THE AMERICAN MATHEMATICAL MONTHLY

Volume 107, Number 7 August-September 2000

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1495 East 100 South, Room 142
University of Utah
Salt Lake City, UT 84112-1114

To submit a manuscript to the MONTHLY, please send your email address and postal address along with 3 copies of the complete manuscript (including all figures with captions and lettering) typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper. Authors who use LaTeX are urged to use article.sty, or a similar generic style, and its standard environments with no custom formatting. The MONTHLY does not accept electronic submissions, so please do not email your manuscript files. See recent articles in the MONTHLY for the style of citations for journal articles and books. When they are available, LaTeX or TeX files are used as an aid to composition, but they are not used to create camera ready copy.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

The complete online MONTHLY, archive at *www.jstor.org* is a valuable resource for both authors and readers; it may be searched online in a variety of ways for any specified keyword(s). MAA members whose institutions do not provide JSTOR access may obtain individual access for a modest annual fee; call 800-331-1622.

See the MONTHLY section of MAA Online for current information such as contents of issues and descriptive summaries of forthcoming articles:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
DONALD ALBERS, Director of Publications

Advertising Correspondence:
Dave Riska, *driska@maa.org*

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center, *maahq@maa.org*

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Recent copies of the MONTHLY are available for purchase through the MAA Service Center, maahq@maa.org, 1-800-331-1622

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT.

Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Defendens Imperium Romanum*: A Classical Problem in Military Strategy

Charles S. ReVelle and Kenneth E. Rosing

INTRODUCTION AND PROBLEM DESCRIPTION. In the third century of the Common Era (CE), when Rome dominated not only Europe, but also North Africa and the Near East, it was able to deploy fifty legions throughout the empire. In this *forward defense* strategy even the furthermost areas of the empire were secured by the on-site presence of an adequate number of legions of the Roman army. However, the empire had lost much of its muscle by the fourth century CE and the forces of Rome had diminished to only about twenty-five legions. It had thus become impossible to station legions in sufficient strength at all of the forward positions of the empire without abandoning the core.

A new *defense in depth* strategy was devised by the Emperor Constantine (Constantine The Great, 274–337) to cope with the reduced power of the empire [**11**]. His defense in depth used local troops to disrupt invasion and deployed mobile Field Armies (FAs) to stop and throw back the intruding enemy, or to suppress insurrection. The earlier *forward defense* strategy had provided a wall around the empire denying any but the most modest of incursions—it even allowed Roman forces to sally into barbarian lands to disrupt invasions as they were being mounted. In place of Rome's forces, the defense in depth strategy substituted local part-time militias (who would be fighting for their own land and families) to slow and fragment any invading barbarian army until the heavier weight of an FA, dispatched from a distant area, could be brought to bear.

Each set of roughly six legions with ancillary cavalry, artillery, etc. forms an FA, a unit of forces whose numbers are sufficient to secure any one of the regions of the empire [**2**]. In the third century CE, Rome's fifty legions or about eight FAs could be allocated so that each of the eight provinces was secured by its own FA. However, by the fourth century CE, only four FAs were available for deployment. The regions of the empire are considered to be connected as shown in Figure 1, where each region is represented as a circle (node). Movement along a line (edge) between regions (nodes) represents a "step" and for a region to be *securable*, an FA must be able to reach it in just one step [**2**].

A region is considered to be *secured* if it has one or more FAs stationed in it already. On the other hand, the region may be *securable*—that is, an FA may be capable of deploying to protect that region in a single step, but only under a special condition. An FA can be deployed from one region to an adjacent region *only* if it moves from a region where there is at least one other FA to help launch it. This is analogous to the island-hopping strategy pursued by General MacArthur in World War II in the Pacific Theater—movement followed the chain of islands already secured by troops left behind.

*Defending the Roman Empire

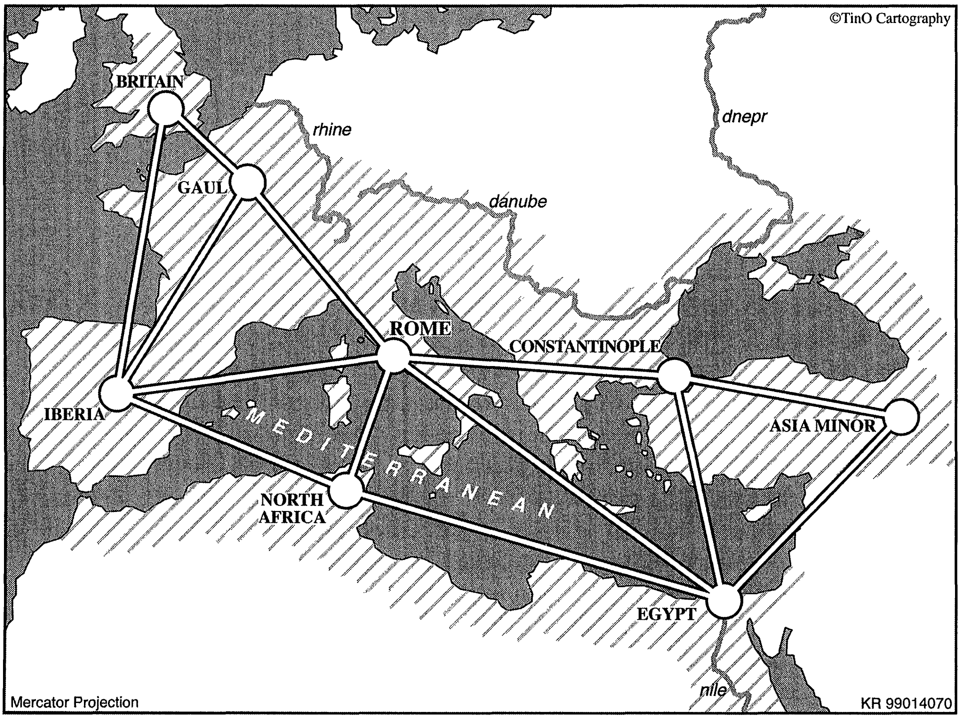


Figure 1. The Empire of Constantine and its Eight Provinces

The challenge for Constantine was to allocate just four FAs to positions in the eight regions of the empire. Constantine chose to place two at Rome, a symbolic as well as strategic choice, and two at his new capital, Constantinople. With this deployment, each region of the empire was either already secured or could be reached by an FA in just one step—with the exception of Britain. To reach Britian with an FA required an FA to move from Rome to Gaul, securing Gaul. Then, a second FA needed to be launched from Constantinople to Rome. Only then could an FA shift from Rome to Gaul, and finally an FA could move from Gaul to Britian, a total of four steps for the field armies. Each of the four steps began from a base that had two FAs present.

Here is another alternative, not necessarily better than Constantine's strategy, but illustrative of the moves that are legitimate. Suppose we place one FA in Gaul, two in Rome, and one in Constantinople. Britian can now be reached in two steps, which consist of moving an FA from Rome to Gaul and moving an FA from Gaul to Britian—better for Britian than before. However, Asia Minor is now *not* reachable in one step, but requires two steps: Rome to Constantinople and Constantinople to Asia Minor. All the rest of the empire is just one step away. It is not clear that this is better than Constantine's strategy. Although the number of steps to the worst-off region has been reduced to two, the number of regions that are more than one step away from any assistance has gone from one to two.

Can a modern analyst do better than Constantine's solution? "Better" may be measured with respect to several criteria. One criterion is the number of regions that cannot be reached in a single step; this number is to be reduced. For Constantine's solution, as we have just seen, that number is one. Another criterion is the number of steps it takes to reach the worst-off node. Again, for Constantine's choice, this number is four, the number of moves needed to reach Britian with assistance. To do better than Constantine, one would need to do better with respect to one of the criteria without degradation of the other. For example, relative to the first criterion, an improvement would be to reach *every* region in just one step. Alternatively, if one can reduce the number of steps to reach the worst-off node *without* increasing the number of regions that are more than one step away, the solution will be better. That is, if an analyst can keep the number of nodes that can't be reached in one step to just one, and can reduce the maximum number of steps to reach the worst-off node to *fewer than* four, then the solution will be better than Constantine's. And of course, if all regions were made either initially secure or reachable in one step, then the empire would be fully protected.

Still a third criterion concerns the consequences of a second war occurring somewhere in the empire. We might want to minimize the number of regions that can't be reached or secured in the event of a second war or, conversely, we might want to maximize the number of regions secured or covered in one step in the event of a second war.

We may envision the fundamental problem as having at least two phases. In the first phase, the number of FAs required may be an unknown, and we ask the question, "What is the least number of FAs to be placed, and where should they be sited, so that all regions of the empire are either secured or securable?" In the second phase, we ask "How should a limited number of FAs (the number available may be insufficient to secure or make securable all of the regions) be deployed to achieve optimally some security objective such as the maximum number of regions made secure or securable?" We present several models that address these criteria in the remainder of this paper.

BASIC FORMULATIONS. The formulations we present next belong to the class of $0,1$ optimization problems that seek yes-or-no location decisions at discrete points. These discrete siting problems, which most often consider distances measured on a network, constitute one of two classes of location problems. The other class typically chooses sites from an infinite space of alternative locations and often makes use of an Euclidean or other distance measure, Together, these two classes of problems make up the family of problems referred to as Location Science or Topothesiology. Several current texts and collections survey these problems and the methods used to solve them; see [**6**], [**7**]. or [**10**].

The Set Covering Deployment Problem. The first formulation is called the Set Covering Deployment Problem (SCDP). This formulation is a novel derivative of a well-known problem, the Location Set Covering Problem ([**16**], [**15**], and [**13**]). In the Location Set Covering Problem with demand nodes and eligible facility sites scattered on a network or a plane, the problem is to find, and site, the least number of facilities so that all points of demand have at least one facility within some distance standard.

In the SCDP formulation, each region must either be *secured* by one or more FAs or *securable* by an FA that can reach the region in a single step from a two-FA region. We seek the least number of FAs to distribute among the regions so that all regions are either secured or securable.

We let:

I = the set of demand areas/deployment sites;

$x_i = 1, 0$; the variable is 1 if region i contains one or two FAs and 0 otherwise;

$y_1 = 1, 0$; the variable is 1 if region i contains two FAs, and 0 otherwise;
$N_i = \{j \mid \text{region } j \text{ is one step from region } i\}$

The objective is

$$\text{Minimize } z = \sum_{i \in I} x_i + \sum_{i \in I} y_i$$

Subject to:

$$x_i + \sum_{j \in N_i} y_j \geq 1 \qquad \text{for every } i \in I \tag{1}$$

$$y_i \leq x_i \qquad \text{for every } i \in I \tag{2}$$

The constraints (1) say that every region must be secured or securable. If $x_i = 1$, then region i is secured—without regard to the presence of any regions one step away that contain two FAs. If $x_i = 0$, then at least one of the regions one step away from region i must have two FAs—this to make region i securable in a single step. Of course, it is possible that for some i, $x_i = 1$ and also there is at least one region one step away from region i that has two FAs. Nonetheless, one of these events is required to occur for each $i \in I$. The constraints (2) say that there cannot be two FAs in a region unless there is at least one FA in the region.

The objective function of the SCDP is the number of regions with one or two FAs plus the number of regions with two FAs. This is exactly the number of FAs deployed. For example, suppose there is just one FA in region A and two FAs in region B. The value of the first term of the objective function is two since both regions A and B have either one or two FAs. The value of the second term is one since only region B has two FAs. The objective function is $2 + 1 = 3$. Minimizing the objective function minimizes the number of FAs deployed subject to the constraints.

If it were felt that an FA could not be isolated (i.e., must have at least one other FA within one step), then we could modify the problem by adding an additional constraint:

$$\sum_{j \in N_i} x_j \geq x_i \qquad \text{for every } i \in I \tag{3}$$

The Maximal Covering Deployment Problem. There is also a natural problem that is complementary to the SCDP. Its goal is to allocate a limited number of FAs to the regions in such a way as to maximize the number of regions that are securable in a single step or already secured by the presence of an FA. We call this problem the Maximal Covering Deployment Problem (MCDP). The MCDP is also a relative of a widely known problem, the Maximal Covering Location Problem [**3**]. The Maximal Covering Location Problem resembles the Location Set Covering Problem in that it assumes that demand nodes and eligible facility sites are dispersed on the plane or network. In contrast to the Location Set Covering Problem, it seeks to site a limited number of facilities in such a way that the greatest number of demand nodes has one or more facilities within a distance standard.

The logical constraints of the MCDP fix in advance the number of FAs, but say that an area is "covered" only if it has an FA present on site (it is secured) or if there is at least one position only one step away that has two FAs located there (it is securable). Coverage of every region is not required but is sought as a goal. One new type of variable needs to be introduced for the MCDP. It is:

$u_i = 1, 0$: the variable is 1 if region i is secured or securable in a single step. it is 0 otherwise.

The problem as structured can be stated as:

$$\text{Maximize } Z = \sum_{i \in I} u_i$$

Subject to:

$$u_i \leq x_i + \sum_{j \in N_i} y_i \qquad \text{for all } i \in I \tag{4}$$

$$y_i \leq x_i \qquad \text{for all } i \in I \tag{5}$$

$$\sum_{i \in I} x_i + \sum_{i \in I} y_i = p \; (p \text{ Field Armies}) \tag{6}$$

Of course, a positive coefficient reflecting the relative importance of each region i could multiply each u_i term in the objective.

The maximal covering deployment model doesn't deal with how far away (the number of steps) any of the uncovered nodes are. That is, if a node is not covered in one step, there is no requirement that it can be covered in two steps, or three steps, etc. Furthermore, the model as stated does not deal with the issue of a second war.

Several methods are available to "solve" integer programs. Each method's utility varies with the characteristics of the program. We choose to solve the MCDP as a relaxed linear program and then impose the integer requirements if necessary. This means that the three sets of zero-one variables are relaxed in the linear program and are allowed to range continuously between zero and one. The replacement definitions are: $0 \leq x_i \leq 1$; $0 \leq y_i \leq 1$; and $0 \leq u_i \leq 1$. The resulting program can be solved easily by any of the numerous linear programming packages available. Some or all of the variables in the solution obtained via linear programming may lie strictly between zero and one. Such non-zero-one solutions are resolved by the technique of branch and bound, an add-on option generally packaged with linear programming solvers.

Branch and bound begins with the linear programming solution and fixes one non-zero-one variable first to 1 and then to 0. Both of the two resulting problems, the one fixing the variable to 0, the other fixing it to 1, are then solved using linear programming. This process of successively fixing non-zero-one variables to 0 or 1 is called *branching* and produces a bifurcated or tree-like structure of sequential solutions. The branching process continues until an integer feasible solution is found.

The objective value of such an integer feasible solution forms a *bound* for the problem. If other nodes are branched from and if solutions with objective values higher than the bound result, none of these solutions can be optimal. In a minimising problem, all solution nodes (solutions that have not yet been branched on) with more costly objectives whose variables are not all 0, 1 can be cut off. That is, they need not be searched further because they can never develop an integer feasible solution whose objective value is less than the one already found. Other integer-infeasible solutions on the tree may be branched from until either a feasible solution with a lower value of the objective is developed (giving a new bound) or until all developed solutions have objective values that exceed the objective value of the best integer solution. Branch and bound cannot be counted on to resolve all problems to 0, 1 solutions efficiently. It has shown itself to be effective on many 0, 1 location problems [**12**], but for certain types of constraint sets, the number of nodes that must be resolved is so huge that practical limits on computer memory or time preclude the use of this method.

We solved the MCDP problem for the empire as shown in Figure 1, with $p = 4$ FAs to be allocated among the eight regions. Six alternate optima were found, all of which protected the entire empire—all regions of the empire were either secured or securable in a single step. Of course, these alternatives could also have been obtained by complete enumeration (evaluating all possibilities) in this small problem or by solving the SCDP. Only one of the six alternatives placed FAs at Rome.

Dantzig cuts were used to generate six alternate optima. Dantzig discovered and developed the technique of cutting away (making invalid) unwanted or integer infeasible solutions in the context of the famous Knapsack Problem [**5**]. These cuts, which we added in the re-solution of the linear programming problem, simply require that the current solution be excluded in all subsequent analysis. In order to apply the cuts, the zero-one variables that are 1 in the current solution are summed, and their sum, in a new problem, is constrained to be 1 less than the number of such variables.

Of the alternate optimal solutions produced by solving the MCDP, the first all zero-one solution (0 cuts, see Table 1) placed two FAs in Iberia and two FAs in Egypt. We then solved the MCDP a second time, adding a constraint that says the sum of the x_i plus the sum of the y_i in Iberia and Egypt is less than or equal to three. This constraint excludes no possibilities other than the previous solution. The result is the second solution (cut 1) listed in Table 1—namely, two FAs in Iberia and two in Constantinople. To expose further alternate optimal solutions would require not only a constraint that excludes the first solution, but one that excludes this second solution as well. That is, we add a constraint that says that the sum of FAs in Iberia and Constantinople is less than or equal to three. The cuts concluded when the first sub-optimal solution was found. In total, six alternate optimal deployments of four FAs were found, each deployment securing or making securable all of the eight regions of the empire.

The optimal solutions of Constantine's problem are shown in Table 1. The first column, headed "cuts", indicates the number of cut constraints required to get the particular solution shown on that particular line. The first line (0 cuts) shows the solution found with no additional constraints. The last column, headed "B & B", shows the number of branch and bound nodes that had to be evaluated in order to arrive at an optimal integer feasible solution. The remaining columns refer to the

TABLE 1. SOLUTIONS: ROMAN EMPIRE

cuts	BRI	IBE	GAU	NAF	ROM	EGY	CON	AMI	B & B
0		2				2			0
1		2					2		1
2		2						2	3
3	1				2			1	3
4	2					2			4
5			2			2			14

BRI = Britian, IBE = Iberia, GAU = Gaul, NAF = North Africa,
ROM = Rome, EGY = Egypt, CON = Constantinople, AMI = Asia Minor

eight regions of the Empire and each is headed by a three-letter abbreviation that is clarified in an underline to the table. The numbers in the body of the table indicate the number of FAs stationed in the various regions. Each of the following tables has the same form.

