hence

$$E_b(q_1,q_2)=q_1 f(n-2,r-x-1;q_1)+(1-q_1)f(n-2,r-1;q_1).$$

The Beta function formulas

$$\int_0^1 x f(n,k;x)\,dx=\frac{k+1}{(n+1)(n+2)} \quad \text{and} \quad \int_0^1 (1-x)f(n,k;x)\,dx=\frac{n-k+1}{(n+1)(n+2)}$$

make it possible to integrate each of E_1, E_a and E_b, producing

$$\kappa_1'=\frac{x}{n-1},\qquad \kappa_a'=\begin{cases}\dfrac{n-x-1}{(n-1)(n-2)} & \text{if } x<r-1, r<n-1,\\[2mm] \dfrac{n+r-2x-1}{2(n-1)(n-2)} & \text{if } x<r-1, r=n-1,\\[2mm] \dfrac{n-x}{2(n-1)(n-2)} & \text{if } x=r-1, r=n-1.\end{cases}$$

$$\kappa_b'=\frac{n-x}{n^2-n},$$

However, if we use $r=n-1$ in the second version of κ_a' and $x=n-2$ in the third, we obtain $\kappa_a'=(n-x-1)/[(n-1)(n-2)]$ in all three cases. Finally we compute $\Sigma_i\kappa_i'=(n^2-x)/[n(n-1)]$ and normalize to get

$$\kappa_1=\frac{nx}{n^2-x},\qquad \kappa_a=\frac{n(n-x-1)}{(n-2)(n^2-x)},\qquad \kappa_b=\frac{n-x}{n^2-x}.$$

TABLE 1 summarizes the power indices in each of the three cases. If we denote the partition of case i by P_i, then we have from the table that for all $x, \kappa_1(P_1)>\kappa_1(P_2)$ and $\kappa_a(P_1)>\kappa_b(P_3)$. Hence this result:

i	$\kappa_1(P_i)$	$\kappa_a(P_i)$	$\kappa_b(P_i)$
1	$\frac{x}{n}$	$\frac{n-x}{n(n-1)}$	———
2	$\frac{x}{n+x}$	———	$\frac{n}{(n+x)(n-1)}$
3	$\frac{nx}{n^2-x}$	$\frac{n(n-x-1)}{(n^2-x)(n-2)}$	$\frac{n-x}{n^2-x}$

Normalized Power Indices for Three Homogeneity Structures

TABLE 1

THEOREM 1. *In an n-player weighted majority game of the form* $[r;x,1,\ldots,1]$ *with* $x<r<n$, *no player can increase his power by unilaterally defecting from the grand homogeneous class.*

Theorem 1 is an intuitively appealing result analogous to Riker's size principle which says that a winning coalition increases the reward to the remaining members by reducing itself towards a minimal winning coalition, [**10**]. The impact of Theorem 1 is that it is never a good idea (as measured by share of power) for a lone individual to leave the grand homogeneous class. Since the maximum value of $\kappa_1(P_1)$ is $(n-2)/n$, which tends to 1 with large n, while $\kappa_1(P_2)$ is bounded above by $1/2$, it follows that the penalty for such a move can be quite severe for player one:

THEOREM 2. *In a completely homogeneous n-player, weighted majority game of the form* $[r;x,1,\ldots,1]$ *with* $x<r<n$, *player one's share of the power is limited only by the number of votes permitted him, whereas if player one defects from the grand homogeneous class unilaterally his share of the power is never as great as half the total power.*

That part of Theorem 2 which pertains to the grand homogeneous class is due to Shapely and Shapiro [**11**]. This theorem, which has the appearance of a paradox in the sense of Brams [**1**, page xv], has important implications for the members of certain weighted voting bodies. Suppose, for instance, that it is necessary to devise a governing board to provide a service function in a metropolitan area. If one large city has an absolute majority of the population, it might be desirable to choose a board of the form $[r; x, 1, \ldots, 1]$ where x is chosen so that the large city is not a dictator but has a share of the power roughly proportional to its population. If power is measured according to the Shapley-Shubik index, this can be accomplished, at least approximately. However, if it is perceived that the smaller players uniformly view player one as an antagonist, it is not possible to provide player one with enough votes to make the game fair.

For example, in Polk County, Iowa, there is one city (Des Moines) and several smaller suburbs. The city and four of the suburbs have joined together to form a metropolitan transit authority (MTA). Des Moines has 3 votes on the governing board of the MTA while the other four members have one vote each; the resulting game is $[4;,3,1,1,1,1]$. TABLE 2 illustrates the

i	$\kappa_1(P_i)$	$\kappa_a(P_i)$	$\kappa_b(P_i)$
1	.600	.100	—
2	.375	—	.155
3	.682	.076	.091

Power Indices for the Game [4;3, 1, 1, 1, 1]

TABLE 2

results in TABLE 1 for this game. As long as player one does not leave the grand class, it has at least 60% of the power. Since 83% of the people represented by the MTA board live in Des Moines, this is a more reasonable share of the power than if Des Moines tries to go it alone against the suburbs.

This note was written while the author was in residence at the University of Iowa.

References

[1] S. J. Brams, Paradoxes in Politics, The Free Press, New York, 1976.

[2] S. J. Brams and P. J. Affuso, Power and size: a new paradox, Theory and Decision, 7 (1976) 29–56.

[3] J. Deegan Jr. and E. W. Packel, To the (minimal winning) victors go the (equally divided) spoils: a new power index for simple n-person games, Math. Assoc. of Amer. Modules in Applied Mathematics, Cornell Univ., 1976.

[4] ______, A New Index of Power for Simple n-Person Games, Internat. J. Game Theory, 7 (1978) 113–123.

[5] P. Dubey and L. S. Shapley, Mathematical properties of the Banzhaf power index, Rand Paper P6016, the RAND Corporation, Santa Monica, CA, 1970.

[6] A. F. Kleiner, A Mathematical Analysis of the Solid Waste Agency Board, Polk County, Iowa, to appear.

[7] W. F. Lucas, Measuring Power in Weighted Voting Systems, Case Studies in Applied Mathematics, Math. Assoc. of Amer., Washington, D. C., 1976. Also appeared as Technical Report No. 227, Department of Operations Research, Cornell Univ., Ithaca, NY 14853.

[8] C. H. Nevison, B. Zicht and S. Schoepke, A naive approach to the Banzhaf index of power, Behaviorial Sci., 23 (1978) 130–131.

[9] P. C. Ordeshook (Editor), Game Theory and Political Science, New York Univ. Press, New York, 1978.

[10] W. H. Riker, The Theory of Political Coalitions, Yale Univ. Press, New Haven, 1962.

[11] L. S. Shapley and N. Z. Shapiro, Values of large games-I: a limit theorem, RM-2648, the RAND Corporation, Santa Monica, CA, 1960. Reprinted in Math. Operations Res., 3 (1978) 1–9.

[12] P. D. Straffin, Power Indices in Politics, Math. Assoc. of Amer. Modules in Applied Mathematics, Cornell Univ., Ithaca, 1976.

[13] ______, Homogeneity, independence and power indices, Public Choice, 30 (1977) 107–118.

Factoring Groups of Integers Modulo n

JOSEPH A. GALLIAN
University of Minnesota, Duluth
Duluth, MN 55812

DAVID J. RUSIN
Princeton University
Princeton, NJ 08540

In a ring with unity the set of elements with multiplicative inverses forms a group called the group of units. In the case of $R(n)$, the ring of integers modulo n, we denote this group by $U(n)$. The elements of $U(n)$ are the positive integers less than or equal to n and relatively prime to n so that the order of $U(n)$ is $\phi(n)$, the Euler phi function of n. Interest in these groups dates back as far as 1801 when they were investigated by Gauss in his classic book *Disquisitiones Arithmeticae*. Today they are especially valuable as examples in an undergraduate abstract algebra course. These groups illustrate in a concrete way the concepts of cyclic and noncyclic groups, isomorphism, internal direct products of subgroups, and external direct (or Cartesian) products of groups. For example, if we denote the internal direct product of two subgroups A and B by $A \times B$, the external direct product of two groups A and B by $A \otimes B$, and the group $\{0,1,2,\ldots,i-1\}$ under addition modulo i by $C(i)$, we see that $U(18) \approx U(9) \approx C(6)$ (exhibiting cyclic groups and isomorphisms); $U(15) \approx U(16) \approx C(2) \otimes C(4)$ (exhibiting noncyclic groups and the external direct product); $U(15) = \{1,4,7,13\} \times \{1,11\}$ (demonstrating the internal direct product); and $|U(16)| = |U(24)|$ but $U(16) \not\approx U(24)$ (exhibiting non-isomorphic abelian groups of the same order). Another fact which makes these groups interesting is that every finite Abelian group is isomorphic to a subgroup of $U(n)$ for infinitely many values of n [**2**, p. 96].

Shanks [**2**, p. 94] gives a method for writing the isomorphism class of $U(n)$ as an external direct product of groups of the form $C(i)$. For example, his method yields $U(105) \approx C(2) \otimes C(2) \otimes C(12)$. A major shortcoming of this technique is that it does not produce the actual subgroups of $U(105)$ which could be used to write $U(105)$ as an internal direct product. His method tells us that there are 12 elements of $U(105)$ which form a cyclic subgroup but not which 12 elements. Similarly, we don't know which set of 4 elements yields the direct factor isomorphic to $C(2) \otimes C(2)$.

In this note we give a simple method which enables one to write $U(n)$ as a product in three different forms: as an internal direct product of subgroups; as an external direct product of groups of the form $U(m)$ where $m|n$; and as an external direct product of groups of the form $C(i)$. In fact, for most integers the process can be carried out in a number of ways and it is ideally suited for implementation by a computer. These methods, in turn, will be generalized to certain subgroups of $U(m)$ in the latter part of the paper.

To begin, we observe that if $n = 1,2,4,p^k$, or $2p^k$ where p is an odd prime and $k \geqslant 1$, then $U(n)$ is cyclic [**2**, p. 92] so there exists no nontrivial factorization of $U(n)$. If n is not of that form, then either $n = 2^k$ with $k > 2$ or n can be written in the form st where $(s,t) = 1$ and $s > 2, t > 2$. We consider the latter case first. If $k|n$, let $U_k(n) = \{x \in U(n) | x \equiv 1 \bmod k\}$. Obviously, $U_k(n)$ is a subgroup of $U(n)$. Our method of factoring $U(n)$ is given in the following result.

THEOREM 1. *Suppose s and t are relatively prime. Then $U(st)$ is the internal direct product of $U_s(st)$ and $U_t(st)$, and $U(st)$ is isomorphic to the external direct product of $U(s)$ and $U(t)$. Moreover, $U_s(st) \approx U(t)$ and $U_t(st) \approx U(s)$, so $U(st) = U_s(st) \times U_t(st) \approx U(t) \otimes U(s)$.*

Proof. If s or t is 1, the result is trivially true so we assume neither is 1. First we show that $U_s(st) \cap U_t(st) = \{1\}$. If $x \in U_s(st) \cap U_t(st)$, then both s and t and therefore st divide $x-1$. Since $0 < x \leqslant st$, it follows that $x = 1$. Because $|U(st)| = \phi(st) = \phi(s)\phi(t) = |U(s)||U(t)|$, it suffices then to show that $U_s(st) \approx U(t)$ and $U_t(st) \approx U(s)$. To this end consider the correspondence $T: U_s(st) \rightarrow U(t)$ defined by $T(x)=r$ where $0 \leqslant r < t$ and $x \equiv r \bmod t$. Standard properties of congruences show that this correspondence is indeed a homomorphism. We will show that T is also 1-1 and onto $U(t)$. Let $x, y \in U_s(st)$ and suppose $T(x)=T(y)$. Then $x \equiv y \bmod s$ and $x \equiv y \bmod t$. Since $(s,t)=1$ it follows that $x \equiv y \bmod st$ so that T is 1-1. Next we show the image of T is contained in $U(t)$. Let $x \in U_s\ (st)$ and write $x = ut + r$ where $0 \leqslant r < t$. Then $(x, st) = 1$ and $T(x) = r$. It follows that $(x,t)=1$ and therefore $(r,t)=1$. Thus $T(x) \in U(t)$. Finally, we show that T is onto $U(t)$. Let $a \in U(t)$. We must find an $x \in U_s(st)$ such that $x \equiv a \bmod t$. Since $(s,t)=1$ there exist integers n and m such that $ns + mt = 1$. Let $x' = a + tm(1-a)$ and write $x' = qst + x$ where $0 \leqslant x < st$. We claim that x has the desired properties. Clearly, $x' \equiv a \bmod t$ and since $x' = a + tm(1-a) = a + (1-a) - (1-a)ns = 1 + (a-1)ns$, we also have $x' \equiv 1 \bmod s$. Thus $(x', s) = 1$. If p is a prime divisor of (x', t) then $p|(a,t)=1$. Thus $(x',t)=1$ also. It follows that $(x', st)=1$ and therefore $(x, st) = 1$. This shows $x \in U(st)$. Now, $x' \equiv 1 \bmod s$ implies $x \equiv 1 \bmod s$ so $x \in U_s(st)$. Finally, $x' \equiv a \bmod t$ implies $x \equiv a \bmod t$ and this completes the proof that $U_s(st)$ is isomorphic to $U(t)$. By symmetry, we have $U_t(st) \approx U(s)$ and the theorem is proved.

As an immediate consequence of the above theorem, we have the following result.

COROLLARY. *Let* $m = n_1 n_2 \cdots n_k$ *where* $(n_i, n_j) = 1$ *for* $i \neq j$. *Then* $U(m) = U_{m/n_1}(m) \times U_{m/n_2}(m) \times \cdots U_{m/n_k}(m) \approx U(n_1) \otimes U_{m/n_1}(m) \times U_{m/n_2}(m) \times \cdots \times U_{m/n_k}(m) \approx U(n_1) \otimes U(n_2) \otimes \cdots \otimes U(n_k)$.

To see how these results work, let's apply them to $U(105)$. We obtain

$$\begin{aligned}
U(105) &= U(15 \cdot 7) = U_{15}(105) \times U_7(105) \\
&= \{1,16,31,46,61,76\} \times \{1,8,22,29,43,64,71,92\} \\
&\approx U(7) \otimes U(15). \\
U(105) &= U(5 \cdot 21) = U_5(105) \times U_{21}(105) \\
&= \{1,11,16,26,31,41,46,61,71,76,86,101\} \times \{1,22,43,64\} \\
&\approx U(21) \otimes U(5). \\
U(105) &= U(3 \cdot 5 \cdot 7) = U_{35}(105) \times U_{21}(105) \times U_{15}(105) \\
&= \{1,71\} \times \{1,22,43,64\} \times \{1,16,31,46,61,76\} \\
&\approx U(3) \otimes U(5) \otimes U(7).
\end{aligned}$$

By keeping in mind the following facts (see [2, p. 93]),

$$\begin{aligned}
&U(2) \approx \{1\}, \qquad U(2^n) \approx C(2) \otimes C(2^{n-2}) \text{ for } n \geqslant 3, \\
&U(4) \approx C(2), \qquad U(p^n) \approx C(p^n - p^{n-1}) \text{ for odd } p,
\end{aligned}$$

we can easily use the corollary above to write the isomorphism class of $U(n)$ as an external direct product of the cyclic groups $C(i)$. For example, $U(105) = U(3 \cdot 5 \cdot 7) \approx U(3) \otimes U(5) \otimes U(7) \approx C(2) \otimes C(4) \otimes C(6)$ and $U(720) = U(16 \cdot 9 \cdot 5) \approx U(16) \otimes U(9) \otimes U(5) \approx C(2) \otimes C(4) \otimes C(6) \otimes C(4)$.