Discussion of the Alternate Optima of the MCDP. The single "Roman solution" deployed two FAs in Rome, one FA in Britian, and one in Asia Minor (Table 1, cut 3). Despite the advantage of placing legions in Rome (hence the term "Roman solution"), this deployment suffers from a reduced capability to respond to a second war occurring somewhere else in the empire. Looking at Figure 1 and envisioning this positioning of FAs, one can see that if a war occurred in any of the five unsecured regions, the response to that war would then leave four regions without protection in the event of a second war. That is, no FA could reach the remaining four regions in a single step. In fact, no FA could be launched anywhere in the empire as no two-FA regions would remain.

On the other hand, the equal optimal to the "Roman solution", which consists of two FAs in Iberia and two FAs in Egypt (Table 1, cut 0) performs better than the Roman solution when it comes to protection in the event of a second war. The worst case situation for this deployment would be a first war in Rome itself or in North Africa. It does not matter whether the response to this first war comes from Iberia or from Egypt. In either situation, two regions would be out of reach in a single step. This is a better outcome than occurs in the Roman solution where up to four regions could not be reached in the event of a second war. The as yet unanswered question is how to discover these robust solutions, robust in the sense that they do well even in the event of a second war.

FURTHER COMPUTATIONAL EXPERIENCE

Pax Britannica. Until well into the 19th century Britian possessed sufficient resources to keep many capital ships in all key regions of interest. With (generally) six "Battle Fleets" (BFs), each composed of roughly twenty ships of the line, Britian pursued a forward defense policy for some 150 years. Around the end of the 19th century, with declining British power, the change to steam propulsion, and the rise of Germany as a maritime power, this strategy had to be revised [**9**]. Even though the increasing mobility and power of the modern ship allowed the reduction of the BF to around eight capital ships each (with their support ships and auxiliaries), Britian had only four BFs in 1900 [**2**]. With six key regions this necessitated a shift to a defense in depth. First Sea Lord John Fisher brought three BFs to home (British) waters and stationed one BF in the Mediterranean. His principle reason was the heightening European tension and the increasing naval powers of Germany [**8**].

The rules of movement of BFs are the same as those of the FAs of the Roman Empire. One BF must be present in a region to launch a second. Now however, since we are all at sea, a step is counted as passing through one of the key regions. The six key regions are identified in Table 2 and their "one step contiguity" is shown in Figure 2; the identifiers of the key regions are across the top and down the left side. An "X" indicates that one region is directly reachable from the other, a one-step move for the BFs.

The strategy of Fisher results in only three of the six regions (Britian, The Mediterranean, and The West Indies) being secured or securable in one step. The worst off region (The Far East) is five steps away [**2**].

TABLE 2. SOLUTIONS: BRITISH EMPIRE

cuts	WIN	BRI	MED	CGH	SAS	FEA	B & B
0			2	2			0
1		1	1	2			0
2		2		2			1
3		2			2		3
4			2	1		1	2
5			2			2	0
6	2					2	8
7		2				2	7
8	2					2	4
9	1	1			2		5
10	2				1	1	6
11			2		2		6
12	2				2		14

WIN = West Indies, BRI-Britian, MED = Mediterranean,
CGH = Cape of Good Hope, SAS = South Asia, FEA = Far East

	WIN	BRI	MED	CGH	SAS	FEA
WIN		X	X	X		
BRI	X		X			
MED	X	X			X	
CGH	X				X	X
SAS			X	X		X
FEA				X	X	

Figure 2. Contiguity Matrix, British Empire. For clarification of abbreviations see Table 2.

Thirteen alternate optimal solutions to the MCDP are displayed in Table 2. All key regions of The British Empire are either secure with the presence of a BF or they can be reached by a BF, launched from a two-BF region, in one step. The greater flexibility (larger number of optimal configurations) is a function of the reduction in the number of regions that have to be covered from eight (Roman Empire) to six (British Empire).

Pax Americana. An application of the MCDP to the strategic network of regions of import to the USA demonstrates that the method is not merely of historical interest. While the USA is not being forced into a defense in depth strategy by declining economic and political power, in this post-colonial, post-right-by-conquest period, certain foreign stationings are problematic or impossible.

The definition of the Unit of Force (UF, a unit analogous to the FA or BF) for the Post-Cold War United States with its plethora of military arms is more problematic. Arquilla and Fredricksen [**2**] base their calculations on a study by Aspin [**1**] (a former secretary of defense). They arrive at four UFs, each consisting of roughly three divisions of infantry, three carrier battle groups, and five air wings with their required auxiliaries and support units.

The 15 equal optimal solutions are given in Table 3 and the one-step contiguity pattern on which the solutions are based is shown in Figure 3. This figure indicates the one-step moves possible for UFs. In all of the solutions in Table 3 the "empire" is either secured or securable with one step from a region hosting two UFs. The large number of equal optimals is hardly surprising when we are dealing with four UFs and a total of only five regions. The regions are named in the underline to the associated Table (Table 3).

TABLE 3. SOLUTIONS: AMERICAN EMPIRE

cuts	USA	EUR	NEA	SAS	EAS	B & B
0	2				2	0
1	2			1	1	1
2	2	1		1		1
3	1	1		2		2
4	2			2		2
5			2	1	1	3
6			2	2		5
7	2		1	1		6
8	1		2		1	1
9	2		2			5
10			2		2	9
11		1	2		1	12
12		1	1		2	12
13		2		1	1	13
14		2			2	12
15		2		2		15

USA = United States, EUR = Europe, NEA = Near East,
SAS = South Asia, EAS = East Asia

	USA	EUR	NEA	SAS	EAS
USA		X	X		X
EUR	X		X		
NEA	X	X		X	
SAS			X		X
EAS	X			X	

Figure 3. Contiguity Matrix, American Empire. For clarification of abbreviations see Table 2.

CONCLUSIONS. We have structured a classical problem on the deployment of military forces as a pair of 0, 1 programming problems. As far as we are aware, neither the Set Covering Deployment Problem nor the Maximum Covering Deployment Problem has previously been defined. Once a new problem is formally stated, other applications frequently occur. For example, several applications have appeared for the Maximum Covering Location Problem, a model developed for locating emergency service in order to maximize the population that can be served within a distance standard; see **[4]** and **[14]**.

We also demonstrate the utility of Dantzig cuts to reveal alternate optimal solutions by making infeasible each solution as it is found. We are not aware of similar uses of Dantzig cuts to expose a sequence of alternate optimal solutions (and stopping when the first sub-optimal is found). This technique could, however, also be useful to explore the sub-optimal region in the immediate neighborhood of the optimal solution. In such a case, as each cut is added, the next-best sub-optimal solution would be produced. The technique can be used to back off from the optimal solution or explore alternate optima in other 0, 1 problems as well.

In the computational experience we have shown that the relaxed linear programming version of the problem solves in 0, 1 variables with either no branch and bound or with only modest amounts of it. In general, we have observed that the greater the number of cuts appended, the greater the amount of branch and bound required to resolve the integer infeasibilities (see Table 1–3).

Perhaps most interesting from the standpoint of combinatorial optimization, however, is the fact that a mathematical program has effectively solved a puzzle of some fame. The reduction of the puzzle to a mathematical programming form suggests the possibility that other puzzles—some of which may have significant applications—have related types of formulations.

REFERENCES

1. L. Aspin, *Report on the BOTTOM-UP REVIEW*, Government Printing Office, Washington, D.C., 1993.
2. J. Arquilla and H. Fredricksen, 'Graphing' an Optimal Grand Strategy, *Mil. Oper. Res. J.* 1 (1995) 3–17.
3. R. Church and C. S. ReVelle, The Maximal Covering Location Problem, *Papers Reg. Sci. Assoc.* 32 (1973) 101–118.
4. C. Chung, Recent Application of the Maximal Covering Location Planning Model, *J. Oper. Res. Soc.* 37 (1986) 735–746.
5. G. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
6. M. Daskin, *Network and Discrete Location: Models, Algorithms, Applications*, Wiley-Interscience, New York, 1995.
7. Z. Drezner, ed., *Facility Location: A Survey of Applications and Methods*, Springer, New York, NY, 1995.
8. J. A. Fisher, *Memories*, Hodder and Stoughton, London, 1919.
9. P. Kennedy, *The Rise and Fall of British Naval Mastery*, Ashfield Press, London, 1976.
10. R. Love, J. Morris, and G. Wesolowsky, *Facilities Location: Models and Methods*, North-Holland, Amsterdam, 1988.
11. E. N. Luttwak, *The Grand Strategy of the Roman Empire*, Johns Hopkins University Press, Baltimore, MD, 1976.
12. C. S. ReVelle, Facility Siting and Integer-Friendly Programming, *Eur. J. Oper. Res.* 65 (1993) 329–342.
13. C. S. ReVelle, C. Toregas, and L. Falkson, Applications of the Location Set Covering Problem, *Geog. Analy.* 8 (1976) 65–77.
14. D. Schilling, J. Vaidyanathan, and R. Barkhi, A Review of Covering Problems in Facility Location, *Loc. Sci.* 1 (1993) 25–56.
15. C. Toregas, and C. S. ReVelle, Binary Logic Solutions to a Class of Location Problems, *Geog. Analy.* 5 (1973) 145–156.
16. C. Toregas, R. Swain, C. S. ReVelle, and L. Bergmann, The Location of Emergency Services, *Oper. Res.* 19 (1971) 1363–1373.

CHARLES S. REVELLE is a professor at The Johns Hopkins University. He took his bachelor's and doctoral degrees at Cornell University. He finds endless entertainment in formulating 0-1 programming problems that describe real situations. He has created 0-1 problems in the field of water management, location science, biological conservation strategies, and forestry management science. He has also written with his wife, Penelope, five undergraduate textbooks on environmental science as well as a book on water resources management and a textbook on mathematical programming. "A window and a light will do for thinking."
Department of Geography and Environmental Engineering
313 Ames Hall, The Johns Hopkins University, Baltimore, MD 21218
revelle@jhu.edu

KENNETH E. ROSING is a recently retired former Reader in Economic Geography, Erasmus University Rotterdam, who received his Ph.D. (jointly) from King's College and University College. He has worked with computers, particularly for the visual representation of data (maps) since the mid 60's, when he began writing algorithms for the first computer generated atlas (published 1970). He believes that research exists for itself and that the serendipital spin-offs of research more than justify the investment. His avocation is developing inventive new tricks to force computers to solve the integer models developed by Professor ReVelle. "A window and a light will do for thinking" (but a Cray in the back room sure helps to check the answers).
Van Polanenpark 23, 4791 BP Klundert, The Netherlands
rosing@few.eur.nl

The Books Studied by Ramanujan in India

Bruce C. Berndt and Robert A. Rankin

1. INTRODUCTION. Since Ramanujan died in 1920, mathematicians have discussed and debated the question, "From which books did Ramanujan learn mathematics while in India?" In his beautiful book highlighting some of Ramanujan's contributions to mathematics, G. H. Hardy devoted almost three pages to this question [**5**, pp. 42–44]. During Ramanujan's short lifetime, his friends and peers evidently did not ask him about the sources of his mathematical knowledge. About this, Hardy [**5**, p. 11] wrote, "I saw Ramanujan almost every day, and could have cleared up most of the obscurity by a little cross-examination. ... I hardly asked him a single question of this kind; I never even asked him whether (as I think he must have done) he had seen Cayley's or Greenhill's *Elliptic Functions*. I am sorry about this now, but it does not really matter very much, and it was entirely natural. In the first place, I did not know that Ramanujan was going to die. He was not particularly interested in his own history or psychology; he was a mathematician anxious to get on with the job. And after all I too was a mathematician, and a mathematician meeting Ramanujan had more interesting things to think about than historical research. It seemed ridiculous to worry him about how he had found this or that known theorem, when he was showing me half a dozen new ones almost every day."

Although the present writers cannot settle all the controversy, we can offer some insights. Our task here is twofold. In the next section, we list the books that Ramanujan definitely had seen, and those that Ramanujan likely knew. For each book, we provide evidence to back up our claim. In the final section, we describe each book and offer a biography of each author. We emphasize that our deliberations are centered only on what Ramanujan might have read in India before departing for England in March, 1914. Of course, after his arrival at Cambridge, he was exposed to considerably more literature.

2. THE BOOKS AND THE EVIDENCE. The first five books listed are ones that Ramanujan definitely had read (at least in part); the next three are ones with which Ramanujan was likely familiar.

S. L. Loney, *Plane Trigonometry*
G. S. Carr, *A Synopsis of Elementary Results in Pure Mathematics*
J. Edwards, *Differential Calculus*
B. Williamson, *An Elementary Treatise on the Integral Calculus*
G. H. Hardy, *Orders of Infinity*
A. G. Greenhill, *The Applications of Elliptic Functions*
A. Cayley, *An Elementary Treatise on Elliptic Functions*
G. B. Mathews, *The Theory of Numbers*

Ramanujan's contemporay biographers, for example, P. V. Seshu Aiyar and R. Ramachandra Rao [**8**, p. xii], record that Ramanujan had borrowed a copy of the second part of Loney's *Plane Trigonometry* from an older student. There is no reason to doubt the veracity of these accounts. Loney's book was a very popular text in India then and was so for many years thereafter.

Similarly, it is well known [**8**, p. xii] that Carr's *Synopsis* served as Ramanujan's primary source of mathematical knowledge. In a future article, we plan to examine this influence carefully. Ramanujan published five papers before leaving for England, and in two of these Ramanujan cited theorems from the second part of Carr's book [**8**, pp. 8, 15].

In Ramanujan's first published paper, he made two references to Edwards's *Differential Calculus* [**8**, pp. 1, 7].

About ten months before leaving for England, Ramanujan received a scholarship from the University of Madras. A stipulation in the scholarship was that Ramanujan was to write *Quarterly Reports* on his research progress. A large portion of the second report is devoted to extensive generalizations of Frullani's integral theorem for which Ramanujan cited Williamson's text [**1**, p. 313]. In his second published paper, Ramanujan also cited Williamson's *Integral Calculus*, here for the duplication formula for the gamma function.

In his first letter to Hardy, written on 16 January 1913, Ramanujan wrote that he had read in Hardy's *Orders of Infinity* [**8**, p. xxiii], [**4**, p. 21] "that no definite expression has been as yet found for the number of prime numbers less than any given number." P. V. Seshu Aiyar, Ramanujan's mathematics teacher at the Government College at Kumbakonam, claims that he had suggested to Ramanujan that he read Hardy's tract [**8**, p. xxii]. However, in a letter [**4**, p. 121] of 17 December 1914 to C. N. Ganapathy Iyer, Lecturer in Mathematics at Presidency College, Madras, Ramanujan remarks, "You must remember that the 'Orders of Infinity' which I found in your room introduced me to Messrs Hardy and Littlewood." In his third *Quarterly Report*, motivated by reading Hardy's book, Ramanujan briefly studied rates of growth of classes of functions [**1**, pp. 326–328].

Ramanujan apparently devoted more research efforts to the theory of elliptic functions than to any other area of mathematics. Although his development of the theory is decidedly different from that of any other author, Ramanujan must have learned about the subject from some text(s) in English. It has been suggested by Hardy and others that Greenhill's and Cayley's books on elliptic functions were the two mostly likely sources for Ramanujan's introduction to the subject. In a letter of 7 April 1936 from K. Ananda Rau to Hardy, Ananda Rau remarks, "I am told that there has been a copy of Greenhill in the Library of the Government College at Kumbakonam, but I have not been able to verify this [**4**, p. 289]." When one of us (BCB) visited the Government College at Kumbakonam's Library in 1984, Greenhill's book was found in the Library's stacks. A library catalogue published by the College in 1933 lists Greenhill's book on elliptic functions, but does not list Cayley's book on the subject. According to Hardy [**5**, p. 20], "The Librarian of the University of Madras has very kindly sent me a copy of the catalogue published in 1914, which makes it plain that the library was better equipped than I had supposed. For example, it possessed two standard French treatises on elliptic functions (Appell and Lacour, Tannery and Molk) as well as the books of Cayley and Greenhill. It seems plain from other evidence that Ramanujan knew something of the English books but nothing of the French ones." When one of us (BCB) visited the University of Madras Library in February, 1999, the present University Librarian, Dr. R. Vengan, had kindly made a list of the mathematics books held by the Library in 1914. Indeed, a large number of books on elliptic functions, mostly in French or German, could be found at the Library in 1914. In particular, the famous treatises on elliptic functions by F. Klein, A. Krazer, and H. Weber were there. Although Greenhill's book is not currently in the Library, records show that it indeed was in the Library prior to 1914. On the other hand, in contrast to Hardy's claim, although Cayley's book is currently in the library, the library's records do not show that the book was present in 1914; according to Dr. Vengan the current copy was not purchased until 1974.

Ramanujan's development of elliptic functions evolved without any use of the theory of functions of a complex variable. Since Greenhill's exposition also does not employ complex analysis, in particular, double periodicity is not mentioned until Chapter VIII, beginning on page 254, it would have been natural for Ramanujan to learn about elliptic functions from Greenhill's book. But there is even stronger evidence that Ramanujan was familiar with Greenhill's text. At three places in his notebooks [**7**], Ramanujan uses modular equations to study the periods of a pendulum [**2**, pp. 212–213, 243–246, 299–301]. This application of elliptic functions is emphasized in Greenhill's treatise; in fact, on page 1 he introduces the subject of elliptic functions by considering the motion of a pendulum. Ramanujan derived well over 100 modular equations, a subject barely touched by Greenhill. However, on pages 327 and 328, Greenhill lists, without proofs, several modular equations due to R. Russell from an article in volume 21 of the *Proceedings of the London Mathematical Society*. In this paper, Russell introduced certain functions, P, Q, and R, of moduli. In his notebooks [**7**], in particular, in Chapters 19 and 20 in the second notebook [**2**] and in the unorganized pages of his first notebook [**3**, pp. 385–391], Ramanujan recorded several modular equations in the notation of Russell. It is therefore most likely that Ramanujan had learned about modular equations of this sort from Greenhill's treatise.

Although Cayley's *Elliptic Functions* was possibly in the University of Madras Library in 1914, we have been unable to discern if this book was in the Library at the Government College at Kumbakonam. In the same letter by Ananda Rau [**4**, p. 289] quoted above, he affirms, "While at Cambridge I remember Ramanujan having with him a private copy of Cayley, and he used to tell me that he found it useful." This is not proof that Ramanujan had also used the book in India, but it would seem likely that he did since almost all of his work on elliptic functions was accomplished in India. Cayley's book is quite different from that of Greenhill. No applications to physics are given. Modular equations are featured, but the viewpoint is unlike that of Ramanujan. Cayley's treatise also contains much material on elliptic integrals, a topic extensively examined by Ramanujan.

Ramanujan likely learned about the theory of numbers from a book, and a likely text would have been G. B. Mathews's *The Theory of Numbers*. However, primarily because Ramanujan evidently had no knowledge of the Riemann zeta-function, its complex zeros, and its connections with the theory of prime numbers, topics thoroughly discussed in Chapter X of Mathews's book, Hardy felt that Ramanujan had not seen it. On the other hand, in a letter [**4**, p. 121] to C. N. Ganapathy Iyer, dated 17 December 1914, in describing some recent work of J. E. Littlewood, Ramanujan writes, "If you see the last paragraph in 'Mathew's [sic] Theory of Numbers' you will find the statement 'Hence we conclude that Gauss's approximation was too great.'" Thus, one of his two closest mathematical friends in Madras did have a copy of Mathews's book, and Ramanujan intimates that this source was well known to both of them. Because the theory of functions of a complex variable is prominently employed in Chapter X, and because Ramanujan evidently had not mastered this subject, Ramanujan probably had not thoroughly read this chapter, and so the lacunae in Ramanujan's knowledge, which were emphasized by Hardy, are understandable.

3. DESCRIPTIONS OF THE BOOKS AND BIOGRAPHIES OF THE AUTHORS

S. L. Loney, *Plane Trigonometry*

After graduating from London University with a Bachelor of Arts degree, Sidney Luxton Loney (1860–1939) entered Sidney Sussex College, Cambridge, to study mathematics. He was classed as third Wrangler in the Mathematical Tripos in 1882 and three years later was appointed to a Fellowship at his College, which he held from 1885 to 1891. In 1888 he was appointed as the first professor of mathematics at Royal Holloway College, London University, which had opened two years earlier at Egham outside London as a college for women students; it is now open to both sexes.

Loney was the author of numerous textbooks on a variety of mathematical subjects. His book on plane trigonometry was published by the Cambridge University Press, and passed through five editions between 1893 and 1925. As a practical man of affairs he was a valued business adviser to the College during its early years. He was a clear, patient, and inspiring teacher. An obituary published by the College notes that he had a strong distrust of innovation, and "to the eager progressive he seemed an inveterate reactionary. But much of his dislike of change sprang from a sentiment for tradition, and even those who differed from him in matters of policy recognized and admired his devotion to the College and its interests."

After his retirement he served for two years as Mayor of the neighboring town of Richmond.

G. S. Carr, *A Synopsis of Elementary Results in Pure Mathematics*

George Shoobridge Carr (1837–1914) entered Gonville and Caius College, Cambridge, in 1876 at the advanced age of 38, and was classed as a Senior Optime (i.e., in the Second Class) in the Mathematical Tripos in 1880. He earned his living in London as a private tutor in mathematics, i.e., as a coach.

The first part of Carr's *Synopsis*, comprising 256 pages and 1174 articles, appeared in 1880; the second larger part, comprising 935 pages and 3243 articles, appeared in 1886. The total number of entries is thus 4417. The number attached to the final entry is 6165, and generally Ramanujan's biographers have written that there are about 6000 entries in Carr's book. However, there are hundreds of gaps in the numbering, ranging from a gap of 1, of which there are numerous instances, to a gap of 170. The publishers were C. F. Hodgson and Son, who were for many years publishers to the London Mathematical Society. The format of the book undoubtedly influenced the way in which Ramanujan recorded his mathematical results in his notebooks. The different articles in the book, some of which are only one line in length, contain occasional hints as to their proofs, but these are generally of a very abbreviated kind.

One feature of the book, which has aroused little or no comment, is the fact that at the end of the second part 32 different mathematical journals are listed, and these are followed by 87 pages indexing different mathematical subjects and the places where they are mentioned in the preceding journals. Thus, pages 873–874 contain a long list of subjects under the heading Elliptic Functions (including modular functions!), although only articles 2125–2132 pertain to a part of that subject, namely elliptic integrals.

Mathematics was not Carr's only interest. In 1896 he published a 115 page booklet entitled *Social Evolution and the Evolution of Socialism: A Critical Essay*.

J. Edwards, *Differential Calculus*

Joseph Edwards (1854–1931) of Sidney Sussex College, Cambridge, was classed as 4th Wrangler in the Mathematical Tripos in 1878. He was then appointed to a Fellowship of his College, which he held for three years. He then moved to Queen's College, London, as Principal and Professor of Mathematics and Physics. Queen's College claims to be the first school in the United Kingdom established for the higher education of women. It gave instruction of a pre-university standard and still exists in Harley Street as an independent day school for girls between the ages of 11 and 18. Edwards's book on the differential calculus was published by Macmillan and ran through numerous editions between 1886 and 1954. He also wrote 2 volumes on integral calculus, published in 1922, also by Macmillan.

B. Williamson, *An Elementary Treatise on the Integral Calculus*

Benjamin Williamson (1827–1916) spent the greater part of his life at Trinity College, Dublin, as a student, Fellow, and Vice-Provost. He held the chair of Natural Philosophy from 1884 to 1890. He was the author of a number of books on mathematical subjects. His treatise on the integral calculus ran through eight editions between 1875 and 1920 and was used by one of the present authors (RAR) when a boy at school. Williamson had a reputation as a good teacher. His books on the differential and integral calculus were designed to impart facility in the practical use of the subject rather than develop its logical structure.

Unlike most Fellows of Trinity in the mid-nineteenth century, he was a layman. A bachelor, he was a genial socially-minded clubman, and an indefatigable diner out in professional, official, and political circles, where he made friends with such public figures as Lord Randolph Churchill.

He was elected a Fellow of the Royal Society in 1879, and his last book, on the mathematical theory of elasticity, appeared in 1894.

G. H. Hardy, *Orders of Infinity: The 'Infinitärcalcul' of Paul Du Bois–Reymond*

This book, published in 1910 as one of the Cambridge Tracts in Mathematics, is concerned with sequences (called scales) of functions, each tending to infinity faster than its predecessor, and, in particular, with the so-called logarithmico-exponential scales. A second edition appeared in 1924. As indicated in the title, the first mathematician to study the subject in detail was Du Bois–Reymond (1831–1889).

Biographies of Hardy (1877–1947) can be found in the authors' book [**4**, pp. 30–33] and in R. Kanigel's book [**6**]. Or, see http://www-history.mcs.st-and.ac.uk/~history/Mathematicians/Hardy.html.

A. G. Greenhill, *The Applications of Elliptic Functions*

Sir George Greenhill (1847–1927) of St. John's College, Cambridge, was second Wrangler in the Mathematical Tripos in 1870, and was appointed to a Fellowship of his College. He later transferred to Emmanuel College. His book on elliptic functions was first published in 1892, and a French translation appeared in 1895. It was reprinted by Dover in 1959. He became Professor of Mathematics at the

Artillery College, Woolwich, in 1876, and was knighted on his retirement from this post in 1908. He served as President of the London Mathematical Society from 1890 to 1892 and was the recipient of the Society's De Morgan Medal in 1902; he was also a Fellow of the Royal Society. His mathematical work was concerned with the applications of elliptic functions to problems in dynamics, hydrodynamics, elasticity, and electrostatics, but he also wrote on complex multiplication. He was an able organist and interested in antiques. For further information, see http://www-history.mcs.st-and.ac.uk/~history/Mathematicians/Greenhill.html.

A. Cayley, *An Elementary Treatise on Elliptic Functions*

Arthur Cayley (1821–1895) of Trinity College, Cambridge, was Senior Wrangler in the Mathematical Tripos in 1842. He afterwards qualified as a lawyer and was called to the bar at Lincoln's Inn in 1849. He confined his legal practice to conveyancing. Cayley held the Sadleirian Professorship at Cambridge from 1863 to 1895. He was a Fellow of the Royal Society and was awarded that society's Copley Medal in 1882. He was also the first recipient of the London Mathematical Society's De Morgan Medal in 1884. He was the author of more than 900 mathematical papers on a variety of subjects, but mainly on algebra and geometry. His treatise on elliptic functions was published in 1876 by University Press, Cambridge, with a second edition appearing in 1895 and a reprinting by Dover in 1961. Cayley was regarded as the greatest mathematician of his time by his countrymen. See also http://www-history.mcs.st-and.ac.uk/~*history/Mathematicians/Cayley.html.*

G. B. Mathews, *The Theory of Numbers*

George Ballard Mathews (1861–1922) entered St. John's College, Cambridge in 1879. At school he had been equally distinguished in Classics and Mathematics and, on coming to Cambridge, his College offered him a Senior Scholarship in either Classics or Mathematics, but he chose to study the latter. He was classed as Senior Wrangler in the Mathematical Tripos in 1883 and was elected to a Fellowship by his College. The following year he was appointed Professor of Mathematics at the University College of North Wales at Bangor.

He was well read not only in mathematics and classics but also in Hebrew and Arabic. He was elected Fellow of the Royal Society in 1897 and wrote a number of papers on geometric and algebraic subjects, including binary cubic forms and lemniscate functions. His book, *The Theory of Numbers*, was published in 1892 at his own expense by Deighton Bell, the Cambridge bookseller. It was intended to be Part I of a more comprehensive treatise, but no further parts appeared. A corrected reprint omitting the words Part I was published by the Chelsea Publishing Company during the 1960s. It is not an easy book to read and contains a chapter on the theory of prime numbers as it existed up to the book's publication, just four years before the first proof of the prime number theorem.

Mathews was the author or reviser of a number of other books on projective geometry, algebraic equations, and determinants. Together with his friend and obituarist Andrew Gray, he published *A Treatise on Bessel Functions and Their Applications in Physics*, which appeared in 1895 and was later translated into Russian.

He resigned his Bangor chair in 1896, being discouraged by the attitude of his students, and returned to a Lectureship in Cambridge. But he returned to Bangor in 1911 to a special Lectureship and resumed some of his professoral duties.

ACKNOWLEDGMENT. We are grateful to J. W. Anderson, T. G. Ramamurthy, K. Srinivasa Rao, Eira J. Scourfield, R. Vengan, and Trevor T. West for the help that they have kindly given.

REFERENCES

1. B. C. Berndt, *Ramanujan's Notebooks, Part I*, Springer-Verlag, New York, 1985.
2. B. C. Berndt, *Ramanujan's Notebooks, Part III*, Springer-Verlag, New York, 1991.
3. B. C. Berndt, *Ramanujan's Notebooks, Part V*, Springer-Verlag, New York, 1998.
4. B. C. Berndt and R. A. Rankin, *Ramanujan: Letters and Commentary*, American Mathematical Society, Providence, 1995; London Mathematical Society, London, 1995.
5. G. H. Hardy, *Ramanujan*, Cambridge University Press, Cambridge, 1940; reprinted by Chelsea, New York, 1960; reprinted by the American Mathematical Society, Providence, 1999.
6. R. Kanigel, *The Man Who Knew Infinity*, Scribner's, New York, 1991.
7. S. Ramanujan, *Notebooks* (2 *volumes*), Tata Institute of Fundamental Research, Bombay, 1957.
8. S. Ramanujan, *Collected Papers*, Cambridge University Press, Cambridge, 1927; reprinted by Chelsea, New York, 1962; reprinted by the American Mathematical Society, Providence, 2000.

BRUCE BERNDT received an A.B. from Albion College in 1961 and a Ph.D. from the University of Wisconsin in 1966 under the direction of J. R. Smart. Except for one year appointments at the University of Glasgow (1966–67) and Institute for Advanced Study (1973–74), he has been at the University of Illinois for his entire academic career. His interest in Ramanujan's work began in early 1974, and he has published five books with Springer-Verlag on Ramanujan's notebooks, for which he was awarded a Steele Prize in 1996. He is now working with George Andrews on similar volumes on Ramanujan's lost notebook.
University of Illinois, 1409 West Green St., Urbana, IL 61801
berndt@math.uiuc.edu

ROBERT A. RANKIN graduated B.A. from Clare College, Cambridge, England in 1937 and enrolled as a research student in the Theory of Numbers and Modular Functions under A. E. Ingham and, later, G. H. Hardy. He graduated Ph.D. in 1940 and spent the following five years of the war working on rockets in West Wales for the Ministry of Supply. After the war he returned to Cambridge as a Lecturer. In 1951 he moved to Birmingham University as Professor of Mathematics, and in 1954 he transferred to Glasgow University, from which he retired in 1982.
University of Glasgow, Glasgow G12 8QW, Scotland, UK
rar@maths.gla.ac.uk

Pinching, Trimming, Truncating, and Averaging of Matrices

Rajendra Bhatia

To M. G. Nadkarni on his sixtieth birthday.

1. INTRODUCTION. What happens to the norm of a matrix when some of its entries are replaced by zeros? This question leads to some interesting mathematics and the goal of this paper is to describe some of it.

The answer to our question depends on two things: the norm that we choose and the position of the entries that are replaced. Let us illustrate this by examples.

The *operator norm* of an $n \times n$ complex matrix A is its norm as a linear operator on the Euclidean space $\mathbb{C}^n$; i.e., $\|A\| = \sup\{\|Ax\| : x \in \mathbb{C}^n, \|x\| = 1\}$. The *Frobenius norm* of A is defined as $\|A\|_2 = (\Sigma_{i,j}|a_{ij}|^2)^{\frac{1}{2}} = (\operatorname{tr} A^*A)^{\frac{1}{2}}$. Both these norms are used frequently in analysis of matrices. They can also be described in terms of the singular values of A—the square roots of the eigenvalues of A^*A enumerated as $s_1(A) \geq \cdots \geq s_n(A)$. We have $\|A\| = s_1(A)$ and $\|A\|_2 = (\Sigma_j s_j^2(A))^{\frac{1}{2}}$.

Let $M(n)$ be the space of all $n \times n$ (complex) matrices. A norm $|||\cdot|||$ on $M(n)$ is said to be *unitarily invariant* if $|||UAV||| = |||A|||$ for all unitary matrices U,V and for all A in $M(n)$. Since $s_j(UAV) = s_j(A)$, the operator norm and the Frobenius norm are unitarily invariant. Another example of such a norm is the *trace norm* defined as $\|A\|_1 = \Sigma_j s_j(A)$. More examples, and properties, of these norms may be found in [**2**, Chapter IV].

Let

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

It is easy to see that $\|A\| = \sqrt{2}$, while $\|B\| = \frac{1}{2}(1 + \sqrt{5})$. Thus, replacing an entry of a matrix by zero can increase its operator norm. On the other hand, the Frobenius norm of a matrix is diminished if any of its entries is replaced by one with smaller absolute value. It is an interesting fact that among all unitarily invariant norms, the Frobenius norm is the only one that has this property; see [**4**, Prop. 3.1]. The matrix B was obtained from A by *triangular truncation*—wiping out the part below the main diagonal. We say more about this operation later on.

There is an interesting operation on $M(n)$ that reduces all unitarily invariant norms. Let $P_1, \ldots, P_k$ be orthogonal projection operators in $\mathbb{C}^n$ whose ranges are orthogonal to each other and whose sum $P_1 + \cdots + P_k = I$. Let

$$\mathscr{C}(A) = \sum_{j=1}^{k} P_j A P_j. \tag{1}$$

This is called a *pinching* of A. If we choose an orthonormal basis for $\mathbb{C}^n$ whose elements successively span the ranges of P_j, $1 \leq j \leq k$, then the matrix of A can be decomposed into blocks in such a way that the diagonal blocks are square and have sizes dim range P_j and the matrix $\mathscr{C}(A)$ is obtained from this by replacing all the off-diagonal blocks by zeros. Especially interesting is the case when the range of each P_j is 1-dimensional. In this case the pinching replaces all off-diagonal entries of A by zeros. The resulting matrix, the *diagonal part* of A, is written as $\mathscr{D}(A)$.

The sum in (1) is reminiscent of a convex combination, and indeed that is one reason for the interest in pinchings—they describe certain averaging operations on operators. There is another sense in which $\mathscr{C}(A)$ can be obtained from A by averaging. Let $\omega = e^{2\pi i/n}$ and let U be the diagonal matrix with entries $1, \omega, \omega^2, \dots, \omega^{n-1}$ down its diagonal. Using the identity $\sum_{j=0}^{n-1}\omega^j = 0$ and elementary algebra one can see that

$$\mathscr{D}(A) = \frac{1}{n}\sum_{j=0}^{n-1} U^j A U^{*j}. \tag{2}$$

The matrix U is unitary and the expression (2) represents $\mathscr{D}(A)$ as an average of n unitary conjugates of A. Using the same idea one can write any pinching $\mathscr{C}(A)$ corresponding to k orthogonal projections as an average of k unitary conjugates of A.

It follows from (2) that $|||\mathscr{D}(A)||| \le |||A|||$ for *every* unitarily invariant norm. This is a consequence of the triangle inequality and the fact that each of the summands in (2) has the same norm as A. More generally, we have $|||\mathscr{C}(A)||| \le |||A|||$ for every pinching $\mathscr{C}$. For the sake of brevity, we write $|||X||| \le |||Y|||$ to mean that for two given matrices X, Y, every unitarily invariant norm of X is bounded by the corresponding norm of Y.