Finally, we consider the possibility that $n = 2^k$ where $k > 2$. In this one case no nontrivial analog of Theorem 1 exists, but we can say something about factorization.

THEOREM 2. *Let* $k > 2$. *Then* $U(2^k)$ *is the internal direct product of* $\{1, 2^k - 1\}$ *and* $\{3^i | 0 \leqslant i < 2^{k-2}\}$. *Furthermore,* $U(2^k)$ *is expressible as a nontrivial external direct product of U-groups if and only if* $2^{k-2}+1$ *is prime.*

Proof. Let A be the subgroup of $U(2^k)$ generated by 3 and let $B = \{1, 2^k - 1\}$. Since $|U(2^k)| = 2^{k-1}$ and $|A| = 2^{k-2}$ [**2**, p. 98], it follows that $U(2^k) = A \times B$ provided that $A \cap B = \{1\}$. For $k = 3$ we see this by inspection. For $k > 3$ note that $2^k - 1$ has order 2 while $3^{2^{k-3}} \equiv 1 + 2^{k-1} \bmod 2^k$ (see [**2**, p. 97]) is a unique element of order 2 in A so that $2^k - 1 \notin A$.

Next suppose that $U(2^k)$ can be written nontrivially as an external direct product of U-groups. Since $U(2^k) \approx C(2) \otimes C(2^{k-2})$ it follows that the product consists of one cyclic factor of order 2 and one cyclic factor, $U(m)$, of order 2^{k-2} [**1**, p. 113]. By Gauss' theorem [**2**, p. 92],

$m=4$, p^n or $2p^n$ where p is an odd prime. If $m=4$, then $2^{k-2}=2$ so that $2^{k-2}+1=3$ and the theorem holds. If $m=p^n$ or $2p^n$, then $2^{k-2}=|U(p^n)|=|U(2p^n)|=(p-1)p^{n-1}$ so that $n=1$ and $2^{k-2}=p-1$ and again the theorem holds. Finally, if $2^{k-2}+1$ is prime, then $U(2^k)\approx U(4)\otimes U(2^{k-2}+1)$.

In view of the foregoing discussion, it is natural to inquire about properties of groups of the form $U_s(st)$ in general, i.e., with no restriction on (s,t). In the remainder of this note we establish a formula for the order of $U_s(st)$ and show how this group can be factored both as an internal direct product of groups of the same form and as an external direct product of groups of the form $C(i)$.

THEOREM 3. $|U_s(st)|=\phi(t')t/t'$ *where* t' *is the largest divisor of* t *which is relatively prime to* s.

Proof. First we note that $(x,st)=1$ if and only if $(x,st')=1$ so that $U_s(st')$ is a subset of $U_s(st)$. Now the correspondence which sends each x in $U_s(st)$ to r where $0\leqslant r<st'$ and $x\equiv r \bmod st'$ is a homomorphism from $U_s(st)$ onto $U_s(st')$ whose kernel is $\{1+kst'|0\leqslant k<t/t'\}$. Thus $|U_s(st)|=|U_s(st')|t/t'=\phi(t')t/t'$ by the so-called first isomorphism theorem and Theorem 1.

COROLLARY 1. $|U_s(st)|=t\cdot\prod_{\substack{p|t\\ p\nmid s}}(1-1/p)$ *where* p *is prime.*

Proof. The corollary follows from the theorem and the well-known formula for $\phi(n)/n$ (see [2, p. 69]).

COROLLARY 2. $|U_s(st)|=\dfrac{|U(st)|}{|U(s)|}$.

Proof. This corollary follows from the previous one and the same formula for $\phi(n)/n$.

THEOREM 4. *If* $t=mn$ *and* $(m,n)=1$, *then* $U_s(st)=U_{sm}(st)\times U_{sn}(st)$.

Proof. Clearly $U_{sm}(st)\times U_{sn}(st)$ is a subgroup of $U_s(st)$ and as in the proof of Theorem 1 we have $U_{sm}(st)\cap U_{sn}(st)=\{1\}$. Finally, it follows from Corollary 1 of Theorem 3 and Theorem 2.5.1 in [1] that $|U_s(st)|=|U_{sm}(st)||U_{sn}(st)|=|U_{sm}(st)\times U_{sn}(st)|$. This completes the proof.

COROLLARY. *Let* $t=p_1^{n_1}p_2^{n_2}\cdots p_r^{n_r}$ *be the prime power decomposition of* t. *Then* $U_s(st)=U_{q_1}(st)\times U_{q_2}(st)\times\cdots\times U_{q_k}(st)$ *where* $q_i=st/p_i^{n_i}$.

Theorem 4 and its corollary show how $U_s(st)$ can be expressed as an internal direct product of subgroups of the same form. In order to write $U_s(st)$ as an external direct product of cyclic additive groups, we need a "cancellation" theorem analogous to the "$U_s(st)\approx U(t)$" portion of Theorem 1. The next theorem is the natural generalization of this.

THEOREM 5. *Let* s' *be the largest divisor of* s *which is relatively prime to* t. *Then* $U_s(st)$ *and* $U_{s/s'}(ts/s')$ *are isomorphic.*

Proof. For each x in $U_s(st)$ define $T(x)=r$ where $0\leqslant r<(ts/s')$ and $x\equiv r \bmod (ts/s')$. We will show that T is the desired isomorphism. Since $x\equiv 1 \bmod s$, we see $x\equiv 1 \bmod s/s'$. Also, since $T(x)\equiv x \bmod t$ and $(t,x)=1$, we have $(T(x),t)=1$ so that T is a homomorphism into $U_{s/s'}(ts/s')$. If $x\in U_s(st)$ and $x\in \operatorname{Ker} T$ then $x\equiv 1 \bmod s$ and $x\equiv 1 \bmod (ts/s')$ also. Thus $x\equiv 1 \bmod st$ (since $[s,(ts/s')]=st$) so that $x=1$ and T is 1-1. Finally, it follows from the definition of s' and Corollary 1 of Theorem 3 that $|U_s(st)|=|U_{s/s'}(ts/s')|$ so that T is onto.

Taken together, Theorems 1 and 5 and the corollary to Theorem 4 reduce the problem of decomposing the group $U_s(st)$ into an external direct product of cyclic additive groups to that of decomposing $U_s(sp^n)$ where p is a prime divisor of s. This final step is accomplished next.

THEOREM 6. *Let* p *be a prime divisor of* s. *Then* $U_s(sp^n)$ *is isomorphic to* $C(p^n)$ *if* $p\neq 2$ *or if* $p=2$ *and* $4|s$; *otherwise,* $U_s(sp^n)$ *is isomorphic to* $C(2)\otimes C(2^{n-1})$.

Proof. Write $s=p^k s'$ where $(p,s')=1$. By Theorem 5 we know $U_s(sp^n)=U_s(p^k s' p^n)$ is isomorphic to $U_{p^k}(p^{k+n})$. By Corollary 1 of Theorem 3, $U_{p^k}(p^{k+n})$ is a subgroup of order p^n of $U(p^{k+n})$. If $p\neq 2$, then $U(p^{k+n})$ and all of its subgroups are cyclic so $U_s(sp^n)\approx C(p^n)$. Now suppose that $p=2$ and $4|s$; then $U_s(s2^n)\approx U_{2^k}(2^{k+n})$ where $k\geqslant 2$. To prove $U_{2^k}(2^{k+n})$ is cyclic, it suffices to show that it contains a unique element of order 2. To this end let $x=1+2^k m$ belong to $U_{2^k}(2^{k+n})$ and assume $|x|=2$. Then $x^2=1+m2^{k+1}+m^2 2^{2k}\equiv 1 \bmod 2^{k+n}$ so that $2^{k+n}|m2^{k+1}(1+m2^{k-1})$. Since $k>1$, $1+m2^{k-1}$ is odd and therefore m must have the form $2^{n-1}u$. If $u\geqslant 2$, then $x=1+2^k m=1+2^k\cdot 2^{n-1}u\geqslant 1+2^{k+n}$. But $x\in U(2^{k+n})$ implies that $x<2^{k+n}$. Thus $u=1$ and therefore $U_{2^k}(2^{k+n})$ is cyclic.

Finally, assume that $p=2$ and $s=2s'$ where $(s',2)=1$. Then, as before, $U_s(sp^n)\approx U_2(2^{n+1})=U(2^{n+1})\approx C(2)\otimes C(2^{n-1})$. This completes the proof.

Let us look at a few examples to illustrate the above theorems. Direct calculations show that $U_8(48)=\{1,17,25,41\}$. Theorem 3 says that $|U_8(48)|=\phi(3)\cdot 6/3=4$ and from Theorems 4, 5, and 6 we obtain:

$$\begin{aligned} U_8(48) &= U_8(8\cdot 6)=U_{24}(48)\times U_{16}(48)=\{1,25\}\times\{1,17\} \\ &\approx U_8(16)\otimes U_1(3)\approx C(2)\otimes C(2). \end{aligned}$$

Similarly, we have $|U_{10}(900)|=\phi(9)\cdot 90/9=60$ and

$$\begin{aligned} U_{10}(900) &= U_{10}(10\cdot 2\cdot 9\cdot 5)=U_{450}(900)\times U_{100}(900)\times U_{180}(900) \\ &\approx U_2(4)\otimes U(9)\otimes U_5(25)\approx C(2)\otimes C(6)\otimes C(5). \end{aligned}$$

Finally, we claim the existence of two isomorphisms; verifications are left as exercises. If each prime divisor of t is a divisor of s_1 as well, then $U_{s_1}(s_1 t)\approx U_{s_2}(s_2 t)$ whenever $s_1\equiv s_2 \bmod t$, and $U_{s_1}(s_1 t)\approx U_d(dt)$ where $d=(s_1,t)$. (For the first isomorphism consider the map $T: U_{s_2}(s_2 t)\to U_{s_1}(s_1 t)$ given by $T(x)=1+(s_1/s_2)(x-1)$.)

This paper was written while the second author was supported by an NSF Undergraduate Research Participation grant at the University of Minnesota, Duluth. The authors would like to thank the referees for their constructive comments.

References

[1] I. N. Herstein, Topics in Algebra, 2nd ed., Xerox, Lexington, Mass., 1975

[2] D. Shanks, Solved and Unsolved Problems in Number Theory, 2nd ed., Chelsea, New York, 1978.

Archimedes' Axioms for Arc-Length and Area

SCOTT E. BRODIE
Rockefeller University
New York, NY 10021

The first book of Archimedes, *On the Sphere and Cylinder*, opens with a formal introduction followed by some definitions and five new axioms which (taken together with those of Euclid) the author apparently felt were necessary for his work. The first two of these axioms ([**1**], pp. 3–4) are typical:

1. Of all lines which have the same extremities, the straight line is the least.
2. Of other lines in a plane and having the same extremities (any two) such are unequal whenever both are concave in the same direction and one of them is either wholly included between the other and the straight line which has the same extremities with it, or is partly included by, and is partly common with, the other; and that [line] which is included is the lesser of the two.

Proof. Write $s=p^k s'$ where $(p,s')=1$. By Theorem 5 we know $U_s(sp^n)=U_s(p^k s'p^n)$ is isomorphic to $U_{p^k}(p^{k+n})$. By Corollary 1 of Theorem 3, $U_{p^k}(p^{k+n})$ is a subgroup of order p^n of $U(p^{k+n})$. If $p\neq 2$, then $U(p^{k+n})$ and all of its subgroups are cyclic so $U_s(sp^n)\approx C(p^n)$. Now suppose that $p=2$ and $4|s$; then $U_s(s2^n)\approx U_{2^k}(2^{k+n})$ where $k\geqslant 2$. To prove $U_{2^k}(2^{k+n})$ is cyclic, it suffices to show that it contains a unique element of order 2. To this end let $x=1+2^k m$ belong to $U_{2^k}(2^{k+n})$ and assume $|x|=2$. Then $x^2=1+m2^{k+1}+m^2 2^{2k}\equiv 1 \bmod 2^{k+n}$ so that $2^{k+n}|m2^{k+1}(1+m2^{k-1})$. Since $k>1$, $1+m2^{k-1}$ is odd and therefore m must have the form $2^{n-1}u$. If $u\geqslant 2$, then $x=1+2^k m=1+2^k\cdot 2^{n-1}u\geqslant 1+2^{k+n}$. But $x\in U(2^{k+n})$ implies that $x<2^{k+n}$. Thus $u=1$ and therefore $U_{2^k}(2^{k+n})$ is cyclic.

Finally, assume that $p=2$ and $s=2s'$ where $(s',2)=1$. Then, as before, $U_s(sp^n)\approx U_2(2^{n+1})=U(2^{n+1})\approx C(2)\otimes C(2^{n-1})$. This completes the proof.

Let us look at a few examples to illustrate the above theorems. Direct calculations show that $U_8(48)=\{1,17,25,41\}$. Theorem 3 says that $|U_8(48)|=\phi(3)\cdot 6/3=4$ and from Theorems 4, 5, and 6 we obtain:

$$\begin{aligned}U_8(48)&=U_8(8\cdot 6)=U_{24}(48)\times U_{16}(48)=\{1,25\}\times\{1,17\}\\&\approx U_8(16)\otimes U_1(3)\approx C(2)\otimes C(2).\end{aligned}$$

Similarly, we have $|U_{10}(900)|=\phi(9)\cdot 90/9=60$ and

$$\begin{aligned}U_{10}(900)&=U_{10}(10\cdot 2\cdot 9\cdot 5)=U_{450}(900)\times U_{100}(900)\times U_{180}(900)\\&\approx U_2(4)\otimes U(9)\otimes U_5(25)\approx C(2)\otimes C(6)\otimes C(5).\end{aligned}$$

Finally, we claim the existence of two isomorphisms; verifications are left as exercises. If each prime divisor of t is a divisor of s_1 as well, then $U_{s_1}(s_1 t)\approx U_{s_2}(s_2 t)$ whenever $s_1\equiv s_2 \bmod t$, and $U_{s_1}(s_1 t)\approx U_d(dt)$ where $d=(s_1,t)$. (For the first isomorphism consider the map $T: U_{s_2}(s_2 t)\to U_{s_1}(s_1 t)$ given by $T(x)=1+(s_1/s_2)(x-1)$.)

This paper was written while the second author was supported by an NSF Undergraduate Research Participation grant at the University of Minnesota, Duluth. The authors would like to thank the referees for their constructive comments.

References

[1] I. N. Herstein, Topics in Algebra, 2nd ed., Xerox, Lexington, Mass., 1975

[2] D. Shanks, Solved and Unsolved Problems in Number Theory, 2nd ed., Chelsea, New York, 1978.

Archimedes' Axioms for Arc-Length and Area

Scott E. Brodie
Rockefeller University
New York, NY 10021

The first book of Archimedes, *On the Sphere and Cylinder*, opens with a formal introduction followed by some definitions and five new axioms which (taken together with those of Euclid) the author apparently felt were necessary for his work. The first two of these axioms ([**1**], pp. 3–4) are typical:

1. Of all lines which have the same extremities, the straight line is the least.
2. Of other lines in a plane and having the same extremities (any two) such are unequal whenever both are concave in the same direction and one of them is either wholly included between the other and the straight line which has the same extremities with it, or is partly included by, and is partly common with, the other; and that [line] which is included is the lesser of the two.

The first axiom is well known. Put in modern language, the second states that if two curves in the plane both intersect a line at the same two points such that each curve, together with the line, bounds a convex set, and if one of these convex sets is a proper subset of the other, then the curve bounding the larger set is the longer of the two (see FIGURE 1). The next two axioms are the analogues, for the relative area of curved surfaces, of the first two axioms. The final axiom asserts the famous "Archimedean property" that, in essence, denies the existence of infinitesimals.

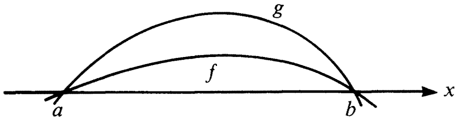


FIGURE 1

We may be confident that these principles were introduced as axioms only because Archimedes was unable to prove them. This note discusses a proof, in the modern context, of Archimedes' second axiom; the fourth axiom may be demonstrated by strictly analogous means in three dimensions.

These axioms play an important role in most of Archimedes' major theorems, which typically determine the area or volume of a geometric figure by means of the double *reductio ad absurdum* argument now referred to as the "method of exhaustion." For such proofs it is necessary to compare the content (arc length, area, or volume) of the figure with that of suitably chosen inscribed or circumscribed figures. In the case of the area of plane figures or the volume of solids, the figures to be compared are generally nested sets, and the comparison is justified by appeal to Euclid's "common notion" that "the whole is greater than the part" ([**3**], Vol. I, p. 155). (For example, see Euclid [**3**], Book XII, prop. 2, "Circles are to one another as the squares on the diameters," or Archimedes [**1**], prop. 28.) On the other hand, when arc length or the surface of solid figures is at issue, the approximating figures are no longer subsets or supersets of the figure under study, and Archimedes was thus forced to introduce different principles to treat such situations (e.g., [**1**], prop. 1 and prop. 28, or [**2**], prop. 1, "The area of any circle is equal to a right-angled triangle in which one of the sides about the right angle is equal to the radius, and the other to the circumference, of the circle," and prop. 3, "The ratio of the circumference of any circle to its diameter is less than $3\frac{1}{7}$ but greater than $3\frac{10}{71}$.")

The hypothesis of Archimedes' second axiom may be readily cast into the language of elementary calculus: suppose that f and g are continous functions, twice differentiable on the open interval (a,b), which satisfy

$$\begin{gathered} f(a)=f(b)=g(a)=g(b)=0 \\ 0<f(x)<g(x) \\ f''(x)<0,\, g''(x)<0, \end{gathered} \tag{1}$$

where the inequalities hold throughout the interval (a,b). To "prove" the axiom, we must show that

$$\int_a^b \sqrt{1+(f'(x))^2}\,dx < \int_a^b \sqrt{1+(g'(x))^2}\,dx. \tag{2}$$

Despite the apparent simplicity of this formulation, it does not readily yield a proof. An insight into the difficulty is given by FIGURE 2. For such curves as are shown there, $\int_a^c \sqrt{1+(f'(x))^2}\,dx > \int_a^c \sqrt{1+(g'(x))^2}\,dx$ for many values of c between a and b. Thus as parameterized by x, the inequality (2) does *not* follow by virtue of any simple inequality between the integrands $\sqrt{1+(f'(x))^2}$ and $\sqrt{1+(g'(x))^2}$.

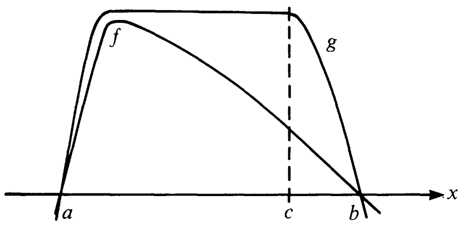


FIGURE 2

This motivates a different choice of parameter: Let $\mathbf{x}$ denote the position vector for the inner curve (i.e., for f) parameterized by its arc length s over the interval from 0 to S. At each point on this curve we introduce the unit tangent vector $\hat{\tau}$ and the unit (outward-pointing) normal vector $\hat{\nu}$. According to the well-known Frenet formulas in the plane,

$$\begin{aligned} d\mathbf{x}/ds &= \hat{\tau} \\ d\hat{\tau}/ds &= \kappa\hat{\nu} \\ d\hat{\nu}/ds &= -\kappa\hat{\tau}, \end{aligned} \tag{3}$$

where κ is the curvature of the inner curve. In this setting, the assumption that the curve $\mathbf{x}$ is concave downward is equivalent to the statement that $\kappa < 0$ on the open interval $(0, S)$.

In order to describe the outer curve g in this context, we introduce a second coordinate scheme for the region of the plane which lies "above" the inner curve $\mathbf{x}(s)$. Consider the set of outward-pointing rays which are normal to the inner curve at the various points $\mathbf{x}(s)$, as s varies from 0 to S. Since the inner curve is concave downward, these rays must diverge in such a way that no two of them intersect. On the other hand, any point within the region above the curve $\mathbf{x}(s)$ and the two rays at the endpoints $\mathbf{x}(0)$ and $\mathbf{x}(S)$ must lie on one of these rays. Thus, letting $\mathbf{y}$ denote the position vector for the outer curve, we may write

$$\mathbf{y}(s) = \mathbf{x}(s) + \eta(s)\hat{\nu}(s) \tag{4}$$

where $\eta(s)$ is a positive number which gives the (least) distance from the point $\mathbf{y}(s)$ to the inner curve $\mathbf{x}$. The shape of the outer curve is thus specified by the function $\eta(s)$ (FIGURE 3). So, rather than supposing (as above) that g is a twice-differentiable function on the interval (a, b), we require instead that $\eta(s)$ be single-valued, differentiable function on the open interval $(0, S)$.

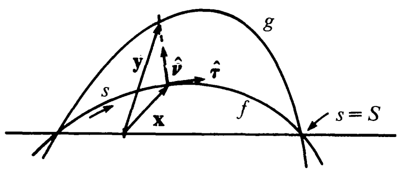


FIGURE 3

(It may be shown by use of the Implicit Function Theorem that the parameterization (4) is valid, and that $\eta(s)$ is differentiable, whenever the function g is differentiable and concave downward. However, the hypothesis that the outer curve is concave downward is actually unnecessary for the "proof" of Archimedes' axiom below. Thus, it is more appropriate to assume directly that the parameterization (4) is valid, with η a differentiable function of s, and to dispense with Archimedes' unnecessary hypothesis that the outer curve is concave downward.)

In place of the conditions (1) on the functions f and g, we now adopt a similar set of conditions on the scalar functions η and κ:

$$\eta(0) = \eta(S) = 0, \quad \eta(s) > 0, \quad \kappa(s) < 0, \tag{5}$$

where the inequalities hold through the interval $(0,S)$. Since the arc length of the outer curve is given by $\int_0^S |dy/ds|ds$, and the arc length S of the inner curve by $\int_0^S 1ds$, in order to prove Archimedes' axiom, we must show that $\int_0^S |dy/ds|ds > \int_0^S 1ds$. Clearly, it suffices to show that in the open interval $(0,S)$,

$$|dy/ds|^2 > 1. \tag{6}$$

To this end, we differentiate (4) and apply (3):

$$\frac{dy}{ds} = \frac{dx}{ds} + \frac{d\eta}{ds}\hat{\nu} + \eta\frac{d\hat{\nu}}{ds}$$
$$= \frac{d\eta}{ds}\hat{\nu} + (1-\kappa\eta)\hat{\tau}.$$

Since $\hat{\nu}$ and $\hat{\tau}$ are orthogonal unit vectors,

$$\left|\frac{dy}{ds}\right|^2 = \frac{dy}{ds}\cdot\frac{dy}{ds} = \left(\frac{d\eta}{ds}\right)^2 + (1-\kappa\eta)^2.$$

Application of the inequalities in (5) yields (6).

There are many reasons why this proof could not be set in the context of Archimedes' time; perhaps the most fundamental is that the Greeks lacked formal definitions of arc-length and surface area (the latter continues to be awkward even today), making it difficult to prove general theorems without reference to specific geometric figures.

References

[1] Archimedes, On the Sphere and Cylinder, Book I, The Works of Archimedes, T.L. Heath, Editor, Dover, New York, 1953, pp. 1–55.

[2] Archimedes, Measurement of a Circle, loc. cit., pp. 91–98.

[3] Euclid's Elements, T.L. Heath, translator, 2nd ed., Dover, New York, 1956.

An Intuitive Approach to Interval Numbers of Graphs

FRANK HARARY
JERALD A. KABELL
University of Michigan
Ann Arbor, MI 48109

The study of multiple interval graphs is one of the most active new areas of research in graph theory. Since multiple interval graphs are special kinds of intersection graphs, we begin by describing the more general (and more well-known) intersection graphs. An **intersection graph** is a graph whose points represent certain distinct nonempty subsets of some set S, two points being adjacent (joined by an edge) if and only if their corresponding subsets have a nonempty intersection. Since it is easily shown that every graph is an intersection graph [**8**, p. 19], one of the natural impulses of a mathematician is to restrict the nature of the set S, or of the subsets, or both. If S is taken to be the real line with the allowable subsets being the finite intervals (either open or closed—it makes no difference), the resulting graph is termed an **interval graph**. (FIGURE 1 contains examples of graphs which are interval graphs, and some that are not.)

Historically, interval graphs were first introduced by Hajós [7] in 1957 from a purely abstract mathematical point of view. They quickly acquired practical significance, however, when Benzer [1] proposed in 1959 to use them as a model of genetic fine structure, and many other where the inequalities hold through the interval $(0,S)$. Since the arc length of the outer curve is given by $\int_0^S |dy/ds|ds$, and the arc length S of the inner curve by $\int_0^S 1ds$, in order to prove Archimedes' axiom, we must show that $\int_0^S |dy/ds|ds > \int_0^S 1ds$. Clearly, it suffices to show that in the open interval $(0,S)$,

$$|dy/ds|^2 > 1. \tag{6}$$

To this end, we differentiate (4) and apply (3):

$$\begin{aligned}\frac{dy}{ds} &= \frac{dx}{ds} + \frac{d\eta}{ds}\hat{\nu} + \eta\frac{d\hat{\nu}}{ds}\\ &= \frac{d\eta}{ds}\hat{\nu} + (1-\kappa\eta)\hat{\tau}.\end{aligned}$$

Since $\hat{\nu}$ and $\hat{\tau}$ are orthogonal unit vectors,

$$\left|\frac{dy}{ds}\right|^2 = \frac{dy}{ds}\cdot\frac{dy}{ds} = \left(\frac{d\eta}{ds}\right)^2 + (1-\kappa\eta)^2.$$

Application of the inequalities in (5) yields (6).

There are many reasons why this proof could not be set in the context of Archimedes' time; perhaps the most fundamental is that the Greeks lacked formal definitions of arc-length and surface area (the latter continues to be awkward even today), making it difficult to prove general theorems without reference to specific geometric figures.

References

[1] Archimedes, On the Sphere and Cylinder, Book I, The Works of Archimedes, T.L. Heath, Editor, Dover, New York, 1953, pp. 1–55.

[2] Archimedes, Measurement of a Circle, loc. cit., pp. 91–98.

[3] Euclid's Elements, T.L. Heath, translator, 2nd ed., Dover, New York, 1956.

An Intuitive Approach to Interval Numbers of Graphs

FRANK HARARY

JERALD A. KABELL

University of Michigan

Ann Arbor, MI 48109

The study of multiple interval graphs is one of the most active new areas of research in graph theory. Since multiple interval graphs are special kinds of intersection graphs, we begin by describing the more general (and more well-known) intersection graphs. An **intersection graph** is a graph whose points represent certain distinct nonempty subsets of some set S, two points being adjacent (joined by an edge) if and only if their corresponding subsets have a nonempty intersection. Since it is easily shown that every graph is an intersection graph [**8**, p. 19], one of the natural impulses of a mathematician is to restrict the nature of the set S, or of the subsets, or both. If S is taken to be the real line with the allowable subsets being the finite intervals (either open or closed—it makes no difference), the resulting graph is termed an **interval graph.** (FIGURE 1 contains examples of graphs which are interval graphs, and some that are not.)

Historically, interval graphs were first introduced by Hajós [**7**] in 1957 from a purely abstract mathematical point of view. They quickly acquired practical significance, however, when Benzer [**1**] proposed in 1959 to use them as a model of genetic fine structure, and many other

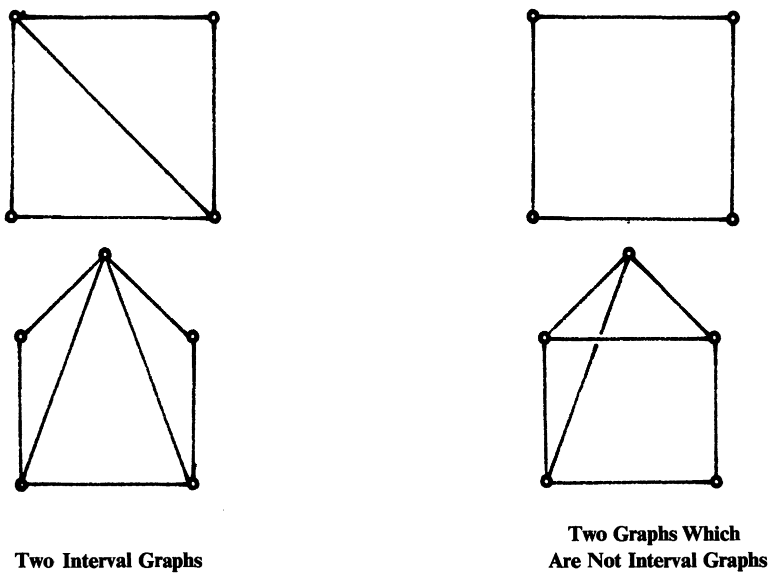


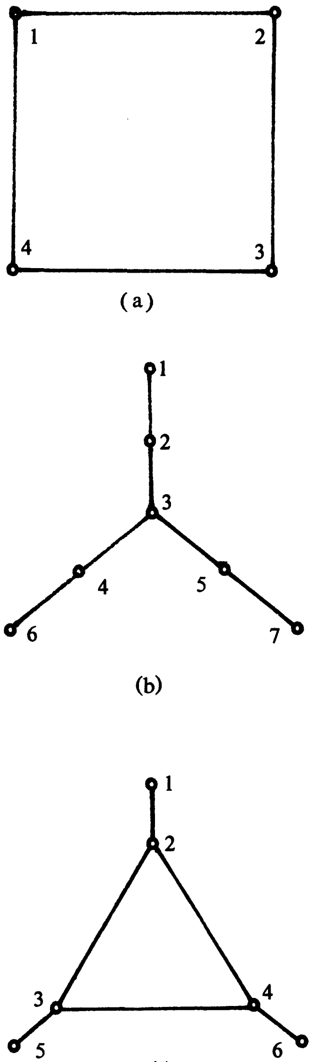
FIGURE 1.

applications have emerged since that time; see, for example, [**2**] or [**12**]. The question of which graphs are interval graphs is considerably less straightforward than the corresponding problem for intersection graphs since it is immediately clear that not all graphs are interval graphs: the smallest counterexample is provided by the cycle of length 4 (see FIGURE 1). The first characterization of interval graphs was given in 1962 by Lekkerkerker and Boland [**11**]. They gave not only a straightforward structural criterion, but also a complete set of forbidden induced subgraphs. Alternative characterizations have since been given by Gilmore and Hoffman [**4**] and Fulkerson and Gross [**3**].