This idea of representing diagonals as averages can be carried much further, as we will soon see. We will replace the sum in (2) by an integral and the roots of unity by trigonometric polynomials. Such representations lead to interesting bounds for norms of matrices obtained by trimming A in different ways.

2. DIAGONALS AS AVERAGES. In addition to the main diagonal of A we consider other diagonals as well. Let σ be any permutation of the indices $\{1,2,\dots,n\}$ and let $\mathscr{S}(A)$ be the matrix obtained from A by replacing all its entries except $a_{j\sigma(j)}$ by zeros. This is a *generalized diagonal* of A obtained by retaining exactly one entry from each row and each column. When σ is the identity permutation $\mathscr{S}(A) = \mathscr{D}(A)$. To each σ corresponds a permutation matrix P_σ that permutes the basis vectors. One can see that $\mathscr{S}(A)$ is just the main diagonal of the matrix AP_σ. Hence $|||\mathscr{S}(A)||| \le |||AP_\sigma|||$. So, by unitary invariance $|||\mathscr{S}(A)||| \le |||A|||$, another instance when norms are diminished when some entries are replaced by zeros.

Now for $1 \le j \le n-1$, let $\mathscr{D}_j(A)$ be the matrix obtained from A by replacing all its entries except those on the jth superdiagonal by zeros. Likewise, let $\mathscr{D}_{-j}(A)$ be the matrix obtained by retaining only the jth subdiagonal of A. (The superdiagonals are the diagonals above the main diagonal and parallel to it; the subdiagonals are the ones below the main diagonal.) To be consistent with this notation, put $\mathscr{D}_0(A) = \mathscr{D}(A)$.

How big are the norms of $\mathscr{D}_j(A)$? Note that for each $1 \le j \le n-1$, the sum $\mathscr{D}_j(A) + \mathscr{D}_{j-n}(A)$ is a generalized diagonal $\mathscr{S}(A)$. Hence $|||\mathscr{D}_j(A) + \mathscr{D}_{j-n}(A)||| \le |||A|||$. Once again, by a permutation we can bring all the nonzero entries of $\mathscr{D}_j(A) + \mathscr{D}_{j-n}(A)$ to the main diagonal. This does not change norms. The norm of a diagonal matrix is certainly reduced if any of its entries is replaced by a zero. (Hint: express the new matrix as an average of the original matrix and another one of equal norm.) Thus

$$|||\mathscr{D}_j(A)||| \leq |||A||| \quad \text{for all } j. \tag{3}$$

Using the triangle inequality, we obtain from this

$$|||\mathscr{D}_j(A) + \mathscr{D}_{-j}(A)||| \leq 2|||A||| \quad \text{for all } j. \tag{4}$$

By the same argument, if $\mathscr{T}_3(A) = D_{-1}(A) + \mathscr{D}_0(A) + \mathscr{D}_1(A)$ is the *tridiagonal part* of A, then

$$|||\mathscr{T}_3(A)||| \leq 3|||A|||. \tag{5}$$

A slightly cleverer argument gives a better inequality. We can write $\mathscr{T}_3(A) = \mathscr{C}(A) + \Gamma(A)$, where $\mathscr{C}(A)$ is a pinching of A by 2×2 blocks and $\Gamma(A)$ is a part of a generalized diagonal of A. For example, if we write the tridiagonal part of a 5×5 matrix A as

$$\begin{bmatrix} \bigcirc & \bigcirc & & & \\ \bigcirc & \bigcirc & \bigstar & & \\ & \bigstar & \bigcirc & \bigcirc & \\ & & \bigcirc & \bigcirc & \bigstar \\ & & & \bigstar & \bigcirc \end{bmatrix}$$

then all the entries represented by circles together constitute a pinching of A into blocks of sizes 2, 2, and 1. The remaining entries represented by stars constitute a part of a generalized diagonal of A. This shows that

$$|||\mathscr{T}_3(A)||| \leq 2|||A|||. \tag{6}$$

Can one improve this further? How about the inequality (4)?

For each real number θ, let U_θ be the diagonal matrix with entries $e^{ir\theta}$, $1 \leq r \leq n$, down its diagonal. Then the (r, s) entry of the matrix $U_\theta A U_\theta^*$ is $e^{i(r-s)\theta} a_{rs}$. Hence, we have

$$\mathscr{D}_k(A) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} U_\theta A U_\theta^* \, d\theta. \tag{7}$$

When $k = 0$, this gives another representation of $\mathscr{D}_0(A)$ as an average over unitary conjugates of A. For other values of k, this expresses $\mathscr{D}_k(A)$ as a "twisted average" over unitary conjugates of A. From this expression we can again derive the inequality (3). We can also use it to write

$$\mathscr{D}_k(A) + \mathscr{D}_{-k}(A) = \frac{1}{2\pi} \int_{-\pi}^{\pi} (2\cos k\theta) U_\theta A U_\theta^* \, d\theta.$$

Hence

$$|||\mathscr{D}_k(A) + \mathscr{D}_{-k}(A)||| \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |2\cos k\theta| \, d\theta |||A|||.$$

It is easy to evaluate this integral. One gets

$$|||\mathscr{D}_k(A) + \mathscr{D}_{-k}(A)||| \leq \frac{4}{\pi} |||A|||. \tag{8}$$

This is an improvement over the inequality (4). Using the same argument we see that

$$|||\mathscr{T}_3(A)||| \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |1 + 2\cos\theta| \, d\theta \, |||A|||.$$

Once again, it is easy to evaluate the integral. We now get

$$|||\mathscr{T}_3(A)||| \leq \left(\frac{1}{3} + \frac{2\sqrt{3}}{\pi} \right) |||A|||. \tag{9}$$

This is an improvement over (6). The constant factor in the inequality (9) is smaller than 1.436, that in (8) is smaller than 1.274.

More generally, consider the *trimming* of A obtained by replacing all its diagonals outside the band $-k \leq j \leq k$ by zeros; i.e., consider the matrices

$$\mathscr{T}_{2k+1}(A) = \sum_{j=-k}^{k} \mathscr{D}_j(A), \quad 1 \leq k \leq n. \tag{10}$$

Then, from (7) we get

$$\mathscr{T}_{2k+1}(A) = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_k(\theta) U_\theta A U^*_\theta \, d\theta, \tag{11}$$

where

$$D_k(\theta) = \sum_{j=-k}^{k} e^{ij\theta} \tag{12}$$

is the *Dirichlet kernel*, a familiar object related to Fourier series. See [**3**, Sec. 2.2] or [**9**, p. 174]. The numbers

$$L_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} |D_k(\theta)| \, d\theta \tag{13}$$

are called the *Lebesgue constants*. It is known that for large values of k, L_k are like $\log k$. For example, one knows that

$$L_k \leq \log k + \log \pi + \frac{2}{\pi}\left(1 + \frac{1}{2k}\right),$$

and that

$$L_k = \frac{4}{\pi^2}\log k + O(1).$$

From (11) we see that

$$|||\mathscr{T}_{2k+1}(A)||| \leq L_k |||A|||. \tag{14}$$

Naive arguments would have shown only that $|||\mathscr{T}_{2k+1}(A)||| \leq (2k+1)|||A|||$. The inequality (14) is a striking improvement, when n, k, and $n-k$ are large.

The trimming operation we have introduced here has an interesting connection with the triangular truncation that we talked of in the Introduction. Let Δ_U be the linear map on the space of matrices (of a fixed size) that takes a matrix B to its upper triangular part; i.e., Δ_U acts by replacing all entries of a matrix below the main diagonal by zeros. Given a $k \times k$ matrix B, consider the matrix $A = \begin{bmatrix} 0 & B^* \\ B & 0 \end{bmatrix}$. The singular values of A are the singular values of B counted twice as often. Hence, $\|A\| = \|B\|$. Note also that

$$\mathscr{T}_{2(k+1)+1}(A) = \begin{bmatrix} 0 & \Delta_U(B)^* \\ \Delta_U(B) & 0 \end{bmatrix}.$$

So, it follows from (14) that

$$\|\Delta_U(B)\| \leq L_{k+1}\|B\|. \tag{15}$$

In other words, the norm of the triangular truncation operator (on the space of $k \times k$ matrices equipped with the operator norm) is bounded by the Lebesgue constant L_{k+1}. We have remarked earlier that $L_{k+1} \approx 4\pi^{-2}\log k$.

It is remarkable that arguments from Fourier series lead to interesting bounds (8), (9), (14), and (15) for norms of matrices. The unexpected appearance of the number π makes them especially attractive. Of course, this appeal would be lost if better bounds were to be found. As it turns out, the bounds (8) and (9) are sharp, as are the bounds (14) and (15) in an asymptotic sense. This is discussed in the next section.

3. EXAMPLES. We show that the bounds (8) and (9) are sharp for the trace norm and, therefore, by a duality argument they are sharp for the operator norm.

Example 3.1. Let B be the tridiagonal $n \times n$ matrix with each entry on its first superdiagonal and the first subdiagonal equal to 1, and all other entries equal to 0. It is not difficult to see that the eigenvalues of B are $2\cos(j\pi/n+1)$, $1 \leq j \leq n$; see [**2**, p. 60]. The matrix $B - 2I$ is the familiar second difference matrix that is used in numerical analysis to discretize the second derivative operator.

Example 3.2. Let $A = E$, the matrix with all entries equal to 1. Then

$$\mathscr{D}_1(A) + \mathscr{D}_{-1}(A) = B,$$

where B is the tridiagonal matrix in Example 3.1. Here

$$\frac{\|B\|_1}{\|A\|_1} = \frac{1}{n}\sum_{j=1}^{n}\left|2\cos\frac{j\pi}{n+1}\right|. \tag{16}$$

Let $f(\theta) = |2\cos\theta|$. The sum

$$\frac{1}{n+1}\sum_{j=1}^{n+1}\left|2\cos\frac{j\pi}{n+1}\right|$$

is a Riemann sum for the function $\pi^{-1}f(\theta)$ over a subdivision of the interval $[0, \pi]$ into $n+1$ equal parts. As $n \to \infty$, this sum and the one in (16) tend to the same limit. This limit is equal to

$$\frac{1}{\pi}\int_0^{\pi}|2\cos\theta|\,d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi}|2\cos\theta|\,d\theta = \frac{4}{\pi}.$$

This shows that the inequality (8) can not be improved if it has to be valid for all dimensions n and for all unitarily invariant norms.

The same example shows that the inequality (9) is also sharp.

Note that in this example A was Hermitian, so the inequalities (8) and (9) are sharp even on the space of all Hermitian matrices.

The duality principle we need says that if T is a linear operator from a Banach space X to another Banach space Y, then its adjoint T^* (a linear map from the dual Y^* to X^*) has the same norm as T.

Let X be the space $M(n)$, or the space $H(n)$ of all Hermitian $n \times n$ matrices. The space X has a natural inner product defined as $\langle A, B\rangle = \operatorname{tr} A^*B$. By the Riesz Representation Theorem every linear functional φ on X is of the form $\varphi(A) = \operatorname{tr} A\Phi$, where Φ is some element of X. The norms $\|\cdot\|$ and $\|\cdot\|_1$ on X are dual to each other; i.e., if X is equipped with the operator norm $\|\cdot\|$, then its dual space X^* is the space X equipped with the trace norm $\|\cdot\|_1$, and vice versa. The adjoint of a linear operator $\mathscr{L}: X \to X$, is the linear operator $\mathscr{L}^*: X \to X$ that satisfies the relation

$$\langle A, \mathscr{L}(B)\rangle = \langle \mathscr{L}^*(A), B\rangle \quad \text{for all } A, B.$$

It is easy to verify that for each k the operator taking a matrix A to $\mathscr{D}_k(A) + \mathscr{D}_{-k}(A)$ is its own adjoint. Hence, its norm as a linear operator on the space $(X, \|\cdot\|)$ is the same as its norm as a linear operator on the space $(X, \|\cdot\|_1)$. Thus the inequalities (8) and (9) are sharp for the operator norm as well. Further, they are sharp even when A is Hermitian.

Example 3.3. Let A be the $n \times n$ matrix with entries $a_{ij} = (i-j)^{-1}$ if $i \neq j$, and $a_{ii} = 0$. With some work [**6**, p. 39] it can be seen that $\|A\| \leq \pi$ and $\|\Delta_U(A)\| \geq \frac{4}{5}\log n$ for large values of n. See the delightful article [**5**] for several examples related to this. This example shows that the norm of the operator Δ_U on $(M(n), \|\cdot\|)$ or $(H(n), \|\cdot\|)$ grows like $\log n$. More elaborate analysis shows that the norm of Δ_U approaches $\pi^{-1}\log n$ as n increases; see [**1**]. Our inequality (15) gives just a little larger number $4\pi^{-2}\log n$ as an asymptotic bound for this norm.

Example 3.4. Let B be a Hermitian matrix of a large order k for which $\|\Delta_U(B)\|$ is close to $\pi^{-1}\log k$; such a matrix exists, as we have remarked in our discussion of Example 3.3. Let $A = \begin{bmatrix} 0 & B \\ B & 0 \end{bmatrix}$. Then $\|A\| = \|B\|$.

Note that

$$\mathscr{T}_{2k+3}(A) = \begin{bmatrix} 0 & \Delta_U(B)^* \\ \Delta_U(B) & 0 \end{bmatrix}$$

and the norm of this matrix is $\|\Delta_U(B)\|$. So, the ratio of $\|\mathscr{T}_{2k+3}(A)\|$ and $\|A\|$ is approximately $\pi^{-1}\log k$, again showing that the bound (14) is almost exact for the operator norm (and by duality for the trace norm).

4. MORE ON AVERAGES. For I, a subset of $\{1, 2, \dots, n\}$, let X_I be the diagonal matrix whose diagonal entry x_{ii} is 1 if $i \in I$, and -1 if $i \notin I$. It can be seen easily that

$$\mathscr{D}(A) = \frac{1}{2^n}\sum_I X_I A X_I. \tag{17}$$

This expression has the advantage of using real diagonal matrices X_I instead of the complex matrices U^k used in (2). Further, these matrices have only ± 1 on the diagonal. So, an analogous expression can be used for matrices over other fields. On the other hand, now there are far more terms involved. This difference is crucial. In [**4**] it was shown that (2) leads to the bound

$$|||A - \mathscr{D}(A)||| \leq 2\left(1 - \frac{1}{n}\right)|||A|||$$

for the off-diagonal part of A, and that this inequality is sharp for the norms $\|\cdot\|$ and $\|\cdot\|_1$. Using (17) one would obtain a weaker inequality here.

Is it possible to obtain a representation for $\mathscr{D}(A)$ using real diagonal conjugations as in (17), but with fewer terms? It has been shown [**4**] that we could not have fewer than n terms in any case.

This question has an amusing connection with a famous problem in the theory of design of experiments. A matrix all of whose entries are $\pm$ 1, and whose columns are mutually orthogonal, is called a *Hadamard matrix*. Do such matrices exist? The 2×2 matrix $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is a Hadamard matrix. Taking m-fold tensor products of this matrix with itself gives Hadamard matrices of order $n = 2^m, m = 1, 2 \dots$. It is not difficult to see that for $n > 2$, a necessary condition for the existence of a Hadamard matrix of order n is that $n = 4k$ for some k. It has been conjectured that this condition is sufficient as well. This conjecture has been proved for $k \leq 106$; see [**10**].

Suppose n is such that a Hadamard matrix of order n exists. Let Y_j, $1 \leq j \leq n$, be the diagonal matrix whose diagonal is the jth column of this Hadamard matrix. Then,

$$\mathscr{D}(A) = \frac{1}{n} \sum_{j=1}^{n} Y_j A Y_j.$$

So, for such values of n, one does have a representation of $\mathscr{D}(A)$ as an average of n real diagonal conjugates of A.

If the conjecture on Hadamard matrices were to have a positive solution, then for all n we could find such a representation for $\mathscr{D}(A)$ with at most $n + 3$ terms.

Although we have concentrated on norms in this article, there is a long tradition in matrix theory of comparing eigenvalues, determinants, and singular values of a matrix to those of its diagonal. Several famous results due to Schur, Hadamard, Mirsky, Fan, Thompson, and others belong to this tradition. The interested reader could find them in the books [**2**], [**7**], and [**8**].

ACKNOWLEDGMENT. I am thankful to T. Ando, R. B. Bapat, M. D. Choi, C. Davis, A. Dey, L. Elsner, J. A. R. Holbrook, and E. Nordgren for illuminating conversations on topics discussed in this paper.

REFERENCES

1. J. R. Angelos, C. C. Cowen, and S. K. Narayan, Triangular truncation and finding the norm of a Hadamard multiplier, *Linear Algebra Appl.* **170** (1992) 117–135.
2. R. Bhatia, *Matrix Analysis*, Springer, New York, 1997.
3. R. Bhatia, *Fourier Series*, Hindustan Book Agency, New Delhi, 1993.
4. R. Bhatia, C. Davis, and M.-D. Choi, Comparing a matrix to its off-diagonal part, *Operator Theory: Advances and Applications* **40** (1989) 151–164.
5. M.-D. Choi, Tricks or treats with the Hilbert matrix, *Amer. Math. Monthly* **90** (1983) 301–312.
6. K. R. Davidson, *Nest Algebras*, Longman, Harlow, 1988.
7. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
8. A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press 1979.
9. W. Rudin, *Principles of Mathematical Analysis*, 2nd ed., McGraw-Hill, New York, 1964.
10. K. Sawade, A Hadamard matrix of order 268, *Graphs Combin.* 1 (1985) 185–187.

RAJENDRA BHATIA has been at the Indian Statistical Institute for several years. His research interests are matrices and linear operators, especially those problems where features of both are present, and where ideas from real and complex analysis play an important role. He admires Monthly articles such as the one cited in this paper, and enjoys working with his several collaborators around the world. His latest visiting appointment was as a Winegard Visiting Professor at the University of Guelph in Canada, where results in this article were first presented.
Indian Statistical Institute, 7, S.J.S. Sansanwal Marg, New Delhi 110016, India
rbh@isid.ac.in

When Is a Family of Sets a Family of Bounded Sets?