The most useful of these characterizations for our purposes is the structural criterion of Lekkerkerker and Boland. To state it, we need a few special definitions. In a **chordal** graph, each cycle of length greater than three contains a **chord**, that is, an edge joining two points of the cycle which are not adjacent along the cycle. A graph is said to be **asteroidal** if it contains three points with the property that between each pair there is a path P such that no point of P (including the endpoints) is adjacent to the third point. Lekkerkerker and Boland showed that *a graph is an interval graph if and only if it is chordal and nonasteroidal.* FIGURE 2 gives examples of three graphs which violate these conditions. These are some of the forbidden subgraphs in the complete set given in [**11**].

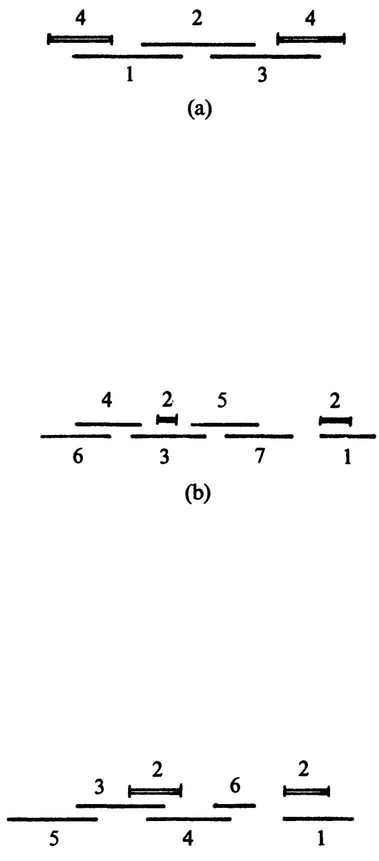
If we relax the requirement that each point correspond to one interval and allow each point to correspond to the union of a finite family of intervals, we obtain a **multiple interval graph.** This type of graph was independently introduced by Trotter and Harary [**13**] and by Griggs and West [**5**]. They have been extensively studied by the present authors [**9**], and by Griggs [**6**]. The first result of this change is that the forbidden subgraphs of FIGURE 2 now all have representations, as shown in FIGURE 3. Note that each of these contains a family of two intervals.

It is easy to see that every graph is a multiple interval graph. All that is necessary is to take a collection of disjoint intervals, one for each point, and from each select a collection of disjoint subintervals, one for each point in the neighborhood of the point represented by the original interval. Now for any multiple-interval representation of a graph G, there will be an integer t such that all the points of G are represented by families of at most t intervals. The minimum value of t over all multiple interval representations of G is then a well-defined numerical invariant of G which has been named the **interval number** of G and denoted by $i(G)$. Clearly, if G is an interval graph, then $i(G)=1$, and conversely. Since the graphs of FIGURE 2 are known



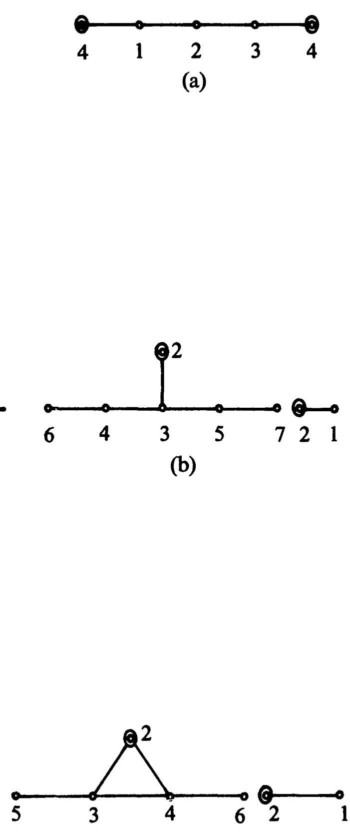
Three Examples of Graphs Which Are Not Interval Graphs.

FIGURE 2.



Multiple Interval Representations of the Graphs in FIGURE 2.

FIGURE 3.



The Interval Graphs Obtained from the Multiple Interval Representations in FIGURE 3.

FIGURE 4.

not to be interval graphs, the representations in FIGURE 3 allow us to conclude that in each of these cases, $i(G)=2$.

There are several different approaches to the problem of determining the interval numbers of specific graphs or families of graphs. Griggs and West [5], and Griggs [6] employ a straightforward, essentially brute-force approach, while Trotter and Harary [13] treat it as a problem in extending functions, where the domain of the function is the point set of the graph, and the range lies in the set of intervals of the real line. The approach outlined below was originally introduced in [9] and so far appears to yield the maximum return on effort invested.

If we temporarily ignore the labelling, we may consider the collections of intervals in FIGURE 3 to be ordinary interval representations of interval graphs. The corresponding interval graphs are displayed in FIGURE 4 with certain distinguished points that carry in each case the only repeated label. Comparing the graphs of FIGURE 4 with the original graphs in FIGURE 2, it is relatively easy to imagine cutting a point in two (in the graphs of FIGURE 2) to obtain the graphs in FIGURE 4. Thinking in these terms, the interval number of a graph becomes the maximum number of pieces into which any point must be cut to convert the graph into an interval graph. We can make this idea more precise by the following useful characterization of interval numbers, which may indeed be considered an alternate definition.

LEMMA. *The interval number of a graph G is the smallest integer t for which there exists an interval graph H and a mapping $\phi: H \to G$ such that*

(i) $\phi(H) = G$;

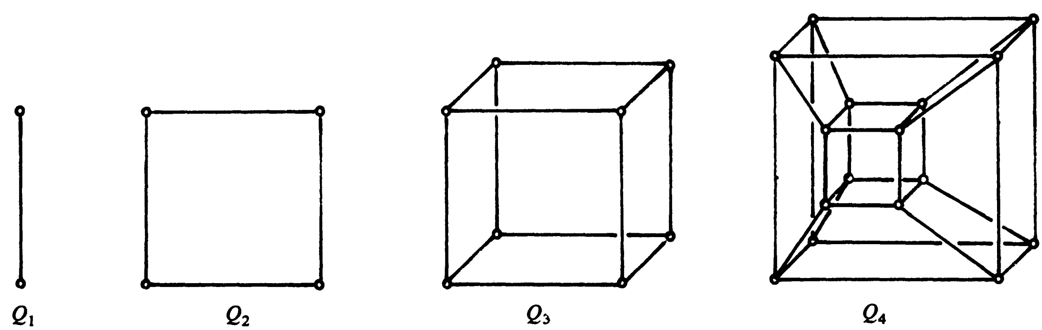
(ii) *ϕ preserves adjacencies, in the sense that the images of two adjacent points are either adjacent or coincident;*

(iii) *for each point v of G, $|\phi^{-1}(v)| \leq t$.*

Proof. Suppose G is a graph with $i(G) = t$, and $V(G) = \{v_1, v_2, \ldots, v_p\}$. Then there exists a multiple interval representation of G in which v_i is represented by the set $I_{i1} \cup I_{i2} \cup \cdots \cup I_{ir_i}$, where each I_{ij} is a finite interval on the real line and $r_i \leq t$. Now define the family $F = \{I_{ij}\}$ and the graph $H = \Omega(F)$, the intersection graph of the family F. We then label the points of $V(H) = \{u_{ij}\}$ so that point u_{ij} represents the interval I_{ij}. This naturally leads us to consider the map $\phi: H \to G$ defined by $\phi(u_{ij}) = v_i$, which is easily seen to satisfy conditions (i)—(iii).

Conversely, let G, H and ϕ satisfy these three conditions, with $V(G) = \{v_1, v_2, \ldots, v_p\}$. Label the points of $V(H)$ by $\phi^{-1}(v_i) = \{u_{i1}, u_{i2}, \ldots, u_{ir_i}\}$. Since H is an interval graph, it has an interval representation $\{I_{ij}\}$ where the interval I_{ij} represents the point u_{ij}. Now define the set $S_i = \cup_j I_{ij}$, and let $F = \{S_i\}$. From this construction it is clear that $\Omega(F) \cong G$ and F is a multiple-interval representation of G in which each S_i consists of at most t finite intervals.

To illustrate the power of the lemma, let us consider the problem of finding the interval numbers of the n-dimensional cubes Q_n. The graph Q_n is simply the skeleton of the n-dimensional hypercube, and may be defined recursively by iterated cartesian products: $Q_1 = K_2$, the complete graph on two points, and $Q_n = Q_{n-1} \times K_2$. FIGURE 5 shows the first four cubes.

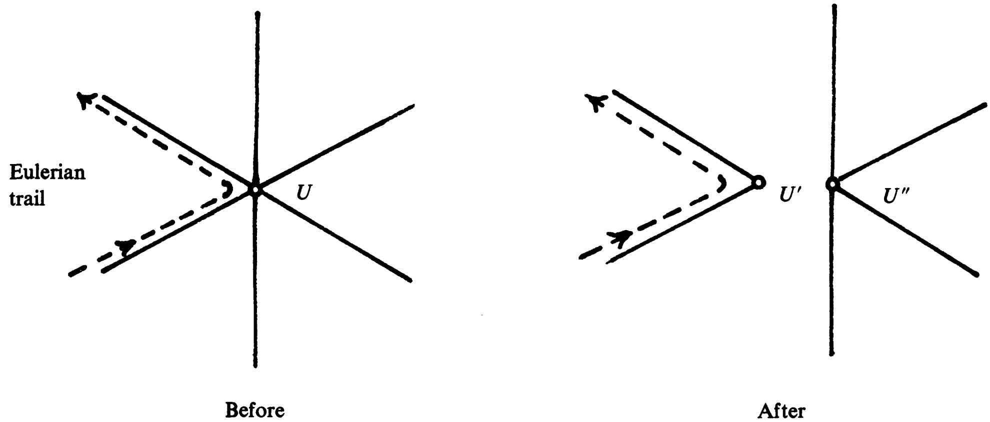


The Small Cubes.

FIGURE 5.

We shall now show that $i(Q_n) = [n/2] + 1$. We begin by considering Q_n where n is even. Since each point has degree n, the graph is eulerian. Now follow a closed eulerian trail and at each point "cut free" the incoming and outgoing lines of the trail without separating them from each other. By cutting free the trail at point u, we mean replacing u by two points u' and u'', where u' is incident with the incoming and outgoing lines only and u'' is incident with the remaining lines at u. This process is illustrated in FIGURE 6. When the trail has been completely traversed, each point will have been cut into $n/2$ pieces and the resulting graph is a cycle. This can be transformed into an interval graph (actually a path) by one additional cut, and so $i(Q_n) \leq n/2 + 1$.

When n is odd, we make use of the fact that Q_n consists of two copies of Q_{n-1} with corresponding points joined. Designate these two cubes as "left" and "right" and trace an

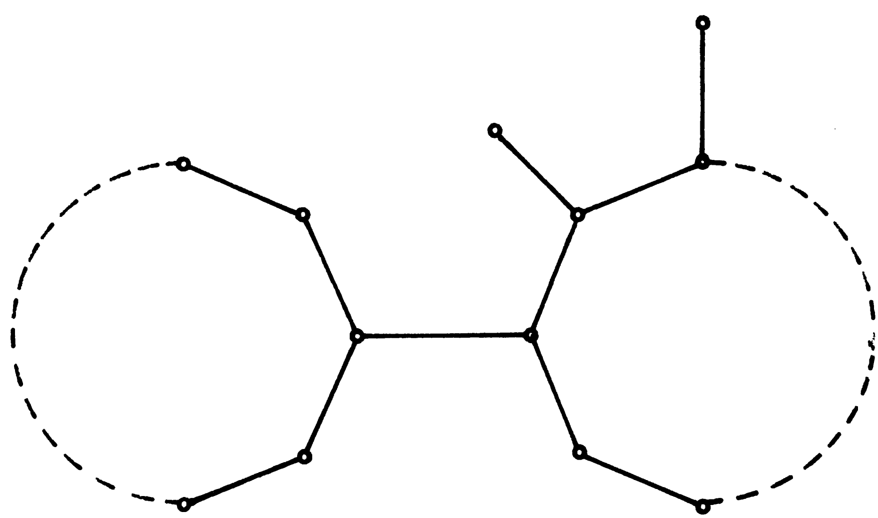


Cutting Free a Trail

FIGURE 6.

eulerian trail in the left Q_{n-1}, cutting the trail free at each point as before but do not cut the initial/terminal point free of the line joining the left and right cubes. Follow the same procedure in the right Q_{n-1}, but at the final occurrence of each point in the trail, do not cut the trail free from the lines which formerly joined the left and right cubes. The result, illustrated in FIGURE 7 is two cycles, one with some endlines, joined by a single line. If we now cut each of the points incident with the line joining the two cycles in such a way as to open the cycles, the resulting graph is an interval graph (in fact, a caterpillar as defined in **[10]**). Careful counting shows that each point of the original Q_n has been cut into either $(n-1)/2$ or $(n-1)/2+1$ points, so we have shown that $i(Q_n) \leqslant (n+1)/2$.

Combining this with the previous result for n even, we find that for all u, $i(Q_n) \leqslant [n/2]+1$. An essentially similar (though somewhat more complex) argument will in fact establish this same upper bound for any n-regular graph! The fact that this is also a lower bound for the cubes is quickly shown by noting that since Q_n contains no triangles, the interval graph H of the lemma can be assumed to contain no triangles, and then applying a simple line counting argument.



An Odd Cube Just Before the Final Cut.

FIGURE 7.

References

[1] S. Benzer, On the topology of the genetic fine structure, Proc. Nat. Acad. Sci. USA, 45(1959) 1607–1620.
[2] J. E. Cohen, Food Webs and Niche Space, Princeton Univ. Press, Princeton, 1978.
[3] D. R. Fulkerson and O. A. Gross, Incidence matrices and interval graphs, Pacific J. Math., 15(1965) 835–855.
[4] P. C. Gilmore and A. J. Hoffman, A characterization of comparability graphs and of interval graphs, Canad. J. Math., 16(1964) 539–548.
[5] J. R. Griggs and D. B. West, Extremal values of the interval number of a graph, SIAM J. Alg. Discrete Methods, to appear.
[6] J. R. Griggs, Extremal values of the interval number of a graph, II, Discrete Math., 28 (1979) 37–47.
[7] G. Hajós, Über eine Art von Graphen, Internat. Math. Nachr., 2(1957) 65.
[8] F. Harary, Graph Theory, Addison-Wesley, Reading, 1969.
[9] F. Harary and J. A. Kabell, Intersection graphs I: The computation of interval numbers, to appear.
[10] F. Harary and A. J. Schwenk, Trees with hamiltonian square, Mathematika, 18 (1971) 138–140.
[11] C. G. Lekkerkerker and J. Ch. Boland, Representation of a finite graph by a set of intervals on the real line, Fund. Math., 51(1962) 45–64.
[12] F. S. Roberts, Discrete Mathematical Models, Prentice-Hall, Englewood Cliffs, N.J., 1976.
[13] W. T. Trotter, Jr. and F. Harary, On double and multiple interval graphs, J. Graph Theory, 3 (1979) 205–212.

The Cobb-Douglas Production Function

GERALD BEER
California State University
Los Angeles, CA 90032

One concept that is invariably considered in a course in mathematical analysis for business or economics students is the notion of a production function for a single product firm. Such a function describes the maximum amount of output that the firm can produce in a fixed period of time as a function of the inputs (or factors of production) that are available to the firm. These factors of production have traditionally been placed in four categories: land, labor, capital, and entrepreneurship. Such a correspondence presupposes the existence of an "efficient manager" [**1**, **9**] who can somehow analyze the production alternatives of the firm with a particular factor allocation and can determine which of them results in maximal output. For example, a firm that manufactures automobiles might be able to utilize alternative production processes in different plants to assemble a vehicle. The production managers of the firm should be able to discern which process or combination of processes should be employed to obtain maximal output with a prescribed set of inputs. (This decision is frequently made using the tools of mathematical programming.) Assuming that the firm has this "efficient manager," the production function is determined by the technology (in the broad sense) that the firm has available. Technological innovation results in a new production function.

The aggregate output of the entire economy can also be modeled by a production function. The best known of these production functions is Cobb-Douglas function [**4**], originally constructed to approximate the output of American manufacturing from 1899 to 1922 as a function of the average number of employed wage earners and the value of fixed capital goods, reduced to dollars of constant purchasing power. We shall refer to these inputs as simply labor and capital. Output was measured by the Day index of physical production [**3**]. Cobb and Douglas assumed that the true production function could be closely approximated by a function of the form

References

[1] S. Benzer, On the topology of the genetic fine structure, Proc. Nat. Acad. Sci. USA, 45(1959) 1607–1620.
[2] J. E. Cohen, Food Webs and Niche Space, Princeton Univ. Press, Princeton, 1978.
[3] D. R. Fulkerson and O. A. Gross, Incidence matrices and interval graphs, Pacific J. Math., 15(1965) 835–855.
[4] P. C. Gilmore and A. J. Hoffman, A characterization of comparability graphs and of interval graphs, Canad. J. Math., 16(1964) 539–548.
[5] J. R. Griggs and D. B. West, Extremal values of the interval number of a graph, SIAM J. Alg. Discrete Methods, to appear.
[6] J. R. Griggs, Extremal values of the interval number of a graph, II, Discrete Math., 28 (1979) 37–47.
[7] G. Hajós, Über eine Art von Graphen, Internat. Math. Nachr., 2(1957) 65.
[8] F. Harary, Graph Theory, Addison-Wesley, Reading, 1969.
[9] F. Harary and J. A. Kabell, Intersection graphs I: The computation of interval numbers, to appear.
[10] F. Harary and A. J. Schwenk, Trees with hamiltonian square, Mathematika, 18 (1971) 138–140.
[11] C. G. Lekkerkerker and J. Ch. Boland, Representation of a finite graph by a set of intervals on the real line, Fund. Math., 51(1962) 45–64.
[12] F. S. Roberts, Discrete Mathematical Models, Prentice-Hall, Englewood Cliffs, N.J., 1976.
[13] W. T. Trotter, Jr. and F. Harary, On double and multiple interval graphs, J. Graph Theory, 3 (1979) 205–212.

The Cobb-Douglas Production Function

Gerald Beer
California State University
Los Angeles, CA 90032

One concept that is invariably considered in a course in mathematical analysis for business or economics students is the notion of a production function for a single product firm. Such a function describes the maximum amount of output that the firm can produce in a fixed period of time as a function of the inputs (or factors of production) that are available to the firm. These factors of production have traditionally been placed in four categories: land, labor, capital, and entrepreneurship. Such a correspondence presupposes the existence of an "efficient manager" [**1**, **9**] who can somehow analyze the production alternatives of the firm with a particular factor allocation and can determine which of them results in maximal output. For example, a firm that manufactures automobiles might be able to utilize alternative production processes in different plants to assemble a vehicle. The production managers of the firm should be able to discern which process or combination of processes should be employed to obtain maximal output with a prescribed set of inputs. (This decision is frequently made using the tools of mathematical programming.) Assuming that the firm has this "efficient manager," the production function is determined by the technology (in the broad sense) that the firm has available. Technological innovation results in a new production function.

The aggregate output of the entire economy can also be modeled by a production function. The best known of these production functions is Cobb-Douglas function [**4**], originally constructed to approximate the output of American manufacturing from 1899 to 1922 as a function of the average number of employed wage earners and the value of fixed capital goods, reduced to dollars of constant purchasing power. We shall refer to these inputs as simply labor and capital. Output was measured by the Day index of physical production [**3**]. Cobb and Douglas assumed that the true production function could be closely approximated by a function of the form

$$f(x,y)=Ax^{\alpha}y^{1-\alpha} \qquad (A>0 \text{ and } 0<\alpha<1)$$

defined on the first quadrant R^2_+ of the coordinate plane where x and y represent units of labor and capital, respectively. The method of least squares yields values of .75 for α and 1.01 for A. With these values for the parameters, the annual deviations of actual production from that predicted by the formula $f(x,y)=1.01x^{.75}y^{.25}$ average only 4.3% for the years 1899–1922. Moreover, in only one of these years does the error in the estimate exceed 11%.

In the 50 years since its introduction, the Cobb-Douglas function $f(x,y)=Ax^{\alpha}y^{1-\alpha}$ has been given broad exposure. In fact, it is now even a common topic in short calculus texts [**2,5,11**]. A major reason for its popularity is its nice mathematical properties. Most significantly, the Cobb-Douglas function is a concave function: for each pair of vectors $u_1=(x_1,y_1)$ and $u_2=(x_2,y_2)$ and each scalar λ in $[0,1]$,

$$f(\lambda u_1+(1-\lambda)u_2)\geq\lambda f(u_1)+(1-\lambda)f(u_2).$$

This algebraic condition is easy to visualize: f is concave if and only if the line segment joining each pair of points on the graph of f lies on or below the graph. Stated formally, f is concave if and only if its **hypograph**, the region $\{(x,y,z):z\leq f(x,y)\}$ under the graph of f, is convex, as shown in FIGURE 1. Similarly, a function f is called **convex** if $\{(x,y,z):z\geq f(x,y)\}$ is a convex set.

We mention two consequences of the concavity of an arbitrary bivariate production function f, both of which can be verified for the particular function $f(x,y)=Ax^{\alpha}y^{1-\alpha}$ without any knowledge of concave function theory. First, such a function is consistent with the law of diminishing returns: if one input is increased indefinitely while the other input quantity is held fixed, then output will eventually increase at a decreasing rate. Second, each level set $\{(x,y):f(x,y)\geq r\}$ is a convex set, and whenever (x,y) is a point in the level set, and $(a,b)\in R^2_+$, then $(x,y)+(a,b)$ is also contained in the level set [**7**]. Thus if inputs are increased, then optimal output will not decrease. For example, if a firm can achieve a certain optimal output with prescribed labor and capital inputs, then the acquisition of more labor will not interfere with production to the extent that the previous optimal output is no longer attainable. In economic jargon, we say that excess inputs are disposable [**8**]. A prototype level set of a concave production function appears in FIGURE 2.

The characteristic shape of the level sets of a concave production function f allows us to solve by geometric means an important optimization problem: to find a combination of labor and

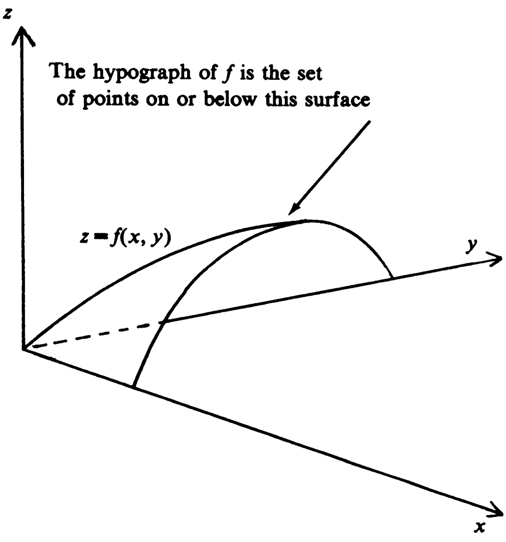


FIGURE 1.

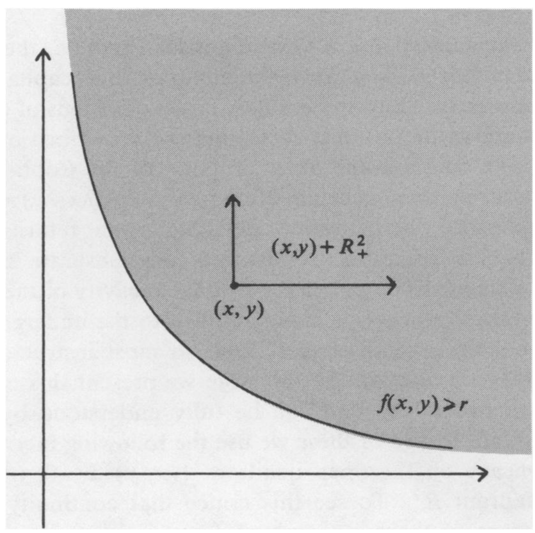


FIGURE 2.

capital that yields a prescribed output r at minimal cost. Suppose each unit of labor can be purchased for p_x dollars and each unit of capital for p_y dollars. The cost of acquiring x units of labor and y units of capital is then $C(x,y)=p_x x+p_y y$ dollars. It suffices to find an isocost curve $C(x,y)=c_0$ that is tangent to the constant production curve $f(x,y)=r$ i.e., an isocost curve that meets the constant production curve at a point (x_0,y_0) where they both have the same slope (see FIGURE 3a). Analytically, tangency occurs if the gradient $\nabla C=(p_x,p_y)$ is a positive scalar multiple of ∇f, which implies that

$$\frac{p_x}{p_y}=\frac{D_1 f(x_0,y_0)}{D_2 f(x_0,y_0)}.$$

This says, in economic terms, that the ratio of the marginal productivities of the factors equals the ratio of their respective prices. In mathematical terms, it is just one form of the method of Lagrange multipliers, well known to every student of advanced calculus. If the production function did not have convex level sets, the existence of a point of tangency need not imply minimal cost, as FIGURE 3b plainly shows.

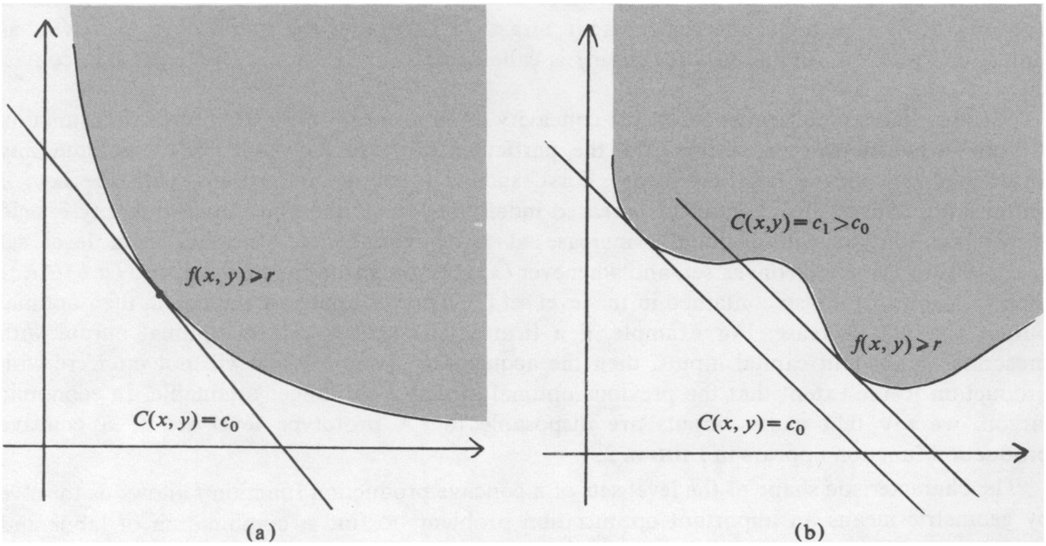


FIGURE 3.

Actually, if f is a Cobb-Douglas function, then for each positive r the constant production curve $\{(x,y):f(x,y)=r\}$ implicitly defines capital as a decreasing convex function of labor and conversely. Thus, increasingly larger additions of one input are needed to compensate for a fixed reduction in the other to maintain a prescribed output level.

We mention one other property of the Cobb-Douglas function that is trivial to verify. It is positively homogeneous of degree one: $f(\lambda x,\lambda y)=\lambda f(x,y)$ for each positive λ. Thus the production process so described exhibits constant returns to scale: a proportional increase in the inputs available to the firm results in a commensurate increase in the maximal possible output.

Despite the significance of the concavity of the Cobb-Douglas function, a formal verification of this property is seldom supplied to the undergraduate business or economics student, for the proof that would suggest itself to most instructors involves knowledge of linear algebra and advanced calculus. In this note we present this obvious but perhaps inappropriate proof along with two others that can be fully understood by a student with a modest quantitative background. In two of these we use the following fact: a continuous function f defined on R^2_+ that is concave on the open quadrant $\{(x,y):x>0,\ y>0\}$ must be concave throughout the closed quadrant R^2_+. To see this notice that continuity ensures that the hypograph of f will be the closure of the hypograph of f restricted to the interior of its domain. Since the closure of a convex set is convex, the hypograph of f must be a convex set. Hence, f is a concave function.

Proof 1. We need only show that f restricted to the interior of R_+^2 is a concave function. Since f has continuous second order partial derivatives on the interior, we can use the following well-known criterion [**6**, **7**, **9**]: f is concave if and only if the Hessian matrix H of f has nonpositive eigenvalues at each point. Since

$$\frac{\partial^2 f}{\partial x^2}=\alpha(\alpha-1)Ax^{\alpha-2}y^{1-\alpha},\qquad \frac{\partial^2 f}{\partial y^2}=\alpha(\alpha-1)Ax^{\alpha}y^{-1-\alpha},$$

$$\frac{\partial^2 f}{\partial x\partial y}=\alpha(1-\alpha)Ax^{\alpha-1}y^{-\alpha},$$

it is easily shown that

$$\det(H-\lambda I_2)=\lambda\left[\lambda-\alpha(\alpha-1)Ax^{\alpha-2}y^{-1-\alpha}(x^2+y^2)\right].$$

Since $0<\alpha<1$, H has one zero eigenvalue and one negative one.

Although this proof is short and direct, it constitutes a "black-box" technique for the intended audience. To appreciate it, the reader must be familiar with the diagonalization of symmetric matrices and quadratic forms as well as the chain rule for vector-valued functions.