Gerald Beer

Let $\langle X, \tau \rangle$ be a metrizable topological space. A metric ρ for X is called *admissible* provided it is compatible with the topology τ for X. Given any admissible metric ρ for the topology, the equivalent metric d defined by $d(x, y) = \min\{1, \rho(x, y)\}$ makes each subset of X d-bounded. Now let $\mathscr{A}$ be a family of subsets of X. In this article we address the following question: when does there exist an unbounded metric d compatible with τ for which each member of $\mathscr{A}$ is a d-bounded set? Since points are always bounded with respect to any admissible metric, and bounded sets are closed under finite unions and are hereditary under inclusion, there is no loss in generality in assuming that (i) $\bigcup \mathscr{A} = X$; (ii) $\{A_1, A_2, A_3, \ldots, A_n\} \subset \mathscr{A} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathscr{A}$; (iii) $A \in \mathscr{A}$ and $B \subset A \Rightarrow B \in \mathscr{A}$. A family of sets satisfying these properties is often called a *bornology* **[1]**. Thus, our question may be recast as follows: when can a bornology in a metrizable space be extended to a nontrivial *metric bornology*. i.e., a bornology consisting of the bounded subsets of X as determined by some unbounded admissible metric?

We pause to introduce some notation and terminology. Let $\langle X, \tau \rangle$ be a metrizable space. We denote the interior, closure, and complement of a subset A of X by int A, cl A, and A^c, respectively. Given an admissible metric d for τ, $x \in X$, and $r > 0$ we denote the open (respectively, closed) ball of radius r and center x by $S_r^d(x)$ (respectively, $B_r^d(x)$). A subset A of X is d-bounded if for some $x \in X$ and $r > 0$, $A \subset S_r^d(x)$; equivalently, $\operatorname{diam}_d(A) = \sup\{d(a_1, a_2) : a_1 \in A, a_2 \in A\} < \infty$. We denote the family of d-bounded sets by $\mathscr{B}(d)$, so that a bornology on X is a metric bornology if it coincides with $\mathscr{B}(d)$ for some admissible metric d. Finally, $d(x, A)$ denotes $\inf\{d(x, a) : a \in A\}$, the usual distance from a point x in X to a subset A of X.

The question we posed at the outset has a rather simple answer. Recall that a family $\mathscr{U}$ of subsets of X is called *discrete* if each point $x \in X$ has a neighborhood V that meets at most one element of $\mathscr{U}$.

Theorem A. *Let $\langle X, \tau \rangle$ be a metrizable topological space, and let $\mathscr{A}$ be a family of nonempty subsets of X. The following conditions are equivalent:*

(i) *there exists an unbounded admissible metric d such that $\mathscr{A} \subset \mathscr{B}(d)$;*
(ii) *there exists a discrete countably infinite family $\{V_n : n \in Z^+\}$ of nonempty open sets in X such that for each $A \in \mathscr{A}$, $A \cap V_n = \varnothing$ for all but finitely many n.*

Proof: (i) $\Rightarrow$ (ii). Given an unbounded metric d for which $\mathscr{A} \subset \mathscr{B}(d)$, choose $x_0 \in X$ and inductively pick $x_n \in X$ such that for each n, $d(x_n, x_0) + 3 < d(x_{n+1}, x_0)$. Then $\{S_1^d(x_n) : n \in Z^+\}$ is the desired discrete family of open sets.

(ii) $\Rightarrow$ (i). Let ρ be a bounded metric for X. For each $n \in Z^+$ choose $x_n \in V_n$ and then $r_n > 0$ such that $B_{r_n}^{\rho}(x_n) \subset V_n$. We define a function $f : X \to [0, \infty)$ in a piecewise manner. On each closed ball $B^{\rho}_{r_n}(x_n)$, define

$$f(x) = n - \frac{n}{r_n} d(x, x_n).$$

Note that f maps the ball into $[0, n]$, $f(x_n) = n$ and if $d(x, x_n) = r_n$, then $f(x) = 0$.

Finally, on the complement of $\bigcup_{n=1}^{\infty} S^{\rho}_{r_n}(x_n)$ let $f(x) = 0$. Continuity of f at a point $w \in X$ is at issue only if $d(w, x_n) = r_n$ for some n, and in this case, we have

$$f(x) = \max\left\{0, n - \frac{n}{r_n} d(x, x_n)\right\}$$

in some neighborhood of w.

We are now ready to produce the desired metric d: $X \times X \to [0, \infty)$:

$$d(x, y) = \rho(x, y) + |f(x) - f(y)|.$$

It is easy to check that such a metric is equivalent to ρ and is thus compatible with the topology. Furthermore, since $d(x_n, x_k) \geq |n - k|$, the metric d is unbounded. If $A \in \mathscr{A}$ there exists $k \in Z^+$ such that $A \cap B^{\rho}_{r_n}(x_n) = \varnothing$ for all $n > k$, and as a result, for all $a_1, a_2 \in A$, we have

$$d(a_1, a_2) = \rho(a_1, a_2) + |f(a_1) - f(a_2)| \leq \operatorname{diam}_{\rho}(X) + k < \infty. \qquad \blacksquare$$

The proof of Theorem A reveals that for $\mathscr{A}$ to be a family of bounded sets, it is sufficient to produce an unbounded continuous real-valued function f on the space whose restriction to each A in $\mathscr{A}$ is bounded. This condition is also necessary, for if d is an unbounded admissible metric such that $\mathscr{A} \subset \mathscr{B}(d)$, then $f(x) = d(x, x_0)$ is an unbounded continuous function that is bounded on each A in $\mathscr{A}$.

What if $\mathscr{A}$ consists of a single subset of X? We have the following result.

Corollary 1. *Let $\langle X, \tau \rangle$ be a metrizable topological space, and let A be a nonempty subset of X. The following conditions are equivalent:*

(i) *there exists an unbounded admissible metric d such that $A \in \mathscr{B}(d)$;*
(ii) *there exists a sequence of distinct points $\langle x_n \rangle$ in X with no cluster point such that x_n lies in* $(\text{cl } A)^c$ *eventually;*
(iii) $(\text{cl } A)^c$ *does not have compact closure.*

Proof: (i) $\Rightarrow$ (ii). Let $\{V_n : n \in Z^+\}$ be a discrete countably infinite family of open sets such that for all n sufficiently large, $A \cap V_n = \varnothing$. Choosing $x_n \in V_n$ gives the desired sequence.

(ii) $\Rightarrow$ (iii). This is obvious.

(iii) $\Rightarrow$ (i). Since points in the closure of a set can be approximated by points in the set itself, by (iii) there is actually a sequence $\langle x_n \rangle$ with distinct terms in the open set $(\text{cl } A)^c$ with no cluster point in X. Let ρ be an admissible metric and for each $n \in Z^+$, define

$$r_n = \min\left\{\frac{1}{n}, \rho(x_n, A), \frac{1}{3}\inf\{\rho(x_n, x_i) : i \neq n\}\right\},$$

The family of closed balls $\{B^{\rho}_{r_n}(x_n) : n \in Z^+\}$ is clearly pairwise disjoint. As a result, discreteness of the family fails if and only if for some $w \in X$, each neighborhood of w hits infinitely many of these closed balls. But since $\lim_{n \to \infty} r_n = 0$, this would force w to be a cluster point of $\langle x_n \rangle$, which is impossible. As a result, $\{B_{r_n}^{\rho}(x_n) : n \in Z^+\}$ and therefore $\{S_{r_n}^{\rho}(x_n) : n \in Z^+\}$ are discrete families. By construction, for each index n, $r_n \leq \rho(x_n, A)$, and so $A \cap S_{r_n}^{\rho}(x_n) = \emptyset$. Apply Theorem A. ■

Theorem B. *Let $\langle X, \tau \rangle$ be a metrizable topological space, and let $\mathscr{A}$ be a family of nonempty subsets of X. The following conditions are equivalent:*

(i) *there exists an unbounded admissible metric d with respect to which $\mathscr{A}$ is uniformly bounded;*
(ii) $(\mathrm{cl} \bigcup \mathscr{A})^c$ *does not have compact closure.*

Proof: Uniform boundedness of $\mathscr{A}$ with respect to a metric d is equivalent to boundedness of $\bigcup \mathscr{A}$. Apply Corollary 1. ■

If $\langle X, \tau \rangle$ is Euclidean n-dimensional space equipped with the usual topology, a family $\mathscr{A}$ of subsets of X is uniformly bounded in some unbounded admissible metric precisely when the complement of the closure of the union of the family is unbounded with respect to usual Euclidean metric.

Corollary 1 following Theorem A of course invites the question: is a family $\mathscr{A}$ of subsets of X a family of bounded sets for some admissible metric if there exists a sequence $\langle x_n \rangle$ without a cluster point, whose terms are eventually outside the closure of each member A of the family? Given the existence of such a sequence, we can find for each $A \in \mathscr{A}$ neighborhoods of the points forming a discrete family of open sets that eventually miss A. The difficulty, then, would seem to be one of producing neighborhoods in a uniform way for the members of $\mathscr{A}$. The following example shows that this difficulty may indeed be insurmountable.

Example. In the plane with the usual topology, for $n \in Z^+$ let $E_n = \{(n, 1/k) : k \in Z^+\}$ and then let $C_n = \mathrm{cl}\, E_n$. For our metrizable space X, we take $X = \bigcup_{n=1}^{\infty} C_n$ as a subspace of the plane. Now consider the following family of subsets of X:

$$\mathscr{A} = \{A \subset X : \exists j \in Z^+ \quad \text{such that } \forall n \geq j\ (n, 0) \notin \mathrm{cl}\, A\}.$$

Note that each C_n belongs to the bornology $\mathscr{A}$, as does any X-closed subset of $\bigcup_{n=1}^{\infty} E_n$. Evidently, the sequence $\langle (\mathrm{n}, 0) \rangle$ is eventually outside $\mathrm{cl}\, A$ for each $A \in \mathscr{A}$. We claim, nevertheless, that there is no unbounded admissible metric d for X such that each $A \in \mathscr{A}$ is d-bounded.

Suppose to the contrary that there exists a metric d such that $\mathscr{A} \subset \mathscr{B}(d)$. Fix $x_0 \in X$ and let $V_n = (B_n^d(x_0))^c$ for $n = 1, 2, 3, \ldots$. Now for each n, V_n must hit infinitely many C_k or else V_n would be contained in a union of finitely many C_k and thus would be d-bounded. Starting with V_1, let $k_1 \in Z^+$ be the smallest positive integer with $V_1 \cap C_{k_1} \neq \emptyset$. Since V_1 is open and $C_{k_1} = \mathrm{cl}\, E_{k_1}$, we may choose a point $e_1 \in V_1 \cap E_{k_1}$. Next, let $k_2 > k_1$ be the smallest integer with $V_2 \cap C_{k_2} \neq \emptyset$, and then choose $e_2 \in V_2 \cap E_{k_2}$. Continuing, we produce a strictly increasing sequence of positive integers $\langle k_i \rangle$ such that $e_i \in V_i \cap E_{k_i}$ for each $i \in Z^+$. By the definition of $\mathscr{A}$ we have $\{e_i : i \in Z^+\} \in \mathscr{A}$, yet for each integer i we have $d(e_i, x_0) > i$. This gives a contradiction to the d-boundedness of each set in $\mathscr{A}$. ■

We can, however, give this positive result.

Theorem C. *Let $\langle X, \tau \rangle$ be a metrizable topological space, and let $\mathscr{A}$ be a family of nonempty subsets of X having a countable subfamily $\{F_n : n \in Z^+\}$ cofinal in $\mathscr{A}$ with respect to set inclusion. The following conditions are equivalent*:

(i) *there exists an unbounded admissible metric d such that $\mathscr{A} \subset \mathscr{B}(d)$;*
(ii) *there exists a sequence of distinct points $\langle x_k \rangle$ in X with no cluster point such that for each $A \in \mathscr{A}$, x_k lies in $(\text{cl } A)^c$ eventually.*

Proof: (i) $\Rightarrow$ (ii). Given such a metric d, fix $x_0 \in X$ and choose for each $k \in Z^+$ a point $x_k \in X$ with $d(x_0, x_{k+1}) > d(x_0, x_k) + 1$. This sequence $\langle x_k \rangle$ satisfies (ii).

(ii) $\Rightarrow$ (i). Without loss of generality, we may assume that the closure of each member of $\mathscr{A}$ is in $\mathscr{A}$ and that $\mathscr{A}$ is closed under finite unions, and as a result, we may assume that $\langle F_n \rangle$ is an increasing sequence of closed sets. Starting with $n = 1$, there exists $j_1 \in Z^+$ such that $x_k \notin F_1$ for all $k \geq j_1$. Arguing as in the proof of the implication (iii) $\Rightarrow$ (i) in Corollary 1, let $\{V_k(1) : k \geq j_1\}$ be a discrete family of open sets such that for all $k \geq j_1$ we have

(a) $x_k \in V_k(1)$;
(b) $F_1 \cap V_k(1) = \varnothing$.

Next choose $j_2 > j_1$ such that $x_k \notin F_2$ for all $k \geq j_2$. Then choose a second discrete family $\{V_k(2) : k \geq j_2\}$ of open sets such that for all $k \geq j_2$ we have

(a) $x_k \in V_k(2)$;
(b) $F_2 \cap V_k(2) = \varnothing$;
(c) $V_k(2) \subset V_k(1)$.

Continuing in this manner we get a strictly increasing sequence $\langle j_n \rangle$ of positive integers and for each n a discrete family $\{V_k(n) : k \geq j_n\}$ of open sets such that for each $n > 1$ and for each $k \geq j_n$ we have

(a) $x_k \in V_k(n)$;
(b) $F_n \cap V_k(n) = \varnothing$;
(c) $V_k(n) \subset V_k(n-1)$.

The family of open sets $\{V_{j_n}(n) : n \in Z^+\}$ fulfills condition (ii) of Theorem A, for if $A \in \mathscr{A}$ there exists some F_k containing A, and so for all $n \geq k$

$$A \cap V_{j_n}(n) \subset F_k \cap V_{j_n}(k) = \varnothing. \qquad \blacksquare$$

Example. In the plane with the usual topology, for each ordered pair (a, b) of reals, let $A(a, b) = \{(x, y) : x \leq a \text{ and } y \leq b\}$. The family $\mathscr{A} = \{A(a, b) : a \in \mathbb{R}$ and $b \in \mathbb{R}\}$ has a countable cofinal subfamily, namely $\{A(n, n) : n \in Z^+\}$. Thus, there exists an unbounded admissible metric with respect to which each such set is bounded, provided we can find a sequence in the plane that escapes each set eventually. Evidently, $\langle (n, n) \rangle$ is such a sequence. The reader can check that the following metric does the job:

$$d((x_1, y_1), (x_2, y_2)) = \min\{1, \max\{|x_1 - x_2|, |y_1 - y_2|\}\} + |\max\{x_1, y_1, 0\} - \max\{x_2, y_2, 0\}|.$$

Given an admissible unbounded metric d for $\langle X, \tau \rangle$, $\{B_n^d(x_0) : n \in Z^+\}$ is a countable subfamily in $\mathscr{B}(d)$ that is cofinal with respect to set inclusion, where x_0

is an arbitrary but fixed point of X. The bornology $\mathscr{B}(d)$ also has these additional properties: $X \notin \mathscr{B}(d)$ and for each $B \in \mathscr{B}(d)$ there exists $B' \in \mathscr{B}(d)$ with $\text{cl}\, B \subset \text{int}\, B'$. These properties turn out to be a characteristic set of properties for a bornology $\mathscr{A}$ to be a nontrivial metric bornology, as was first observed by S.-T. Hu; see [**2**] or [**3**. p. 191]. That is, a bornology $\mathscr{A}$ for a metrizable space coincides with $\mathscr{B}(d)$ for some unbounded admissible metric d if and only if

(a) $\mathscr{A}$ contains a countable subfamily cofinal with respect to set inclusion;
(b) $X \notin \mathscr{A}$;
(c) for each $A \in \mathscr{A}$ there exists $A' \in \mathscr{A}$ with $\text{cl}\, A \subset \text{int}\, A'$.

Condition (c) cannot be replaced by the weaker condition $A \in \mathscr{A} \Rightarrow \text{cl}\, A \in \mathscr{A}$, for the family $\mathscr{A}$ of finite subsets of the rationals equipped with the relative topology fails to be a metric bornology. The counterexample preceding Theorem C shows that condition (a) also does not follow from the remaining ones.

Hu's result, which ought to be standard in topology texts but is not, has a straightforward proof, which we outline. First, it is not hard to construct a cofinal sequence of closed sets $\langle A_n \rangle$ in $\mathscr{A}$ such that for each $n \in Z^+ A_n \subset \text{int}\, A_{n+1}$. Using Urysohn's Lemma, construct for each n a continuous function f_n into the unit interval that maps A_n to zero and $(\text{int}\, A_{n+1})^c$ to one, and let $f = f_1 + f_2 + f_3 + \cdots$. Then f is finite-valued and continuous and $f|A$ is bounded if and only if A is a subset of some A_n. Starting with a bounded admissible metric ρ for X, it is clear that the metric d defined by $d(x, y) = \rho(x, y) + |f(x) - f(y)|$ satisfies $\mathscr{A} = \mathscr{B}(d)$.

As an application of Hu's Theorem, consider a noncompact locally compact separable metrizable space X, and let $\mathscr{A}$ denote the bornology of relatively compact subsets of X, i.e., the family of subsets of X that have compact closure. Then $X \notin \mathscr{A}$, and by local compactness alone, the closure of each $A \in \mathscr{A}$ has a compact neighborhood. For the remaining condition, let ρ be an admissible metric for X, and let $\{x_n : n \in Z^+\}$ be a countable dense subset of X. For each n, let $T_n = \{r : r \text{ is a positive rational and } B_r^\rho(x_n) \text{ is compact}\}$. We have the following representation:

$$\mathscr{A} = \left\{ A : \exists\, k \in Z^+ \text{ such that } A \subset \bigcup_{n=1}^{k} B_{r_n}^{\rho}(x_n), \text{ where } r_n \in T_n \text{ for } n \le k \right\}.$$

It is now clear that $\mathscr{A}$ has a countable cofinal subset and is thus a metric bornology, which means that there is an admissible metric d for X whose closed balls are all compact, a fact first observed by H. Vaughn [**4**]. Conversely, local compactness and separability of the space are forced by the compactness of closed balls (a sigma compact metrizable space is separable [**5**, p. 126]).

We indicate a proof of the nontrivial part of Theorem A that relies on Hu's theorem. Let $\{V_n : n \in Z^+\}$ be a discrete countably infinite family of nonempty open subsets of X such that for each $A \in \mathscr{A}$, $A \cap V_n = \varnothing$ for all but finitely many n. Let ρ be an admissible metric and choose $x_n \in V_n$ and then $r_n > 0$ such that for each n, $B_{r_n}^\rho(x_n) \subset V_n$. For each $k \in Z^+$ let F_k be the following closed set:

$$F_k = \bigcup_{n=k}^{\infty} B_{k^{-1} r_n}^{\rho}(x_n).$$

Evidently, $\text{int}\, F_k \supset F_{k+1}$ and $\bigcap_{k=1}^{\infty} F_k = \varnothing$. One can now check that

$$\mathscr{B} = \{ B \subset X : \exists\, k \in Z^+ \text{ such that } B \cap F_k = \varnothing \}$$

is in fact a metric bornology (a countable cofinal subset within $\mathscr{B}$ is $\{F_n^c : n \in Z^+\}$). Finally, if $A \in \mathscr{A}$ there exists $k \in Z^+$ such that $A \cap (\bigcup_{n=k}^{\infty} V_n) = \varnothing$ and so $A \cap F_k = \varnothing$. This shows that $\mathscr{A} \subset \mathscr{B}$, as required.

Hu's result leads to a second completely transparent answer to our initial question.

Theorem D. *Let $\langle X, \tau \rangle$ be a metrizable topological space, and let $\mathscr{A}$ be a family of nonempty subsets of X. The following conditions are equivalent:*

(i) *there exists an unbounded admissible metric d such that $\mathscr{A} \subset \mathscr{B}(d)$;*
(ii) *there exists a cover $\{E_n : n \in Z^+\}$ of X by proper closed subsets such that $E_n \subset \operatorname{int} E_{n+1}$ and such that each $A \in \mathscr{A}$ is contained in some E_n.*

Proof: (i) ⇒ (ii). Fix $x_0 \in X$ and for each E_n take $B_n^d(x_0)$.
(ii) ⇒ (i). Given such a closed cover, let $\mathscr{B}$ be the following bornology:

$$\mathscr{B} = \{A \subset X : \exists\, n \in Z^+ \text{ such that } A \subset E_n\}.$$

Then $\mathscr{B}$ satisfies Hu's axioms for a metric bornology and contains $\mathscr{A}$. ∎

REFERENCES

1. H. Hogbe-Nlend, *Bornologies and Functional Analysis*, Mathematical Studies 26, North-Holland, Amsterdam, 1977.
2. S.-T. Hu, Boundedness in a topological space, *J. Math. Pures Appl.* 28 (1949) 287–320.
3. ______, *Introduction to General Topology*, Holden-Day, San Francisco, 1966.
4. H. Vaughn, On locally compact metrizable spaces, *Bull. Amer. Math. Soc.* 43 (1937), 532–535.
5. S. Willard, *General Topology*, Addison-Wesley, Reading, MA, 1968.

GERALD BEER received his Ph.D. from UCLA in 1971 under Frederick Valentine. He is professor of mathematics at California State University, Los Angeles, where he also serves as faculty athletics representative to the NCAA. His research interests lie in set-valued analysis, topology, and convexity. A lack of fluency notwithstanding, he is known to lecture in French. Beer is fortunate to have three sons, the eldest of which, Brendan, recently graduated in mathematics from the United States Naval Academy.
California State University, Los Angeles, Los Angeles, CA 90032
gbeer@calstatela.edu

Statistics and Mathematics: Tension and Cooperation

David S. Moore and George W. Cobb

1. INTRODUCTION. It has become a truism, at least among statisticians, that while statistics is a mathematical science, it is not a subfield of mathematics. We even have aphorisms to express some ways in which our science differs from mathematics. George Box: "All models are wrong, but some are useful." George Cobb: "In mathematics, context obscures structure. In data analysis, context provides meaning." David Moore: "Mathematical theorems are true; statistical methods are sometimes effective when used with skill." That versions of these aphorisms apply whenever mathematics models phenomena in another field only emphasizes that statistics is another field. Cobb and Moore [**2**] discuss the implications of this fact for the teaching of statistics.

Our present focus is neither the scientific distinction between mathematics and statistics nor the teaching of these subjects. We are concerned with the environment in which the fields now operate and the implications of our environment for our future prospects and for opportunities for synergy. We believe in particular that increased cooperation between the American Statistical Association (ASA) and the MAA could serve the interests of both mathematicians and statisticians.

Both mathematics and statistics are of course served by other major societies, the American Mathematical Society (AMS) and the Institute of Mathematical Statistics (IMS). These are in fact the most important societies in supporting academic research, and as such they play a large role in the culture of the two fields. The AMS is particularly significant, if only because most mathematicians are academics and most statisticians are not. Our recommendations focus on the ASA and the MAA because these groups have been more active in areas such as undergraduate education and because more statisticians are involved in the MAA than in the AMS.

We begin with a simple thesis: statistics has cultural strengths that might greatly assist mathematics, while mathematics has organizational strengths that can provide shelter for academic statistics, shelter that may be essential for its survival. Better relations between these two connected fields could help both. It is mainly cultural differences that prevent closer relations. We might say that mathematics is French, while statistics is resolutely Anglo-Saxon. The French, proud of a long history and high culture, are wary of aggressive Anglo-Saxon pragmatism. In the words attributed to a French diplomat, "It works in practice, but does it work in theory?" Cultural change is never easy, but there are signs that mathematicians are recognizing the need for some change. In doing so, they may (inadvertently?) rescue statistics from oblivion.

In developing our thesis, we serve up some potentially unpalatable implications, both for statisticians and for mathematicians. Academic statisticians will have to give up dreams of autonomy and seek shelter within the organizational strengths of mathematics. Worse yet, just at a time when various developments have made the practice of statistics less dependent on mathematics than in the past, it may be time to recognize that mathematics offers statistics intellectual as well as institutional protection. Thus the statisticians who approach our table are to be served humble pie. What do we offer our mathematician colleagues? We urge them to see their culture as others see it, and change. Worse yet, we offer our own field, statistics, as a model for that change. How do others see mathematics? The AMS recently spent seven years interviewing the chairs of research mathematics departments and their deans. Here [**6**, p. 65] is what the deans said:

> The prevalent theme in every discussion was the insularity of mathematics. Mathematicians do not interact with other departments or with faculty outside mathematics, many deans claimed, and they view this as a problem both for research and for teaching. In many cases, deans contrasted mathematics with statistics, which they pointed out had connections everywhere.

In suggesting that mathematics has become insular and statistics imperiled, we invite debate, but we attempt to argue from data and hope others will do likewise.

2. STATISTICS IS DIFFERENT. We begin by outlining the ways in which statistics is the healthier discipline. Among the encouraging vital signs, we find increasing enrollment and a consensus on teaching, more non-academic employment and links to many academic fields, and a positive response to technological change.

Increasing enrollment. The 1995 CBMS survey [**19**] details the enrollment situation: even if we look only at mathematics and statistics departments, the number of students taking elementary statistics is growing rapidly while, roughly speaking, everything else mathematics departments do is eroding.

Elementary Statistics Enrollment (thousands)

	Math depts	Stat depts	2-year colleges	Total
1990	87	30	54	171
1995	115	49	72	236

According to the same survey [**19**, p. 90] the ratio of statistics enrollment to calculus I enrollment in two-year colleges (where mathematics department enrollment is increasing) rose from 56% in 1990 to 82% in 1995. In the last three years, the birth and rapid growth of Advanced Placement statistics—roughly 7,500 exams in 1997, 15,486 in 1998, 25,240 in 1999, and 38,000 expected in 2000—means that for the first time large numbers of academically ambitious secondary school students are being exposed to modern data-centered statistics. Some will decide to study more of it.

Meanwhile, the numbers of mathematics majors and upper division enrollment have eroded seriously [**9**]:

Mathematics Enrollment (thousands)

	1985	1990	1995	% Change 1990–1995	% Change 1985–1995
Calculus	637	647	539	−17	−15
Advanced	138	119	96	−19	−30

The upward trend in statistics enrollment takes place in the context of broad agreement about what should be taught, especially in the beginning course.

Consensus on teaching. Although it would be rash to claim that any group of academics has reached a consensus on any subject, statisticians have by academic standards a clear consensus on the important subject of what we should teach beginners. As Richard Scheaffer says [**27**, p. 156], "With regard to the content of an introductory statistics course statisticians are in closer agreement today than at any previous time in my career." Contrast the wars over "reform calculus" among mathematicians.

Doctoral programs in statistics have occasioned more debate, but have evolved toward a general pattern of three equal tracks of core material: probability, statistical theory, and methodology. Equal treatment for methodology (generally both classical and computer-intensive) is relatively recent. Contrast the fact that the AMS Committee on the Profession has been unable to issue a report on the question "should the AMS advocate for more interdisciplinary research and correspondingly broadened graduate programs, and if so, how?" because of disagreement among its members [**17**].

Even in responding to the invitation of the National Council of Teachers of Mathematics to comment on the draft revision of its important *Standards* [**22**], the lack of consensus among mathematicians can be seen by contrasting the MAA group's statement that "we worked hard, and successfully, to obtain consensus on our reports" [**26**] with the AMS response that it "found that we are far from being of one mind on many issues" [**12**]. The ASA group quickly agreed on united reports.

Why is statistics enrollment increasing? Why has a consensus on teaching emerged? Causes are no doubt multiple, but both trends are surely linked to a growing recognition that statistics is a subject whose goal is to solve real-world problems. This sense of the subject is also reflected in its links to other academic subjects, and in the opportunities for non-academic employment for those who study statistics.

More non-academic employment. The 1999 AMS-ASA-IMS-MAA survey [**18**] shows continuing differences between the employment of new Ph.D.s in core mathematics and in statistics. We thank James Maxwell of AMS for providing an advance copy of this report. For 1998–1999 doctoral recipients employed in the United States:

Nonacademic employment of new Ph.D.s by field of thesis

Statistics	PAN	All others
$\frac{83}{167} = 49.7\%$	$\frac{64}{157} = 40.8\%$	$\frac{79}{435} = 18.2\%$

Here PAN aggregates the fields of probability, applied mathematics, and numerical analysis. "All others" may be taken as core mathematics. The distinctions among fields in non-academic employment, though still marked, have in fact somewhat diminished in recent years as more core mathematics Ph.D.s have sought non-academic employment. There is also considerable year-to-year variation: the percentages of non-academic employment in the three groups were 54.7%, 46.5%, and 29.4% for 1997–1998 doctoral recipients. However, this variation does not change the relative standings.

Data from the same survey show that the percentage of Ph.D.s going to women remains higher in statistics than in mathematics:

Female Ph.D.s in the mathematical sciences

Statistics	All other fields
35.7%	25.7%

Though data are lacking, we conjecture that the difference is related to the greater opportunities for non-academic employment in statistics.

The survey data on employment understate the contrast between fields because they pertain only to recipients of the Ph.D. Statistics has a meaningful and very employable professional masters degree designed to train statisticians for non-academic work. Most statistics departments distinguish the professional M.S. program from the first portion of the Ph.D. program (which may also lead to an M.S. degree). Although data are hard to come by because working statisticians with an M.S. degree do not commonly join professional societies, it appears that in terms of counts of active professionals, mathematics and statistics are disciplines of similar size. Statistics is a much smaller field in academe, but mathematics is the smaller field once we leave campus.

Statisticians' greater opportunities for non-academic employment have parallels on campus, where statisticians collaborate with colleagues in many other areas.

Links to many fields. Mathematicians are justifiably attracted by the abstract beauty of their subject, which Bertrand Russell characterized as cold and austere, like the beauty of sculpture. But in the past century, and especially in the last 50 years, pursuit of abstract beauty has often meant turning away from connections to other subjects. Mathematicians have become notorious, as their deans told the AMS interviewers, for reluctance to talk with researchers in other disciplines. Even a panel of very senior mathematicians [**24**] agrees: "Communication between mathematical scientists and other scientists is poor the world over."

The panel of course fudged a bit: "mathematical scientists" other than core mathematicians often work in environments where communication is essential. This is certainly true of non-academic statisticians, that is, of most statisticians. Statistical work in the private sector is almost always interdisciplinary, centered on the needs of clinical trials, market research, or industrial process improvement rather than on the internal development of statistics itself. Academic statisticians have gotten religion more slowly, but many now display the enthusiasm of converts. Statistics departments have long operated consulting services that both serve their campuses and train graduate students. As the culture of academic statistics has changed, leading-edge research looks less like research in mathematics and increasingly interdisciplinary and methodological, often driven by problems from other fields. It is common to find highly regarded young statistics faculty doing research in genomics, computational finance, neuroscience, data compression, and so on. Roughly half the faculty in the top-rated Stanford department hold joint appointments in other disciplines. At Carnegie Mellon, all statistics Ph.D. students are required to complete a semester-long interdisciplinary project in which they collaborate with someone outside their own department. The National Research Council's report *Modern Interdisciplinary University Statistics Education* [**23**] grew out of a symposium held almost a decade ago. In the last dozen years the NRC's

Committee on Applied and Theoretical Statistics has held symposia and published proceedings on a variety of emerging areas of research where statistics meets one client field or another.

Although statistics has been more outward-looking than mathematics throughout the past century—the only century of its existence as a separate field—its current state of energetic outreach and fruitful linkage to other subjects is a comparatively recent development, one that owes much to computers.

Energized and redirected by computing. The change in the culture and content of academic statistics has occurred in large part because of technology, which has revolutionized the teaching of statistics, reshaped the practice of statistics, and created new types of research questions.

Before computers, teachers of statistics had essentially two choices—derivations or algorithms. Some courses emphasized proofs, and lost all but the most mathematically inclined; others emphasized arithmetic, and bored all but the most desperate to pass. Neither emphasis did much to show statistics as practically useful or intellectually interesting. Perhaps it is no wonder that consensus emerged comparatively quickly once computers made it possible for students to work with real data.

The practice of statistics also felt the profound effect of computing. Now that you could do one analysis quickly, it was possible to analyze a data set in more than one way and compare the results. Once you had entered the data into a computer, the marginal effort required for additional analyses was minimal. Multiple analyses led to greater emphasis on assumption-checking and model-fitting, which in turn opened up research areas in diagnostics and statistical graphics.

Academic statistics responds (eventually) to changes in professional practice because of the importance of non-academic employment in the field and the tradition of offering consulting on campus. But the primary driver of change has been the intellectual challenge of new types of problems.

Academic statistics in the 1950s and 1960s looked inward, drawn by hopes that inductive inference could be satisfactorily mathematized. The outward pull of statistical practice is now much more noticeable. One example: 40 years ago Bayesians were working to show that their approach had desirable properties not shared by non-Bayesian methods. Although the arguments were mathematically rigorous, they converted few practitioners. Bayesian posterior distributions were simply too hard to compute in applied settings of even modest complexity. High-speed computing made it possible to revisit the computational challenge in a new environment, and academic research responded. In the last decade, Markov Chain Monte Carlo and other computer-intensive methods have made the computation of Bayesian posterior distributions much more nearly automatic. Bayesian methods are now increasingly used in practice, and "Bayesian computation" has become an active research specialty.

Computational resources that by the standards of the past are infinite and free, combined with scanning technologies in other fields that produce previously unthinkable volumes of data, open new problems concerning analysis of massive data sets; for example, see **[10]** and **[7]**. Clever use of computing has simultaneously changed the way small sets of data are analyzed, as when resampling methods allow large-scale computation to replace hard-to-justify model assumptions **[5]**. The result has been a dramatic shift in research emphases as technology moves statistics back towards its roots in data analysis and scientific inference. Links to other fields have strengthened, as freedom from restrictive assumptions has brought greater flexibility and utility to what statisticians can offer colleagues in the sciences. At the same time, computationally intensive methods (bootstrap, neural nets, wavelets) have come to be seen as core statistics. One apparent corollary is that applied work in statistics is less dependent than in the past on mathematically derived analytic solutions. We argue, however, that this important new freedom should not be mistaken for independence from mathematics.

3. STATISTICS IS, ALAS, DIFFERENT. Should statisticians gloat? Not at all. Each of the strengths we have noted has a darker side. The picture that emerges when all the pieces come together is one of organizational weakness. The advantages of statistics over mathematics in our current environment are cultural, and cultural strength rarely outweighs organizational weakness. God is on the side of the big battalions. Mathematics has the big battalions, and statistics has a few guerrillas scattered about the academic jungles. Mathematics is far likelier than statistics to have an extended, if not prosperous, future.

Weak organizational players. Statistics enrollment may be increasing, but it remains the case that many more statistics courses are taught in mathematics departments than in statistics departments. Introductory statistics is also taught in departments of psychology and economics and in schools of business and engineering.

Statisticians active in reform have reached a consensus on teaching, but they lack institutional power. Neither of their major professional organizations matches the MAA's emphasis on teaching. The IMS primarily supports academic researchers, while the ASA has a majority of non-academic members and an emphasis on serving working statisticians. The ASA's Section on Statistical Education is one of 21 sections, and only about 1,000 of ASA's roughly 17,000 members belong to it.