Proof 2. In contrast to the first proof, our second method requires no mathematics beyond a semester of calculus. It is based on the application of a simple inequality: if (x_1,y_1) and (x_2,y_2) are in the first quadrant and $0<\alpha<1$, then

$$x_1^\alpha y_1^{1-\alpha}+x_2^\alpha y_2^{1-\alpha}\leqslant(x_1+x_2)^\alpha(y_1+y_2)^{1-\alpha}.$$

This clearly holds if either point lies on a coordinate axis. Otherwise, we follow the route suggested by Roberts and Varberg [**6**]. Since $g(t)=\log(1+e^t)$ has a positive second derivative, it is a convex function and we have

$$\log[1+e^{\alpha s+(1-\alpha)t}]\leqslant\alpha\log(1+e^s)+(1-\alpha)\log(1+e^t).$$

Since the exponential function is an increasing function, applying it to both sides of the last inequality yields

$$1+e^{\alpha s+(1-\alpha)t}\leqslant(1+e^s)^\alpha(1+e^t)^{1-\alpha}.$$

Next the substitutions $s=\log(x_2/x_1)$ and $t=\log(y_2/y_1)$ transform the above inequality into

$$1+(x_2/x_1)^\alpha(y_2/y_1)^{1-\alpha}\leqslant\left(1+\frac{x_2}{x_1}\right)^\alpha\left(1+\frac{y_2}{y_1}\right)^{1-\alpha}.$$

Multiplication of both sides of the last inequality by $x_1y_1^{1-\alpha}$ produces the desired result. Finally, we apply this inequality to verify that the Cobb-Douglas function is concave:

$$\begin{aligned}f[\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2)]&=A[\lambda x_1+(1-\lambda)x_2]^\alpha[\lambda y_1+(1-\lambda)y_2]^{1-\alpha}\\&\geqslant A\left[(\lambda x_1)^\alpha(\lambda y_1)^{1-\alpha}+([1-\lambda]x_2)^\alpha([1-\lambda]y_2)^{1-\alpha}\right]\\&=\lambda f(x_1,y_1)+(1-\lambda)f(x_2,y_2).\end{aligned}$$

Our third proof is based on geometry. It is similar to the second one in that no knowledge of advanced mathematics is required. However, unlike the second proof, it is intuitive and involves no tricks. The superiority of this method to the other two gives support to the following dictum: Whenever possible, attack problems in convexity with geometry.

Proof 3. Let $E=\{(x,y,z):f(x,y)\geqslant 1 \text{ and } z\leqslant 1\}$. Clearly, E is a subset of the hypograph of f. We mentioned earlier that $\{(x,y):f(x,y)=1\}$ implicitly defines capital as a convex function of labor; so, $\{(x,y):f(x,y)\geqslant 1\}$ is a convex set. It follows that E is a convex set, for it is the intersection of the vertical cylinder $\{(x,y,z):f(x,y)\geqslant 1\}$ with the half-space $\{(x,y,z):z\leqslant 1\}$.

Now the set of all open rays emanating from the origin through a convex set is again a convex set. We claim that (x,y,z) lies on a ray emanating from the origin through the convex set

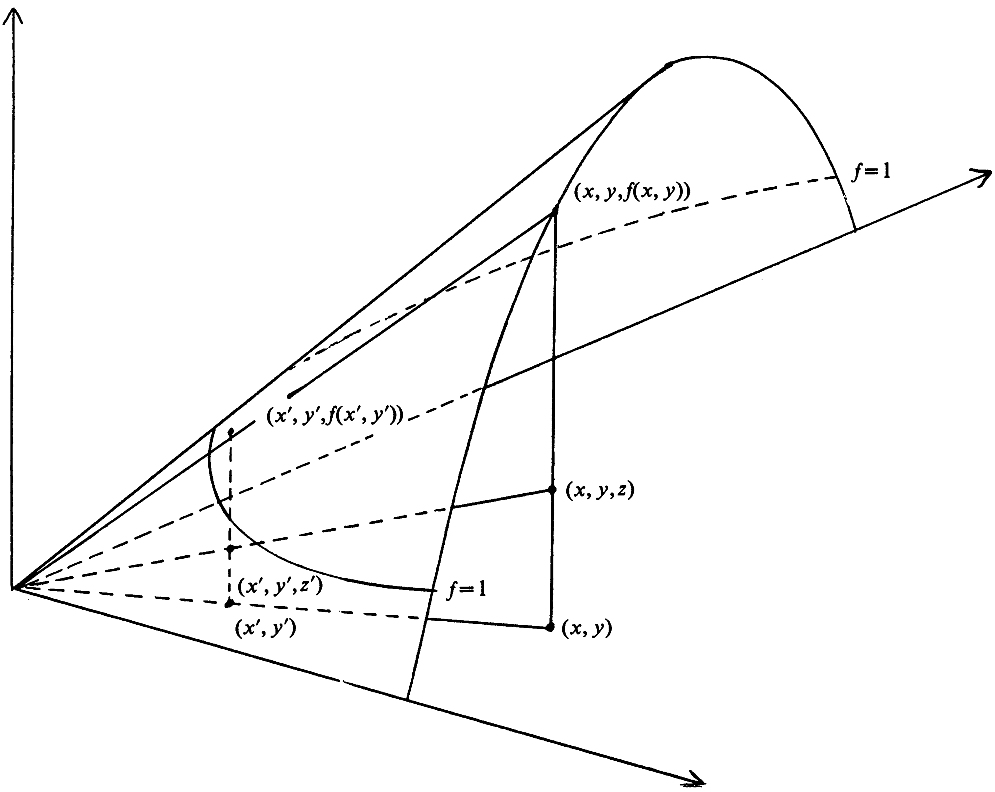


FIGURE 4.

E if and only if x and y are both positive and (x,y,z) belongs to the hypograph of f. To see this, suppose first that $(x,y,z)=t(x',y',z')$ where $t>0$ and (x',y',z') is in E. Since $z'\leqslant 1\leqslant f(x',y')$, the positive homogeneity of f ensures that $z\leqslant f(x,y)$ so that (x,y,z) is in the hypograph (see FIGURE 4). It is evident that x and y must be both positive because $\{(x,y):f(x,y)\geqslant 1\}$ is contained in the interior of the first quadrant.

Conversely, if (x,y,z) is in the hypograph and x and y are positive, then $z\leqslant f(x,y)$ and $f(x,y)>0$. Since $[f(x,y)]^{-1}z\leqslant 1$, the positive homogeneity of f again implies that $[f(x,y)]^{-1}(x, y,z)$ is in E. We conclude that (x,y,z) is a positive multiple of a member of E and the claim is established. This shows that f restricted to the interior of the first quadrant is a concave function; it follows, as noted above, that f is concave throughout R^2_+.

References

[1] W. J. Baumol, Economic Theory and Operations Analysis, Prentice-Hall, Englewood Cliffs, N. J., 1965.

[2] G. Beer, Applied Calculus for Business and Economics, Winthrop, Cambridge, Mass., 1978.

[3] E. E. Day and W. M. Persons, An index of the physical volume of production, Rev. Econ. and Stat., 2 (1920) 328–29.

[4] P. H. Douglas, The Theory of Wages, Macmillan, New York, 1934.

[5] L. Goldstein, D. Lay, and D. Schneider, Calculus and its Applications, Prentice-Hall, Englewood Cliffs, N. J., 1977.

[6] A. W. Roberts and D. E. Varberg, Convex Functions, Academic Press, New York, 1973.

[7] R. T. Rockafellar, Convex Analysis, Princeton Univ. Press, Princeton, N. J., 1970.

[8] R. W. Shephard, The Theory of Cost and Production Functions, Princeton Univ. Press, Princeton, N. J., 1970.

[9] J. Stoer and C. Witzgall, Convexity and Optimization in Finite Dimensions I. Springer-Verlag, New York, 1970.

[10] A. Takayama, Mathematical Economics, The Dryden Press, Hinsdale, Il., 1974.

[11] D. Zill, Introductory Calculus for Business, Economics, and Social Science, Wadsworth, Belmont, Ca., 1977.

PROBLEMS

DAN EUSTICE, Editor
LEROY F. MEYERS, Associate Editor
The Ohio State University

Proposals

To be considered for publication, solutions should be mailed before October 1, 1980.

1089. Determine the highest power of 1980 which divides

$$\frac{(1980n)!}{(n!)^{1980}}.$$

[*M. S. Klamkin, University of Alberta.*]

1090. It is well known that if n is prime, then for every pair of relatively prime integers a and b the gcd of $(a^n-b^n)/(a-b)$ and $(a-b)$ is 1 or n. Find a corresponding result valid for every integer $n \geq 1$ and every pair of distinct integers a and b. [*Tom M. Apostol, California Institute of Technology.*]

1091. Let (x,y) denote the gcd of integers x and y. If a and b are relatively prime integers with $a>b$, prove that for every pair of positive integers m and n we have

$$(a^m-b^m, a^n-b^n)=a^{(m,n)}-b^{(m,n)}.$$

[*Tom M. Apostol, California Institute of Technology.*]

1092. Let D be the disk $x^2+y^2<1$. Let the point A have coordinates $(r,0)$ where $0<r<1$. Describe the set of points P in D such that the open disk whose center is the midpoint of $\overline{AP}$ and whose radius is $AP/2$ is a subset of D. [*Roger L. Creech, East Carolina University.*]

ASSISTANT EDITORS: DON BONAR, *Denison University;* WILLIAM A. MCWORTER, JR., *The Ohio State University. We invite readers to submit problems believed to be new. Proposals should be accompanied by solutions, when available, and by any information that will assist the editors. Solutions to published problems should be submitted on separate, signed sheets. An asterisk (*) will be placed by a problem to indicate that the proposer did not supply a solution. A problem submitted as a Quickie should be one that has an unexpected succinct solution. Readers desiring acknowledgement of their communications should include a self-addressed stamped card. Send all communications to this department to Dan Eustice, The Ohio State University, 231 W. 18th Ave., Columbus, Ohio 43210.*

Solutions

Does X or Y know (x,y)? September 1978

1051. A game involves a quizmaster and two players, X and Y. The quizmaster chooses an ordered pair of real numbers (x,y) and tells x to player X and y to player Y. The quizmaster also tells the players that (x,y) is in the set $A=\{(x_i,y_i):i=1,2,\ldots,n\}$. The quizmaster then asks X and Y alternately if they know (x,y). Find a characterization of the set A which guarantees that either X or Y will eventually know (x,y). [*A. K. Austin, The University of Sheffield.*]

Solution: Each turn in the game, i.e., each statement by X and the corresponding statement by Y, must eliminate some ordered pairs of A from consideration if the game is to conclude. The game ends only when all pairs but one have been eliminated by previous statements and by knowledge of one coordinate. The statement "I don't know" by X eliminates any pairs (among those not already eliminated) for which the x_i is unique. Likewise the statement "I don't know" by Y eliminates any remaining pairs with unique y_i. The game will never conclude if a turn occurs in which both players say "I don't know" and no pairs are eliminated, for both will then continue answering *ad infinitum*. Thus in such a non-terminating game A is reduced to a subset in which each x and y occurs in more than one pair. Hence for successful termination A is characterized by the property that *every subset of A containing (x,y) must contain at least one pair with a unique x coordinate or a unique y coordinate.*

ZANE C. MOTTELER
Michigan Technological Institute
Houghton, Michigan

Also solved by S. F. Barger, Victor Pambuccian (Romania), and the proposer.

Isomorphic Boolean Rings September 1978

1052. Show that Boolean rings (idempotent commutative rings with identity) are isomorphic if their multiplicative semigroups are isomorphic. [*F. David Hammer, Santa Cruz, California.*]

Solution: Let f denote an isomorphism between the multiplicative semigroups of Boolean rigns R and R'. Let 0 and $0'$ denote the additive identities of R and R', respectively. Clearly, $f(0)=0'$. (Suppose $f(a)=0'$. $f(0)=f(a\cdot 0)=f(a)\cdot f(0)=0'\cdot f(0)=0'$.) For elements a and b in R, there is an element x in R such that

$$f(a+b)=f(a)+f(b)+f(x). \tag{1}$$

Multiplying both sides of (1) by $f(ab)$, we obtain

$$\begin{aligned} f(ab)f(a+b)&=f(ab)(f(a)+f(b)+f(x))\\ f(ab+ab)&=f(ab)+f(ab)+f(abx)\\ f(0)&=0'+f(abx). \end{aligned}$$

Thus, $abx=0$.

Multiplying both sides of (1) by $f(ax)$, we obtain

$$\begin{aligned} f(ax)f(a+b)&=f(ax)(f(a)+f(b)+f(x))\\ f(ax+abx)&=f(ax)+f(abx)+f(ax)\\ f(ax+0)&=0'. \end{aligned}$$

Thus, $ax=0$.

Multiplying both sides of (1) by $f(x)$, we obtain

$$f(x)f(a+b)=f(x)(f(a)+f(b)+f(x))$$
$$f(ax+bx)=f(ax)+f(bx)+f(x)$$
$$f(bx)=f(0)+f(bx)+f(x).$$

Thus, $f(x)=0'$ and so $f(a+b)=f(a)+f(b)$. f is an isomorphism between the rings R and R'.

WILLIAM F. FOX
Moberly Junior College

The solution above shows that the multiplicative identity is not needed.

Also solved by James B. Anderson, S. F. Barger, Philip M. Benjamin (Taiwan), Theodore S. Bolis, Duane M. Broline, Brian A. Davey (Australia), Elwyn H. Davis, F. J. Flanigan, Victor Pambuccian (Romania), St. Olaf College Problems Group, and the proposer.

Mean and Intermediate Value Properties — September 1978

1053. Let $f(x)$ be differentiable on $[0,1]$ with $f(0)=0$ and $f(1)=1$. For each positive integer n, show that there exist distinct $x_1, x_2, \ldots, x_n$ such that $\sum_{i=1}^{n} 1/f'(x_i)=n$. [*Peter Ørno, The Ohio State University*.]

Solution I: Set $a_0=0$ and $a_n=1$. Since $f(x)$ is differentiable on $[0,1]$, it is also continuous and by the Intermediate Value Property for each k, $1 \leqslant k \leqslant n$, there is an a_k such that $a_{k-1}<a_k \leqslant 1$ with $f(a_k)=k/n$. By the Mean Value Theorem there is an x_k, $a_{k-1}<x_k<a_k$, such that

$$f'(x_k)=\frac{f(a_k)-f(a_{k-1})}{a_k-a_{k-1}}=\frac{1}{n(a_k-a_{k-1})}.$$

Therefore $\sum_{k=1}^{n} 1/f'(x_k)=\sum_{k=1}^{n} n(a_k-a_{k-1})=n(a_n-a_0)=n$. Since the a_k are strictly increasing and $a_{k-1}<x_k<a_k$, the points x_k are distinct.

ROBERT CLARK, senior
Central High School
Philadelphia, Pennsylvania

Solution II: Consider the range R of $f'(x)$. By the Mean Value Theorem, 1 is in R. Either $f'(x)=1$ for all x, which results in a trivial solution, or R contains a value greater than 1 and a value less than 1. All intermediate values are also in R by the Intermediate Value Property for derivatives. Thus R contains some neighborhood N centered about 1. For each $a>1$ contained in N, there exists an $a'<1$ also in N such that $1/a+1/a'=2$. We can select $[n/2]$ of these pairs from N, adding the value 1 if n is odd. For these a_i, $\sum_{i=1}^{n} 1/a_i=n$. Each $a_i=f'(x_i)$ for a unique x_i, thus $\sum_{i=1}^{n} 1/f'(x_i)=n$.

JOHN AMPE, senior
Oak Park and River Forest H. S.
Oak Park, Illinois

Also solved by Peter Andrews, Theodore S. Bolis, Paul Bracken (Canada), T. Y. Chow, C. S. Davis (England), Santo Diano, Michael J. Dixon, Den Ducoff, Michael W. Ecker, Michael Finn, William F. Fox, David Hammer, George C. Harrison & Mou-Liang Kung, Johnny Henderson, Eli L. Isaacson, Lewis Lum, Russell Lyons, James McKim & Gerald Wildenberg, Zane C. Motteler, Victor Pambuccian (Romania), Steve Ricci, St. Olaf Problems Group, M. S. K. Sastry, Richard Schauer, Bill Scherer, Bob Sien, Dale Smith, J. M. Stark, Gerald Thompson, Ronald S. Tiberis, Stanley Wagon, and the proposer. There were two unsigned solutions.

1054. (a*) Show how to construct triangle ABC by straightedge and compass, given side a, the median m_a to side a, and the angle bisector t_a to side a.

(b*) Show how to construct triangle ABC by straightedge and compass, given angle A, m_a, and t_a. [*Jerome C. Cherry, Santa Maria, California.*]

Solution: (a) We shall give an algebraic solution. Place the given side a on the x-axis of a rectangular Cartesian frame of reference, with the midpoint of the side at the origin. For convenience we choose our unit of distance equal to $a/2$ and designate m_a and t_a more simply by m and t, respectively. Let (x,y) denote the coordinates of vertex A, and $(-z,0)$ those of the foot of the angle bisector t. Then, utilizing the distance formula of analytic geometry and the fact that the angle bisector divides side a into parts proportional to the other two sides of the triangle, we have

$$x^2+y^2=m^2, \tag{1}$$

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

$$\frac{(x+1)^2+y^2}{(x-1)^2+y^2}=\frac{(1-z)^2}{(1+z)^2}. \tag{3}$$

Setting $y^2=m^2-x^2$, from (1), in (2) and (3), and simplifying, we obtain

$$2xz+z^2+m^2=t^2, \tag{4}$$

$$\frac{1+2x+m^2}{1-2x+m^2}=\frac{(1-z)^2}{(1+z)^2}. \tag{5}$$

Eliminating x from (4) and (5) we obtain, after simplifying,

$$z^4-(t^2+m^2+1)z^2+(m^2-t^2)=0, \tag{6}$$

a quadratic in z^2. By standard constructions, we can construct segments of lengths t^2+m^2+1 and $(m^2-t^2)^{1/2}$. Again, by standard constructions we can construct a segment of length z^2, and then one of length z. Once z is found, the sought triangle is easily constructed.

Note. The devotee of the game of Euclidean constructions is not really interested in the actual mechanical construction of the sought triangle, but merely in the assurance that the construction is possible. To use a phrase of Jacob Steiner, the devotee performs his construction "simply by means of the tongue," rather than with actual instruments on paper. As soon as equation (6) is achieved, and recognized as a quadratic in z^2 with suitably constructible coefficients, the problem is finished. From this point of view, the problem is, in reality, essentially a pillow problem.

(b) Again we shall give an algebraic solution. Place vertex A at the origin of a rectangular Cartesian frame of reference, with t_a lying along the positive x-axis. To simplify the notation, designate t_a and m_a more simply by t and m. Let k, $-k$, s denote the slopes of the sides of angle A and of an arbitrary line L through the point $(t,0)$, and let (x_m,y_m) denote the coordinates of the midpoint of the segment cut off on line L by the sides of angle A. The equation of L and of the degenerate conic made up of the sides of angle A are, respectively,

$$y=s(x-t) \quad \text{and} \quad y^2=k^2x^2.$$

Eliminating y we obtain

$$x^2(s^2-k^2)-2s^2tx+s^2t^2=0. \tag{7}$$

Designating the roots of quadratic (7) by x_1 and x_2, we find

$$x_m = \frac{x_1 + x_2}{2} = \frac{s^2 t}{s^2 - k^2} \tag{8}$$

whence

$$y_m = \frac{sk^2 t}{s^2 - k^2}. \tag{9}$$

Eliminating s from (8) and (9) we obtain

$$y_m^2 = k^2 x_m^2 - k^2 t x_m. \tag{10}$$

But we also have

$$x_m^2 + y_m^2 = m^2. \tag{11}$$

Eliminating y_m from (10) and (11) we obtain

$$x_m^2(k^2+1) - k^2 t x_m - m^2 = 0. \tag{12}$$

But, since $k = \tan(A/2)$, this reduces to

$$x_m^2 - t(\sin^2(A/2)x_m - m^2(\cos^2(A/2)) = 0. \tag{13}$$

Simple constructions yield segments of lengths $t(\sin^2(A/2))$ and $m(\cos(A/2))$. A standard construction then yields x_m. Once x_m is found, the sought triangle is easily constructed.

Note. As in the former problem, the devotee quits the game as soon as equation (13) is attained. To find the point (x_m, y_m) we seek an intersection of a hyperbola and a circle. In general, to find an intersection of a conic and a circle is beyond the Euclidean tools. In our case, however, the circle is specially placed with respect to the hyperbola—its center lies at a vertex of the hyperbola.

HOWARD EVES
University of Maine

Also solved by Walter Bluger (Canada), Theodore D. Bolis, Erhard Braune (Austria), the Case Western Reserve University Problem Solving Group, Jordi Dou (Spain), Alexander Oliveira, Paul Smith (Canada), and the proposer. One incorrect solution was received. Braune supplied a reference (for part (a)) to Problem E1915, Amer. Math. Monthly, 75 (1968) 190. J. Garfunkel supplied a reference (for part (b)) to Problem E1375, Amer. Math. Monthly, 67 (1960) 185–186. Thébault's solution to E1375 was purely synthetic.

Bolis, the CWRU PSG, and Dou found necessary and sufficient conditions for solution. They are: for part (a), $m_a > t_a$ and either $2m_a \leq a$ or $4m_a(m_a - t_a) < a^2 < 4m_a(m_a + t_a)$; for part (b), $m_a > t_a$ and $0 < A < \pi$.

Binomial Distribution **November 1978**

1055. Find the limit as $n \to \infty$ of

$$a_n(p) = \sum_{k=0}^{n} \binom{2n+1}{k} p^k (1-p)^{2n+1-k}, \; 0<p<1.$$

[*Andreas N. Philippou, University of Patras.*]

Solution: (i) If $p = 1/2$, then $a_n(p) = \Sigma_{k=0}^{n} \binom{2n+1}{k} / 2^{2n+1} = 1/2$.

(ii) If $p>1/2$, then

$$0\leqslant a_n(p)\leqslant p^n(1-p)^{n+1}\sum_{k=0}^{n}\binom{2n+1}{k}=4^np^n(1-p^{n+1})=b_n(p).$$

But for $p>1/2$, $4p(1-p)<1$. Thus $b_n(p)\to 0$ as $n\to\infty$ and therefore $a_n(p)\to 0$ as $n\to\infty$.

(iii) If $p<1/2$, then from the identity $a_n(p)=1-a_n(1-p)$, it follows that $a_n(p)\to 1$ as $n\to\infty$.

PHILIP TODD, graduate student
Georgia Institute of Technology

Also solved by Theodore Bolis, Paul Bracken (Canada), Michael Chamberlain, Michael W. Ecker, Peter Flusser & Bill Scherer, John A. Gillespie, Jerrold W. Grossman, G. A. Heuer, J. D. Hiscocks (Canada) Robert A. Leslie, Robert F. Ling, Roger B. Nelsen, B. Pedler (Australia), Eberhard L. Stark (Germany), J. M. Stark, Glenn T. Vickers (England), and the proposer. Several solvers used the fact that $a_n(p)$ is the probability that there will be more failures than successes in $2n+1$ Bernoulli trials and applied the Central Limit Theorem. E. Stark used the fact that $a_n(p)$ is the $(2n+1)$th Bernstein polynomial for the characteristic function of $[0,1/2]$ and applied the convergence theorem for such expansions.

Measure for Measure **November 1978**

1056. "Oh, drat!" exclaimed the meteorologist stormily. "I've just anchored my new rain gauge onto a cement post, and it seems to be crooked."

"What does your rain gauge look like?" asked his friend, the math student.

"It's in the shape of a circular cylinder 8 centimeters in diameter with height-markings all around its sides. Its axis is only 3 degrees off-vertical, but this will affect the amount of rain entering the top, and besides, which height-marking should I use? The water-level will look tilted. I'm very discouraged about this whole business."

"Are you interested in measuring extremely light rains?" asked his friend.

"Not really. Anything less than a half-centimeter is too hard to measure accurately anyway, so I just record it as being a 'trace of precipitation'."

"I think I can help you," said the math student.

Tell the meteorologist how to correct the readings on his crooked rain gauge. [*Daniel A. Moran, Michigan State University*.]

Solution: Let a and b be the largest and smallest readings on the tilted rain gauge. $(a+b)/2$ is invariant with respect to the angle of tilt from the vertical position, provided that b can be read. Vertical rainfall meets an opening with effective area $16\pi\cos 3°$. Let h be the reading of a vertical rain gauge with the same specifications. Then $h(16\pi\cos 3°)=16\pi(a+b)/2$, i.e., $h=((a+b)/2)\sec 3°$. This is the rainfall to be reported.

Minimum reading $b=0$ occurs when $a=8\tan 3°$. Then $h=4\tan 3°\sec 3°<1/2$ and our meteorologist can simply report a "trace of precipitation" if b can't be read.

A. DICKSON BRACKETT
SUNY at Oneonta

Also solved by Thomas E. Elsner, Howard Eves, Kenneth M. Gustin, R. E. King, Lew Kowarski, P. Pedler (Australia), and the proposer. Eves also suggested making the bottom level by pouring in a sufficient amount of some hard-setting waterproof liquid.

REVIEWS

Paul J. Campbell, Editor
Beloit College

Pierre Malraison, Editor
Control Data Corp.

Assistant Editor: Eric S. Rosenthal, Princeton University. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of the mathematics literature. Some reviews of books are adapted from the Telegraphic Reviews in the American Mathematical Monthly.

Hofstadter, Douglas R., Gödel, Escher, Bach: An Eternal Golden Braid, Basic Books, 1979; xxi + 777 pp, $18.50.

Subtitle: "A metaphorical fugue on minds and machines in the spirit of Lewis Carroll." This is the most exciting and most brilliant popular book on mathematics to appear in the last 25 years. For a synopsis of its contents, see Martin Gardner's column in the July 1979 *Scientific American*, pp. 16-24. A paperback edition of the book is forthcoming.

Kolata, Gina Bari, *Cathleen Morawetz: The mathematics of waves*, Science 206 (12 Oct. 1979) 206-207.

Biography and career of an applied mathematician at the Courant Institute, "one of the very few women of her generation to succeed in mathematics."

Efron, Bradley, *Computers and the theory of statistics: Thinking the unthinkable*, SIAM Review 21:4 (October 1979) 460-480.

Survey for "nonstatisticians" (who however have a command of undergraduate mathematical statistics) of some advances in statistics. Topics are chosen to illustrate how the advent of the computer has affected the growth of statistical theory; they include nonparametric methods, the jackknife, the bootstrap, cross-validation, robustness, and censored data. The 'unthinkable' mentioned in the title is simply the thought that one might be willing "to perform 500,000 numerical operations in the analysis of 16 data points..."

Gardner, Martin, *Mathematical games: The random number omega bids fair to hold the mysteries of the universe*, Scientific American 241:6 (November 1979) 2-34, 206.

This article wrestles with what it means for an integer (or real number) to be random. "Incompressibility" turns out to be the key idea: a number is random if its shortest description (in terms of a Turing machine program) is about as long as the number itself, in binary bits. One such random number is Ω, the probability that a universal Turing machine will halt on random input (Ω depends on the particular machine). "[I]f the first thousand digits of Ω were known, they would, at least in principle, suffice to decide most of the interesting open questions in mathematics."

d'Espagnat, Bernard, *The quantum theory and reality*, Scientific American 241:6 (November 1979) 158-181, 206.

"The doctrine that the world is made up of objects whose existence is independent of human consciousness turns out to be in conflict with quantum mechanics and with facts established by experiment." One or more of realism, inductive inference, and Einstein separability (no influence can propagate faster than the speed of light) must fail, as together they lead to a simple set-theoretic inequality contradicted by experiment.

Clements, M.A. (Ken), *Sex differences in mathematical performance: An historical perspective*, Educational Studies in Math. 10:3 (August 1979) 305-322.

Comparison of attitudes in England and Australia toward mathematics for women, 1835-1912.

Batschelet, E., et al., *On the kinetics of lead in the human body*, J. Math. Biology 8 (1979) 15-23.

The authors set up a conceptually simple model resulting in three linear nonhomogeneous differential equations. They estimate parameters from field data and solve the system numerically (using eigenvalues). Finally, they investigate dynamics of lead concentrations under hypothetical changes in environmental exposure. Altogether, the article illustrates mathematical modelling at its best.

Morris, Scot, *Games: Festschrift for the master gamesman*, Omni 2 (October 1979) 176-177, 163.

Devoted to Martin Gardner, this month's column offers anecdotes, "magic" tricks, and a selection of 10 of Gardner's problems.

You cannot be a twentieth century man without maths, The Economist, October 27, 1979, pp. 107-114.

A wide ranging survey of contemporary mathematics, touching on calculus, topology, gauge symmetries, game theory, catastrophe theory, and fuzzy data, with constant commentary on how such esoteric topics are important to modern business. "If society is not to be split into a magic circle of mathematicians and a mass of laymen, manipulated and bemused, the challenge of post-medieval maths must be faced."

Bradley, Milton N., *The game of Go--the ultimate programming challenge?* Creative Computing 5:3 (March 1979) 90-99.

An excellent summary of the rules of Go, complete with examples and a very nice bibliography. Beginning suggestions for programming Go on small boards in a mini-computer.

Bloor, David, *Polyhedra and the abominations of Leviticus*, British J. History Science 11 (1978) 245-272.

Bloor brings together two books with the common theme of how people respond to things which do not fit the classifications and organizing principles of accepted ways of thinking. One is Imre Lakatos' *Proofs and Refutations: The Logic of Mathematical Discovery* (Cambridge U Pr, 1976); the other is an anthropological work, Mary Douglas' *Natural Symbols: Explorations in Cosmology* (Harmondsworth, 1973). Bloor applies Douglas' "grid-group" concepts to Lakatos' analysis of the history of the Euler-Descartes conjecture about polyhedra (v-e+f=2), and the result is an important contribution to a sociological approach to mathematics.

Kolata, Gina Bari, *Trial and error game that puzzles fast computers,* Smithsonian 10 (November 1979) 90-96.

A very elementary introduction to hard (NP-complete) problems, with applications to unbreakable codes.

Robson, Ernest and Wimp, Jet, Against Infinity: An Anthology of Contemporary Mathematical Poetry, Primary Press, 1979; v + 90 pp, $17, $8.95 (P).