Statistics does have links to many fields, but this is because it is a methodological discipline rather than a core substantive area. As a result, statistics in most colleges is distributed among many core disciplinary departments (including mathematics), and is taught and practiced by faculty whose main training and interests lie elsewhere. Most large research universities have separate statistics departments, but these are typically small relative to other sciences and are rarely viewed as essential in the way that a mathematics department is essential. Statistics departments are natural targets when budget crunches demand downsizing.

This is also true outside academe. Although there are many more non-academic than academic statisticians, many corporations (Dupont, Lipton, Corning, Kodak, ...) have reduced or eliminated separate statistics units. Technology enables engineers and other directly productive employees to do much more statistics than in the past. In a survey of the state of statistics in business and industry, Gerry Hahn and Roger Hoerl of General Electric speak of "statistics without statisticians" **[11]**. Statistics is increasingly universal, but "simply performing a statistical analysis is no longer a marketable task—anyone with a laptop can do that." The situation is not entirely bleak, as discussants to Hahn and Hoerl's paper note, but the strong employment situation for statisticians is disproportionately concentrated in the pharmaceutical industry and should not be taken for granted.

The organizational weakness of statistics, growing as it does out of the methodological nature of the subject and its consequent close ties to many fields, raises a troubling question.

Does the field have a core?. The parochialism of mathematicians may be unwise, but it is explained and in part justified by the long history and continuing triumphs of a deep discipline. Mathematicians, and even some others, understand that this inward-looking field produces both profound beauty and "unreasonably effective" tools for other sciences. Statistics, in contrast, coalesced in this century from beginnings in many fields and may be about to dissipate back into many fields. It isn't clear what "statisticians" engaged in market research and molecular biology have in common. To the extent that applied problems—that is to say, research topics—in one client discipline become specialized and thus different in substance from problems in other client disciplines, statisticians become more identified with the area of their applied work than with statistics as a free-standing discipline. A pessimistic (or perhaps simply realistic) vision of the future sees statistics dissipating back into other fields.

Undergraduate mathematics has experienced a similar, though milder, dispersal. Garfunkel and Young [8] documented the migration of the teaching of mathematics beyond calculus to other academic departments. That their work has gone largely unremarked, while the CBMS survey shows continuing attrition in advanced mathematics enrollment, is a disturbing cultural indicator. The dearth of students concentrating in mathematics, which has not gone unremarked, is another indicator of poor health. If academic statistics departments risk extinction, mathematics risks a future in which it looks to others much like philosophy: an old and respected discipline engaged in intense investigation of questions of interest only to itself. Mathematics departments will no doubt continue to have larger service teaching loads than philosophy departments, cold comfort though this may be.

The threat from information technology. Although statistics has been re-energized and redirected by computing, it now risks being engulfed by information technology. Friedman [7] and others have noted that new areas in which statistical ideas offer promise are being pursued more vigorously by non-statisticians than by statisticians. Information technology is now the most important methodology for most scientific fields, displacing both statistics and mathematics from their traditional roles. Mathematics provides methodology on the side, so to speak, and will therefore survive. Statistics is inherently a methodological discipline, and is therefore at risk.

4. A BASIS FOR SYNERGY: IN THEORY, IT CAN WORK. Why shouldn't statistics simply allow itself to be reabsorbed by its client fields, or, alternatively, swallowed whole by information technology? From the other direction, why should mathematicians want to retain statistics under their organizational umbrella? Given existing cultural differences, real cooperation seems unlikely unless there are shared commitments at a deep level. Fortunately, there are. In what follows, we discuss four of them.

Undergraduate mathematics courses. Statisticians depend on, and care about, the undergraduate program in mathematics. Indeed, many faculty at top graduate departments of statistics care more about whether applicants have taken a good course in real analysis than about whether they have taken a good course in statistical methods. Although the ASA is mounting an initiative to strengthen undergraduate programs in statistics, this will in no way reduce the importance of undergraduate mathematics to statisticians. Quite the contrary. A student headed for any of the better graduate programs in statistics should have taken at least seven courses in mathematics: three in calculus, plus one each in linear algebra, probability, mathematical statistics, and real analysis. More is better. (Contrast this with how little mathematics is currently required for computer science.) If undergraduate programs in statistics grow, mathematics enrollment will benefit.

Goals for undergraduate teaching and learning. The growth of statistics enrollment is part of a larger pattern, with parallel growth in computer science and economics. The combined increases in these areas more than offset the erosion in mathematics enrollment. One inward-looking response by mathematics shrugs off the pattern as part of a shift toward a vocationalism unworthy of the Liberal Arts. Another view is this: traditional mathematics exposition and teaching often presents the subject's abstractions as completed structures. Examples come after, to serve as illustrations. Students receive a finished product, but don't participate in the *process* of building it. What statistics and other quantitative subjects offer, and traditional mathematics courses often do not, is more experience with the process of searching for patterns at a low level of abstraction before formulating a more abstract statement and then assessing its validity. Mathematicians rely on this process, of course, and there is no obstacle to their making it a more explicit part of the undergraduate mathematics curriculum. At Mount Holyoke College, for example, all mathematics majors are required to take a sophomore-level course devoted to this process, entitled "Laboratory in Mathematical Experimentation" [**21**]. When students experience the process, as opposed to just its products, the barriers between learning and research are lowered in healthy ways.

This observation is of course far from original. Constructivism migrated long ago from schools of education to departments of mathematics, and a parallel reform movement has urged science teachers to make their laboratory assignments more a process of discovery—an experience of scientific method in action—than a ritual confirmation of canonical results. The key point for our argument is that mathematicians and statisticians share a commitment to a process of pattern searching, generalization, and verification that operates at a deep level, despite surface differences.

Mathematicians who care about teaching are struggling with a variety of issues, among them the roles of technology, applied context, and the prerequisite structure that relates courses to each other [**25**]. Statisticians who care about teaching face parallel questions, but because computing is more central to what statisticians do, computers began to change approaches to teaching in statistics much earlier and more profoundly than in mathematics. For example, Minitab, a statistical software package explicitly designed to be used in teaching, was developed in the 1960s. By the early 1970s, many university departments were using computers in some of their courses and beginning to think systematically about how to bring them to the introductory level. By the mid-1980s, some introductory textbooks included computer exercises. Statisticians did not reach consensus on the introductory course without a struggle, but our struggle began much earlier than in mathematics and has been going on longer, with a greater sense of urgency. Thus statistics may be able to offer some useful models to mathematics.

The power of mathematics in statistics. Statistics, like physics and economics but unlike algebraic topology or probability theory, values mathematical understanding as a means to an end, not as an end in itself. Like physics and economics, but unlike sub-fields of mathematics, statistics has a subject matter of its own, quite apart from mathematics. While it is thus true that statistics is not a subfield of mathematics, we suggest that continued emphasis on this truth fights a battle that has been won and risks losing a more important battle. Even if statistics resists fragmentation and dispersal among its client disciplines, it still risks absorption by information technology. Ironically, a major reason for that threat is the increased intellectual distance of contemporary statistics from mathematics. Technology has greatly expanded what statistics can do, but it has also made advanced statistical tools usable by people with little knowledge of either mathematics or statistics.

Twenty years ago, selection of variables in multiple regression and several-way analysis of variance with unbalanced data were specialist topics. People who attempted such analyses were likely to have taken several statistics courses. Then, as now, statisticians insist that they have insights that go beyond suggesting appropriate methodology and that some grasp of these insights is required for skillful use of statistical software. However, today's software is much easier to use, and many introductory courses give students extensive practice. Whereas once there was a good chance that a person using multiple regression software would know a fair amount of statistics and undergraduate mathematics, that is no longer the case.

We do not deny that accessibility and wide use of statistical methods are good for science and society as a whole. But there are costs associated with this good. Elaborate studies using complex statistics, when done without much background in mathematical or statistical thinking, are prone to errors that lie deeply buried in the awful details. This is of course true whenever complex mathematical models are automated and then used without adequate grounding in mathematics and in the substance that the mathematics describes. It seems clear that *in the debate over the relative roles of thought and automated methods, statisticians and mathematicians are natural allies*.

Statistics is distinguished from "mere computing" by its extensive use of mathematical models. Classical statistical inference is largely based on the "general linear model" combining n-dimensional Euclidean geometry for structure with Gaussian distributions to model variation. The recent history of statistics is studded with examples where existing mathematical structures provide a powerful and elegant new understanding, or where the process of seeking a natural level of generality unifies a diverse collection of examples whose connections had not previously been seen clearly. The use of differential geometry to understand exponential families of distributions is a good example of the first kind (e.g., [**16**]); the emergence of the EM algorithm is an instance of the second [**3**].

We have earlier [**2**] given examples demonstrating that mathematical knowledge is insufficient for statistical understanding. We now offer a more advanced example that illustrates that mathematical knowledge is often necessary.

An example. The following example illustrates first, the shortcomings of "old-style" statistical methods, which require simplifying assumptions to make a problem analytically tractable; second, a computer intensive alternative, which requires no such simplifying assumptions, but which nevertheless makes substantial use of mathematics; third, an unexpected role for (elementary) abstract algebra; and finally, an illustration of a trap for the unwary user who neglects essential mathematics. This example is adapted from a recent article by Diaconis and Sturmfels [**4**]. The mathematics is theirs, the moral subtext is ours.

Consider testing independence of the row and column variables in a two-way table of counts. For example, to test for association between birthday and death day, let o_{ij} be the number of people in a population of interest who were born in month i and died in month j $(i, j = 1, 2, \ldots, 12)$. Diaconis and Sturmfels analyze such a table for 82 descendants of Queen Victoria. Compute row and column sums o_{i+} and o_{+j} and the grand total o_{++}, and use these to find the expected counts under the assumption of independence, $e_{ij} = o_{i+}o_{+j}/o_{++}$. Now compare observed and expected, by computing the Pearson statistic $X^2 = \Sigma(o_{ij} - e_{ij})^2/e_{ij}$. X^2 is a measure of how far the observed values fall from what we would expect if birth and death months were independent. The p-value is the proportion p of tables, among all those with the same row and column sums, for which the value of X^2 equals or exceeds the value computed from the data. Thus the p-value measures how surprising we should find the data if we believe the hypothesis of independence. If the p-value is small, we have strong evidence against that hypothesis.

How shall we calculate the proportion p? The traditional method, taught in most introductory courses on statistical methods, is the chi-square test. Make the simplifying assumption that the cell counts are large enough to justify an asymptotic argument that replaces multinomial distributions with Gaussian approximations. Then the p-value is approximately equal to the probability that a chi-square random variable with $121 = (12 - 1) \times (12 - 1)$ degrees of freedom is greater than or equal to X^2.

This is "old-style" statistics, in that simplifying assumptions that may or may not fit the data at hand are needed to render the problem analytically tractable. In this case, a lot gets lost in the rendering. One version of the simplifying assumption often used in practice is that most expected cell counts e_{ij} are at least 5; see [**20**, section 3.2.5] for a discussion. For our data, with an average cell count of $82/144$, the assumption is not appropriate. The old-style approach, "Make simplifying assumptions until you have turned the problem into one you can solve analytically", has sometimes appeared to put "use mathematics" ahead of "get the job done right." Reaction against this assignment of priorities is part of what has energized declarations of independence by statisticians.

For very small data sets, you don't need the chi-square approximation because exact enumeration is possible. "Fisher's exact test" uses the hypergeometric distribution as a shortcut, but the method is equivalent to enumeration. Enumeration avoids the unrealistic assumption of large expected cell counts, but unfortunately even computer-aided enumeration is practical only in small problems. There are simply too many possibilities. Diaconis and Sturmfels give a 4×4 data set with 592 observations for which 1.2×10^{18} tables have the same row and column totals. Alternatively, or in addition, there may be specific constraints that make enumeration difficult. For tables in three or more dimensions, for example, the hypothesis of interest may not be as simple as independence of the individual dimensions, and the constraints may involve several overlapping two-dimensional arrays of totals to be matched. Or there may be "structural zeros" in the data, e.g., in a table with dimensions for gender and for site of cancer. The fact that computer-aided enumeration is not often used for problems of this sort highlights the point that in statistics "computer-intensive" means something more subtle than just very-high-speed plodding.

The current computer-intensive method for computing p-values for this problem is known as Markov Chain Monte Carlo, the same method mentioned earlier for computing Bayesian posterior distributions. To apply it here, define a connected, aperiodic, reversible Markov Chain whose states are all the tables of size 12×12 that meet the required constraints (e.g., with non-negative integer entries having the given row and column totals). One standard way to do this is to assign transition probabilities as follows. Pick two rows i and j and two columns r and s at random. Move from the current data table by adding 1 to cells (i, r) and (j, s) while subtracting 1 from (i, s) and (j, r); stay in place if the chosen move would produce a negative entry. This chain has a known equilibrium distribution, which can be modified (by accepting or rejecting moves using coin flips with suitable bias) to give the desired sampling distribution for the set of tables. Run the (modified) chain long enough to reach equilibrium, then sample from it, say, every 50 steps. Approximate the p-value by the proportion of sampled tables whose X^2 value equals or exceeds the one computed from the data.

Clearly, this elegant computer-intensive method could not have been developed without mathematics. More to the point, even its use is made easier, and in some instances made possible, by mathematical understanding that is not at all obvious. A major contribution of the paper by Diaconis and Sturmfels is to show that defining the Markov Chain (finding a Markov basis) is equivalent to finding the generators of an ideal in a polynomial ring. Thus methods of computational algebra, and a computer system like Maple, can be used to set up the Markov Chain.

As a final point, we repeat an observation of Diaconis and Sturmfels, that connectedness of the chain really matters, and cannot be taken for granted. They cite a published example where inattention to this detail led to an incorrect p-value.

Applied problems as an inspiration for research. In theory, at least, mathematics and statistics share a commitment to developing new tools to solve applied problems. In statistics, attention to the needs of other fields has always been responsible for much of the profession's evolution. In the long history of mathematics as well, applied problems were traditionally the source of most new mathematics. In theory, yes; in statistics, yes; in history, yes. In practice, in mathematics, at the present time, research appears to be disproportionately driven by the internal development of mathematics itself. We hope that mathematicians will find the history of their own subject persuasive, the specter of academic philosophy galvanizing, and therefore, the example of statistics useful.

5. CAN SYNERGY WORK IN PRACTICE. In theory, as we have seen, there is a substantial basis for cooperation:

- Graduate programs in statistics need undergraduate programs in mathematics, at a time when mathematics programs need to be needed.
- Despite intellectual differences, mathematics and statistics both depend on the process of working from the concrete to the abstract, and can learn from each other's successes and failures in teaching this process to undergraduates. In addition, because statisticians were confronted much earlier by computer-driven challenges in curriculum and pedagogy, mathematicians may be able to benefit from their experience.

- Statistics can benefit from embracing more openly the importance of mathematical thinking, particularly in research and practice near its boundary with computer science, where statistics is especially threatened.
- At the same time, mathematics can benefit from the experience of statistics over the last 30 years, during which computing has redirected energies back towards the roots of statistics in applied science. A parallel shift in orientation for mathematics would be consistent with its nineteenth-century history of ties to the sciences.

Can cooperation work in practice? The recent experience of mathematics at the University of Rochester (**[13]**, **[14]**, **[15]**) offers some encouragement. The organizational strength of mathematics, via the Rochester Task force of the AMS, played an influential role in rescuing the department from a threatened elimination of the Ph.D. program and 50% cut in faculty. The key to the compromise "Renaissance Plan" for the mathematics department emerged from what the university's president described **[13]** as "a series of unprecedented conversations between math faculty and the administration and between math faculty and their colleagues in other departments. That, in turn, led to the Department of Physics and Astronomy's offer to promote linkages by joint appointments and a new dedication on the part of the mathematics faculty to strengthen undergraduate instruction and their ties to other departments . . . ."

In the remainder of this section, we offer two clusters of remarks on cooperation between mathematics and statistics, first recounting some good beginnings, then suggesting some promising continuations.

Good beginnings. Cooperation between professional societies can help bridge the culture gap and help mathematics and statistics to profit from their complementary strengths. Substantial cooperation between ASA and MAA is already in evidence.

Joint Committee. In 1992 the MAA and ASA formed a joint committee on undergraduate statistics. This group has done useful things, and statisticians within the MAA have done more: workshops for mathematicians teaching statistics, volumes on statistics published by the MAA, lobbying for adequate treatment of statistics in MAA guidelines for departments. The MAA has been generally receptive to the concerns of its statistician members.

SIGMAAs. The ASA has an active "isolated statisticians group" that supports statisticians teaching in mathematics departments, where they lack statistical colleagues. Activities include meetings at the Joint Statistical Meetings and elsewhere and an active listserve (isostat@oberlin.edu). The MAA is now forming SIGMAAs, special interest groups analogous to ASA's sections. A statistics education SIGMAA will be among the first formed.

Guidelines. The MAA *Guidelines for Programs and Departments in Undergraduate Mathematical Sciences* can greatly influence the reception offered statistics and statisticians in mathematics departments. The current draft offers a warm reception. For example, it urges recognition of consulting as scholarly work and stresses that instruction in statistics and other border areas should be developed and supervised by faculty fully trained in the area. Acceptance of these recommendations by the MAA would be an important step in changing culture.

Annual meetings. The ASA/MAA joint committee has contributed to a greater presence of statistics at the Joint Mathematics Meetings. The MAA has been generally hospitable to statistical sessions and short courses, and in each of the last few years there has been both a well-attended minicourse related to the teaching of statistics and a well-attended contributed paper session with enough papers to fill six hours spread over three days.

Suggestions for the future. Here are some suggestions for the MAA, for the AMS and the MAA jointly representing the mathematics profession, or for mathematical and statistical societies working together, to build from our promising beginnings.