An attempt to bridge the "irreconcilable disparity between the backgrounds and interests of poets and mathematicians."

Brams, Stephen J., Spatial Models of Election Competition, EDC/UMAP, 1979; v + 94 pp, $4 (P).

Adapted from the author's *The Presidential Election Game,* with the addition of twenty-odd exercises, this short monograph is intended for class use in an analytically-oriented political science course; only high-school mathematics is assumed.

Dertouzos, Michael L. and Moses, Joel (Eds.), The Computer Age: A Twenty-Year View, MIT Press, 1979; xvi + 491 pp, $25.

Notable among these 20 essays probing prospects for the individual, for society, and for technology are the contributions of N.P. Negroponte ("The return of the Sunday painter") and H.A. Simon ("The consequences of computers for centralization and decentralization").

Bezuszka, Stanley, et al., Perfect Numbers, Boston College Press, 1980; 169 pp, (P).

One of a series of "Motivated Math Project Activity Booklets" written for students and teachers. One half the book is devoted to tables, including the first 600 each of abundant and deficient numbers, and all 27 known Mersenne primes and corresponding perfect numbers, in their entirety. (The 27th perfect number has not quite 27,000 digits.)

Andree, Josephine P. and Andree, Richard V., Sophisticated Ciphers, CRYPTO Project (Room 423, 601 Elm, Norman, OK 73019), 1978; 149 pp, (P). (Preliminary trial edition.)

One of a series of five mini-courses on problem-solving and logical thinking; the other four include two more elementary booklets on ciphers and one each on cryptarithms and logical thinking. This booklet offers up-to-date information on developments since 1970, plus examples drawn from U.S. Civil War messages. An instructor's manual is available.

Renfrew, Colin and Cooke, Kenneth L. (Eds.), Transformations: Mathematical Approaches to Culture Change, Academic Press, 1979; xxii + 514 pp, $39.50.

This volume is concerned with the historical development of human societies, and the contributors are mostly mathematicians and archaeologists. The models employed include optimization, dynamical systems, simulation, catastrophe theory, and weighted digraphs.

Neugebauer, G.H., *Fitting curves to data,* Machine Design 52 (6 Sept. 1979) 91-95.

Least-squares fits "applied to workaday problems." Particularly interesting is the discussion of non-linear curves that translate to linear curves when graphed on appropriate paper.

Lehman, Hugh, Introduction to the Philosophy of Mathematics, Rowman & Littlefield, 1979; xi + 177 pp, $18.

Do mathematical entities exist? On what basis can we claim to arrive at mathematical knowledge? These are the two primary questions in mathematical philosophy. Not surprisingly, Lehman answers "yes" to the first; to the second, he asserts that mathematical knowledge is empirical--not intuitive, hypothetical (formal), or analytic. The book is blessed with clear writing, good organization, and a very detailed table of contents.

Booss, Berhelm and Niss, Mogens (Eds.), Mathematics and the Real World, Birkhäuser, 1979; 136 pp, $22 (P).

Proceedings of an International Workshop at Roskilde University (Denmark). What is a real, good, important mathematical problem? What are the mathematical needs of developing countries? How can new mathematics be better communicated to those who might need it? What should mathematicians do about the deepening currents of pragmatism ("Abstract mathematics is so to speak difficult to sell today and seems not to be able to lead to a quick useful reward and high profit rates.")? The authors describe a malaise, a syndrome; but the diagnosis is vague and a cure uncertain.

Boltianskii, V.G., Hilbert's Third Problem. Transl: R.A. Silverman. V.H. Winston, 1978; xii + 228 pp, $19.95.

Hilbert's Third Problem is nice, because like Fermat's Last Theorem it has a fairly elementary statement: show that the volume of a pyramid does not depend solely on its base and its height. Even nicer, the problem has been solved. Boltianskii leads the reader (assumed to be an advanced undergraduate or graduate student) from precise definitions of area and decomposition of polygons to corresponding notions for volume and decomposition of polyhedra and a clear modern proof.

Bueker, Robert (Ed.), Professional Opportunities in the Mathematical Sciences, Tenth Edition, MAA, 1978.

This booklet should be handy on the desk of every mathematics teacher and placed in the hands of every student contemplating a career in a mathematical science.

Moore, David S., Statistics: Concepts and Controversies, Freeman, 1979; xv + 313 pp, $6.95 (P).

Excellent either as text or supplement for an elementary statistics course, depending on one's goals. Computation and technique are de-emphasized (no t-tests, no chi-squared) in favor of teaching students to *think*--about sampling, experimentation, measurement and index numbers, as well as the traditional visual representations of data, descriptive statistics, plus probability and statistical reasoning. Examples are taken from current news materials; exercises compose 30% of the book; and the style is crisp, personal, and enjoyable. An informative instructor's manual is available.

Yaglom, I.M., A Very Simple Non-Euclidean Geometry and Its Physical Basis, Springer-Verlag, 1979; xviii + 307 pp, $19.80.

The geometry is "Galilean," that associated with the Galilean principle of relativity that no mechanical experiment conducted within a physical system can disclose uniform motion of the system. The author paints a beautiful and fascinating picture, interweaving geometry and kinematics and comparing the simple Galilean geometry with its more complicated Euclidean and non-Euclidean cousins. The book is remarkably rich in ideas, and the exposition is elegant.

NEWS & LETTERS

HEINZ BAUER RECEIVES 1980 CHAUVENET PRIZE

Professor Heinz Bauer of the University of Erlangen-Nürnberg was awarded the 1980 Chauvenet Prize for noteworthy exposition for his paper "Approximation and Abstract Boundaries" which appeared in the *American Mathematical Monthly* 85 (1978) 632-647. The prize, represented by a certificate and a check for $500 is the twenty-eighth award of the Chauvenet Prize since its inception by the Mathematical Association of America in 1925. It was presented to Professor Bauer at the January 1980 annual meeting of the M.A.A. in San Antonio, Texas.

Bauer was born January 31, 1928, at Nürnburg, Germany. Between 1948 and 1953, he studied at the University of Erlangen and the Université de Nancy, France. He was awarded the Ph.D. (summa cum laude) by the University of Erlangen in 1953 and, until 1956, served as Assistant Professor there. He has held Associate and Full Professorships at the University of Hamburg and, since 1965, has served as Full Professor at the University of Erlangen-Nürnberg.

Professor Bauer is involved in research in integration theory, functional analysis (convexity and approximation theory), potential theory, and Markov processes. He is the author of more than fifty papers and three books, and has served as editor of both *Inventiones Mathematicae* (1966-79) and *Math. Annalen* (1970-).

The paper for which Professor Bauer received the Chauvenet Prize discusses three famous theorems of P.P. Korovkin that concern uniform approximation of functions. David P. Roselle, secretary of the M.A.A., noted that these theorems are presented in a well-chosen setting and are superbly illuminated with a collection of examples and applications. The paper is accessible to any graduate student who has learned about the Lebesgue integral.

JOB OPENING: ASSOCIATIVE EXECUTIVE DIRECTOR FOR THE MATH. ASSOC. OF AMER.

The Mathematical Association of America has an opening for an Associate Executive Director at its headquarters in Washington, D.C., beginning July 1, 1980 or as soon as possible thereafter. Applicants should have interest and ability in administration and should be willing to exercise independent initiative. A Ph.D. in mathematics or a mathematical science is desirable. Salary will be commensurate with the experience and training of the candidate.

Applicants should send a curriculum vitae, and should also arrange to have three letters of recommendation sent directly to:

Prof. Dorothy L. Bernstein, President
Mathematical Association of America
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

The Mathematical Association of America is an equal opportunity employer.

HENRY ALDER RECEIVES 1980 DISTINGUISHED SERVICE AWARD

The Mathematical Association of America bestowed its Award for Distinguished Service to Mathematics on Henry L. Alder of the University of California at Davis during its annual meeting in San Antonio, Texas. In presenting the award, M.A.A. secretary David Roselle read the following citation:

> Henry has had tremendous and profound influence on all of mathematics and, more than any other person, he has charted the course of the Mathematical Association of America for the past twenty years. There are mathematicians who have added greatly to mathematics through their research activities.

Still others have contributed through their teaching. Finally, some mathematicians are willing and energetic workers on behalf of the professional organizations. But there are few among us who fill all three of these rolls. Thus, it is appropriate that we today honor Professor Henry L. Alder, an able researcher, an honored teacher, and a man who has made unparalleled contributions of time and energy to the professional activities of mathematicians.

Born in Duisburg, Germany, in 1922, Henry Alder moved with his family to Zürich, Switzerland, in 1933. After graduating from the Kantonschule in Zürich in 1940, he studied chemistry for a semester at the Eidgenössische Technische Hochschule. When his family moved to the United States in 1941 he enrolled immediately in the University of California at Berkeley, where he received his A.B. degree in 1942 and, after service in the U.S. Air Force, his Ph.D. in mathematics in 1947. In 1948 he joined the faculty of the University of California at Davis where he rose through the ranks, becoming Professor of Mathematics in 1965.

Henry Alder served as Secretary of the M.A.A. from 1960 through 1975, and as President in 1977-78. He is currently a member of the CBMS Council, of the United States Commission on Mathematical Instruction, of the Program Committee for the International Congress on Mathematical Education IV, and is Vice-Chairman of the Council of Scientific Society Presidents.

STATISTICS CONFERENCE

The Eighth Annual Mathematics and Statistics Conference at Miami University, Oxford, Ohio, will be held September 26-27, 1980. The theme for this year's conference will be "Statistics." Featured speakers will include Robert V. Hogg of the University of Iowa and William H. Lawton of the Eastman Kodak Company.

There will be sessions of contributed papers and a poster session which should be suitable for a diverse audience of statisticians, mathematicians, and students. Abstracts should be sent by June 1, 1980 to Professor John Skillings, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056. Information regarding preregistration and housing may also be obtained from Professor Skillings.

CONFERENCE ON UNDERGRADUATE MATHEMATICS

The fifth annual conference on undergraduate mathematics will be held on March 14 and 15, 1980, at Fort Lewis College in Durango, Colorado. Sponsored by the *Journal of Undergraduate Mathematics*, the meeting will include papers written by undergraduate students together with papers presented by the following faculty: William E. Brown, University of the Pacific; Burton W. Jones, University of Colorado; Robert A. Rubin, Whittier College; and Fred Stevenson, University of Arizona.

NIMBI ADDENDUM

D. Brooke has brought to our attention that Last-Player-Lose (LPL) and Last-Player-Win (LPW)-"Nimbi" were posed as a problem by K. Scherer ("Nimbi," *J. Rec. Math.* 9 (1976-77) 212; and 10 (1977-78) 226). Scherer's "Nimbi" is played the same as the game Nimbi discussed on pp. 21-25 of this issue of *Mathematics Magazine*, except that in the former the tokens removed from any row need not be contiguous. Thus "Nimbi" is just another variation of 2-dimensional Nim defined on p. 26 of our note. Having used a computer, M.R.W. Buckley noted (in *J. Rec. Math.* 11 (1978-79) 218-219) that LPL-"Nimbi" is a first-player-win, whereas LPW-"Nimbi" is a second-player-win. (The arguments are omitted because of their length, according to an editorial comment.)

Aviezri S. Fraenkel
Weizmann Institute of Science
Rehovot, Israel

Hans Herda
Boston State College
Boston
Massachusetts 02115

1979 WILLIAM LOWELL PUTNAM MATHEMATICAL COMPETITION

A-1. Find positive integers n and $a_1, a_2, \ldots, a_n$ such that

$$a_1 + a_2 + \ldots + a_n = 1979$$

and the product $a_1 a_2 \ldots a_n$ is as large as possible.

A-2. Establish necessary and sufficient conditions on the constant k for the existence of a continuous real valued function $f(x)$ satisfying $f(f(x)) = kx^9$ for all real x.

A-3. Let $x_1, x_2, x_3, \ldots$ be a sequence of nonzero real numbers satisfying

$$x_n = \frac{x_{n-2}x_{n-1}}{2x_{n-2} - x_{n-1}} \quad \text{for } n = 3, 4, 5, \ldots .$$

Establish necessary and sufficient conditions on x_1 and x_2 for x_n to be an integer for infinitely many values of n.

A-4. Let A be a set of $2n$ points in the plane, no three of which are collinear. Suppose that n of them are colored red and the remaining n blue. Prove or disprove: there are n closed straight line segments, no two with a point in common, such that the endpoints of each segment are points of A having different colors.

A-5. Denote by $[x]$ the greatest integer less than or equal to x and by $S(x)$ the sequence $[x], [2x], [3x], \ldots$. Prove that there are distinct real solutions α and β of the equation $x^3 - 10x^2 + 29x - 25 = 0$ such that infinitely many positive integers appear both in $S(\alpha)$ and in $S(\beta)$.

A-6. Let $0 \le p_i \le 1$ for $i = 1, 2, \ldots, n$. Show that

$$\sum_{i=1}^{n} \frac{1}{|x - p_i|} \le 8n\left(1 + \frac{1}{3} + \frac{1}{5} + \ldots + \frac{1}{2n-1}\right)$$

for some x satisfying $0 \le x \le 1$.

B-1. Prove or disprove: there is at least one straight line normal to the graph of $y = \cosh x$ at a point $(a, \cosh a)$ and also normal to the graph of $y = \sinh x$ at a point $(c, \sinh c)$.

B-2. Let $0 < a < b$. Evaluate

$$\lim_{t \to 0} \left\{ \int_0^1 [bx + a(1-x)]^t \, dx \right\}^{1/t}$$

[The final answer should not involve any operations other than addition, subtraction, multiplication, division, and exponentiation.]

B-3. Let F be a finite field having an odd number m of elements. Let $p(x)$ be an irreducible (i.e., nonfactorable) polynomial over F of the form

$$x^2 + bx + c, \qquad b, c \in F .$$

For how many elements k in F is $p(x) + k$ irreducible over F?

B-4. (a) Find a solution that is not identically zero, of the homogeneous linear differential equation

$$(3x^2+x-1)y'' - (9x^2+9x-2)y' + (18x+3)y = 0.$$

Intelligent guessing of the form of a solution may be helpful.

(b) Let $y = f(x)$ be the solution of the nonhomogeneous differential equation

$$(3x^2+x-1)y'' - (9x^2+9x-2)y' + (18x+3)y = 6(6x+1)$$

that has $f(0) = 1$ and $(f(-1) - 2)(f(1) - 6) = 1$. Find integers a, b, c such that $(f(-2) - a)(f(2) - b) = c$.

B-5. In the plane, let C be a closed convex set that contains $(0,0)$ but no other point with integer coordinates. Suppose that $A(C)$, the area of C, is equally distributed among the four quadrants. Prove that $A(C) \le 4$.

B-6. For $k = 1, 2, \ldots, n$ let $z_k = x_k + iy_k$, where the x_k and y_k are real and $i = \sqrt{-1}$. Let r be the absolute value of the real part of

$$\pm\sqrt{z_1^2 + z_2^2 + \ldots + z_n^2}\,,$$

Prove that

$$r \le |x_1| + |x_2| + \ldots + |x_n| .$$

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036

MATHEMATICS MAGAZINE VOL. 53, NO. 1, JANUARY 1980