Teach statistics well. Consider these data for comparable departments on the percentage of elementary statistics enrollments treated to two "good practices" [**19**, p. 71]:

Elementary Statistics (1995)

	Ph.D. Math Depts	Ph.D. Stat Depts
Tenure-track instructor	29%	46%
Computer assignments	42%	61%

The performance of the statistics departments is nothing to be proud of, but the gap between mathematics and statistics departments is clear. Statistics is almost the only offering of mathematics departments that is growing. Self-interest alone suggests that it be given greater attention.

The ASA/MAA joint committee has taken the lead here; see its report [**1**]. The ASA Board of Directors has unanimously endorsed a short version as guidance for beginning instruction in statistics. We hope that the MAA will look closely at this issue as well. Because more statistics teaching is done by mathematics departments than by statistics departments, statisticians have a major stake in how mathematics departments teach statistics. Mathematics departments should want to offer modern data-centered elementary statistics courses to satisfy what is now their fastest-growing clientele.

Undergraduate programs. If statistics *courses* are attracting more student interest and the rapid rise of AP Statistics promises more interest, development of undergraduate statistics *curricula* is the next step. In all but large institutions, where full undergraduate major programs in statistics are possible, the focus may be on interdisciplinary concentrations in statistics that draw on quantitative elements from several disciplines. Mathematics departments can take the lead on their campuses, and ASA and MAA might cooperate to provide thoughtful guidelines.

Consulting and joint appointments. Statisticians have learned how to organize and fund successful consulting services, and joint appointments are common in statistics departments. Consulting services are a good indicator of how well a statistics department connects with its campus, as they not only serve many disciplines but also lead almost inevitably to joint research. A campus that, like most, lacks a separate statistics department offers an opportunity to the mathematics department: start a consulting service. Statisticians on the mathematics faculty may be the first to draw business, but many areas in mathematics are relevant to the work of scholars in other fields. Of course, the department must be willing to recognize consulting on and off campus as scholarly work, a change of attitude for many mathematicians.

Research and graduate programs. Statistics graduate programs are vastly different from those of the 1960s, with much more emphasis on computational, methodological, and applied research. Most statisticians understand that "applied research" is applied *to* something; this is not traditional usage in mathematical circles. The resistance of many mathematicians to "more interdisciplinary research and correspondingly broadened graduate programs" (to borrow the phrasing of the AMS Committee on the Profession) paves the road to becoming another philosophy department. Due to the differing nature of the disciplines, the culture of the relatively few leading research departments in mathematics and statistics may well diverge. Leading statistics departments already have strong programs of methodological and interdisciplinary research. Their counterparts in mathematics, most of which are at institutions with separate statistics departments, tend to concentrate on core mathematics. Other mathematics departments, especially those in institutions without separate statistics departments, may find it attractive to offer graduate concentrations in statistics, operations research, and applied mathematics. Their graduate programs will naturally then evolve in directions that no longer follow the path of leading departments. Continued AMS discussion aimed at legitimizing graduate programs that do not attempt to imitate the few leading research institutions would be welcome.

Publications. It may be that individual mathematics and statistics societies lack the scale needed to succeed in the new environment of on-line publication, which favors groups that can offer libraries searchable archives of large numbers of journals and other publications. Exploration of possibilities for cooperation among societies is natural.

Annual meetings. A good start has been made, as noted earlier, but much remains to be done. The January Joint Mathematics Meetings are important both for statisticians seeking academic employment and for mathematics departments seeking to hire statisticians, as the meetings are the natural locus for job interviews. The Joint Statistical Meetings, held in August, cannot fill this role. The timing of the two major meetings illustrates once more that mathematics is primarily an academic discipline and that statistics is not. ASA can contribute here by inaction: it should encourage attendance at the mathematics meetings rather than, as has been suggested, sponsor a winter statistics meeting. AMS and MAA can lend force to ASA's inaction by doing still more to fill the vacuum left for statisticians by the absence of ASA meetings in January.

A final plea. Having made several pleas to mathematicians, we end with a simple plea to academic statisticians: **Join the MAA.** To succeed in and change a foreign culture, immigrants must be willing, at least partially, to assimilate.

6. CONCLUSION. Mathematics, a core discipline, looks inward and risks being seen as increasingly irrelevant. Statistics, a methodological discipline, looks outward but risks being swallowed by information technology. Both professions have a stake in the survival of statistics as a subject informed and structured by mathematics. To mathematics, statistics offers not only the example of an outward looking culture, but also entree to new problems ripe for mathematical study. To statistics, mathematics offers not only the safe harbor of organizational strength, but intellectual anchorage as well: mathematical understanding is an essential part of what distinguishes statistical thinking from most of the rest of information technology. Increased cooperation between mathematical and statistical professional associations can lead the societies, their members, and their disciplines in healthier directions.

REFERENCES

1. George W. Cobb, Teaching statistics, in L. A. Steen (ed.), *Heeding the Call for Change: Suggestions for Curricular Action*, Mathematical Association of America, Washington, DC, 1992, pp. 3–43.
2. George W. Cobb and David S. Moore, Mathematics, statistics, and teaching, *Amer. Math. Monthly* **104** (1997) 801–823.
3. Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. B* **39** (1997) 1–38.
4. Persi Diaconis and Bernd Sturmfels, Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.* **26** (1998) 363–397.
5. Bradley Efron and Rob Tibshirani, Statistical data analysis in the computer age, *Science* **253** (1991) 390–395.
6. John Ewing, (ed.), *Towards Excellence: Leading a Mathematics Department in the 21st Century*, American Mathematical Society, Providence, RI, 1999.
7. Jerome Friedman, Data mining and statistics: what's the connection? http://www-stat.stanford.edu/~jhf/, 1997.
8. Solomon A. Garfunkel and Gail S. Young, *Math Outside of Math: A Study of Mathematics Enrollments Outside of Mathematics Departments*, COMAP, Lexington, MA, 1989. See also, by the same authors, Mathematics outside of mathematics departments, *Notices Amer. Math. Soc.* **37** (1990) 408–411.
9. Solomon A. Garfunkel and Gail S. Young, The sky is falling, *Notices Amer. Math. Soc.* **45** (1998) 256–257.
10. Clark Glymour, David Madigan, Daryl Pregibon, and Padhraic Smyth, Statistical themes and lessons for data mining, *Data Mining and Knowledge Discovery* **1** (1997) 11–28.
11. Gerry Hahn and Roger Hoerl, Key challenges for statisticians in business and industry, *Technometrics* **40** (1998) 195–200.
12. Roger Howe, The AMS and mathematics education: the revision of the "NCTM Standards," *Notices Amer. Math. Soc.* **45** (1998) 243–247.
13. Allyn Jackson, Breakthrough at Rochester: Ph.D. program restored, *Notices Amer. Math. Soc.* **43** (1996) 562–563.
14. Allyn Jackson, Beyond Rochester: Mathematics departments come to grips with university downsizing, *Notices Amer. Math. Soc.* **43** (1996) 572–575.
15. Allyn Jackson, Whatever happened to Rochester? Two years later, mathematics is getting accolades, *Notices Amer. Math. Soc.* **44** (1997) 1463–1464.
16. Robert E. Kass and Paul W. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley, New York, 1997.
17. Joseph Lipman, 1997 report of the Committee on the Profession, *Notices Amer. Math. Soc.* **45** (1998) 739–740.
18. Don O. Loftsgaarden, James W. Maxwell, and Kinda M. Remick, 1999 Annual survey of the mathematical sciences, first report, *Notices Amer. Math. Soc.* **47** (2000) 231–243.
19. Don O. Loftsgaarden, Donald C. Rung, and Ann E. Watkins, *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States*, 1995 *CBMS Survey*, MAA Reports No. 2, Mathematical Association of America, Washington, DC, 1997.

20. David S. Moore, Tests of chi-squared type, in Ralph B. D'Agostino and Michael A. Stephens, (eds.), *Goodness-of-Fit Techniques*, Marcel Dekker, New York, 1986, pp. 63–95.
21. Mount Holyoke College, *Laboratories in Mathematical Experimentation: A Bridge to Higher Mathematics*, Springer-Verlag, New York, 1997.
22. National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards For School Mathematics*, NCTM, Reston, VA, 1989.
23. National Research Council, *Modern Interdisciplinary University Statistics Education*, National Academy Press, Washington, DC, 1994.
24. National Science Foundation, *Report of the Senior Assessment Panel for the International Assessment of the U.S. Mathematical Sciences*, document NSF 98-95, Washington, DC, 1998.
25. Donald B. O'Shea and Harriet S. Pollatsek, Do we need prerequisites?, *Notices Amer. Math. Soc.* **44** (1997) 564–570.
26. Kenneth A. Ross and the President's Task Force on the NCTM Standards, Report of the MAA ARG, *MAA Online*, http://www.maa.org/features/maa-dnctm99.html, 1999.
27. Richard Scheaffer, Discussion of David S. Moore, New pedagogy and new content: the case of statistics, *Int. Statist. Rev.* **65** (1997) 156–158.

DAVID S. MOORE is Shanti S. Gupta Distinguished Professor of Statistics at Purdue University. He received his A.B. from Princeton and his Ph.D. from Cornell, both in mathematics.His scholarly interests have moved from statistical large sample theory toward statistics education, driven both by middle age and by strong interests in technology. He was 1998 president of the American Statistical Association.
Purdue University, West Lafayette, IN 47907-1399
dsmoore@stat.purdue.edu

GEORGE W. COBB is Robert L. Rooke Professor in the Department of Mathematics and Statistics at Mount Holyoke College. He has an A.B. in Russian from Dartmouth, MS in Biometry from the Medical College of Virginia, and Ph.D. in Statistics from Harvard. Like his co-author Moore, Cobb is being driven somewhere by middle age, but he is still waiting to discover where that may be. He has chaired the ASA/MAA Joint Committee on Undergraduate Statistics, and currently serves on the National Research Council's Committee on Applied and Theoretical Statistics.
Mount Holyoke College, South Hadley, MA 01075
gcobb@mhc.mtholyoke.edu

NOTES

Edited by **Jimmie D. Lawson and William Adkins**

Nilpotent Numbers

Jonathan Pakianathan and Krishnan Shankar

Introduction. One of the first things we learn in abstract algebra is the notion of a cyclic group. For every positive integer n, we have $\mathbb{Z}_n$, the group of integers modulo n. When n is prime, a simple application of Lagrange's theorem yields that this is the *only* group of order n. We may ask ourselves: what other positive integers have this property? In this spirit we call a positive integer n a *cyclic number* if every group of order n is cyclic. We define *abelian* and *nilpotent* numbers analogously. Recall that a group is nilpotent if and only if it is the (internal) direct product of its Sylow subgroups; see [**7**, 126].

This is not a new problem; the cyclic case is attributed to Burnside and has appeared in numerous articles, [**9**], [**4**], [**1**], [**2**]. The abelian case appears as a problem in an old edition of Robinson's book in group theory; see also [**6**] and the nilpotent case was also done quite some time ago (see [**5**], [**6**]). In this article we give an arithmetic characterization of the cyclic, abelian, and nilpotent numbers from a single perspective. Throughout this paper $\mathbb{Z}_n$ denotes the cyclic group of order n.

Nilpotent numbers. The smallest non-prime cyclic number is 15. This follows from [**3**, Proposition 6.1, p. 98] where it is shown that for primes p and q, if $p > q$, then pq is a cyclic number if and only if $q \nmid (p-1)$. Motivated by this arithmetic criterion we make the following definition.

Definition. A positive integer $n = p_1^{a_1} \cdots p_t^{a_t}$, p_i distinct primes, is said to have *nilpotent factorization* if and only if $p_i^k \not\equiv 1 \bmod p_j$ for all integers i, j and k with $1 \le k \le a_i$.

Examples of numbers with nilpotent factorization are all powers of prime numbers and pq where $p > q$ are prime and $q \nmid (p-1)$. For example, the number $21 = 3 \cdot 7$ does not have nilpotent factorization since $7 \equiv 1 \bmod 3$. It turns out that this rather strange looking property characterizes nilpotent numbers.

Theorem 1. *A positive integer n is a nilpotent number if and only if it has nilpotent factorization.*

Proof: Suppose $n = p_1^{a_1} \cdots p_t^{a_t}$ is a positive integer without nilpotent factorization. Then there exist i, j, and k with $1 \le k \le a_i$ such that $p_i^k \equiv 1 \bmod p_j$. Note that p_i and p_j are necessarily distinct so after relabelling we may assume $p_1^k \equiv 1 \bmod p_2$ for some $1 \le k \le a_1$. Let E be the elementary abelian group consisting of the direct product of k copies of $\mathbb{Z}_{p_1}$ i.e., $E = \mathbb{Z}_{p_1}^k$. E can also be viewed as a k-dimensional vector space over $\mathbb{F}_{p_1}$, the finite field with p_1 elements (isomorphic to $\mathbb{Z}_{p_1}$ as a group). Then the group of vector space automorphisms of E is $\mathrm{Aut}(E) \cong GL_k(\mathbb{F}_{p_1})$. The latter is the group of $k \times k$ matrices with entries in $\mathbb{F}_{p_1}$ and non-zero determinant modulo p_1. The order of $GL_k(\mathbb{F}_{p_1})$ is $(p_1^k - 1)(p_1^k - p_1) \cdots (p_1^k - p_1^{k-1})$. By assumption $p_1^k \equiv 1 \bmod p_2$, so $p_2 | (p_1^k - 1)$ and hence p_2 divides $|GL_k(\mathbb{F}_{p_1})|$. Then $\mathrm{Aut}(E)$ has a subgroup isomorphic to $\mathbb{Z}_{p_2}$ by Cauchy's theorem and we may form a non-trivial semi-direct product, $E \rtimes \mathbb{Z}_{p_2}$. Now consider the group

$$G = \left(E \rtimes \mathbb{Z}_{p_2}\right) \times \mathbb{Z}_{p_1}^{a_1 - k} \times \mathbb{Z}_{p_2}^{a_2 - 1} \times \mathbb{Z}_{p_3}^{a_3} \times \cdots \times \mathbb{Z}_{p_t}^{a_t}.$$

By construction, G is a group of order n. In a nilpotent group, elements in Sylow subgroups corresponding to distinct primes commute with each other. The elements of E all have order p_1 and they don't commute with the elements of $\mathbb{Z}_{p_2}$ in the semi-direct product $E \rtimes \mathbb{Z}_{p_2}$, by construction. Hence G is not nilpotent and consequently n is not a nilpotent number.

For the converse, we wish to show that if n has nilpotent factorization, then it is a nilpotent number. Suppose this is not true. Let n be the smallest positive integer with nilpotent factorization that is not a nilpotent number. Then there exists a group G of order n that is not nilpotent. If H is any proper subgroup of G, then $|H|$ has nilpotent factorization also. H must be nilpotent, since we assumed n to be the smallest non-nilpotent integer with nilpotent factorization. So G is a non-nilpotent group with every proper subgroup nilpotent. By a theorem of O. J. Schmidt [**9**, 9.1.9. p. 251], such groups are rather special and we must have $n = |G| = p^a q^b$, where p, q are distinct primes and $a, b \geq 1$.

Let n_p and n_q denote the number of Sylow p-subgroups and Sylow q-subgroups, respectively, of G. By Sylow's theorem, $n_p \equiv 1 \bmod p$, but it is also equal to the index of the normalizer, $N_G(S_p)$, of some Sylow p-subgroup S_p in G. Now $S_p \subset N_G(S_p) \subset G$. So the order of $N_G(S_p)$ is $p^a q^k$ for some integer k, and has index $q^{b-k} = n_p \equiv 1 \bmod p$ in G. By assumption $|G| = p^a q^b$ has nilpotent factorization, which forces $b - k = 0$. This implies $N_G(S_p) = G$ and hence S_p is unique and normal in G. The same argument applied to q shows that the Sylow q-subgroup, S_q, is also unique and normal. Hence, $G \cong S_p \times S_q$, which contradicts our assumption that G was not nilpotent. So if n has good factorization, then it must be a nilpotent number. ■

We will see that this also characterizes cyclic and abelian numbers since we have the containments

$$\textit{cyclic groups} \subset \textit{abelian groups} \subset \textit{nilpotent groups}.$$

Recall that a positive integer $n = p_1^{a_1} \cdots p_t^{a_t}$ is said to be *cube–free* if $a_i \leq 2$ for all i. It is said to be *square–free* if $a_i = 1$ for all i.

Abelian numbers. Given a prime p, there is always a non-abelian group of order p^3. For example,

$$T_p = \left\{ \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} : a, b, c \in \mathbb{Z}_p \right\},$$

where addition and multiplication of entries is performed modulo p, is one such group for every prime p. So an abelian number is necessarily cube–free. We claim that n is an abelian number if and only if it is a cube–free number with nilpotent factorization.

Suppose n is a cube–free nilpotent number and let G be a group of order n. Then G is nilpotent and $G \cong S_{p_1} \times \cdots \times S_{p_t}$, i.e., G is isomorphic to the product of its Sylow subgroups. Since n was assumed to be cube−free, each S_{p_i} has order p_i or p_i^2 and is hence, abelian. G is then abelian, being a product of abelian groups, and n is an abelian number.

Conversely, if n is an abelian number, then it must be a nilpotent number and hence it has nilpotent factorization. We noted that n is necessarily cube−free; if not, then there exists a prime p such that $p^3|n$. Then $T_p \times \mathbb{Z}_{n/p^3}$ is a non-abelian group of order n, contradicting the assumption that n is an abelian number. This completes the argument and establishes our claim.

Cyclic numbers. We now claim that n is a cyclic number if and only if it is a square−free number with nilpotent factorization. The argument here is along the same lines as for the abelian case once we note that $\mathbb{Z}_p \times \mathbb{Z}_p$ is a non-cyclic group of order p^2.

This characterization is equivalent to another well known characterization of cyclic numbers. Let $\varphi(n)$ be the Euler totient function of n. It counts the number of positive integers less than or equal to n that are relatively prime to n. For $n = p_1^{a_1} \cdots p_t^{a_t}$,

$$\varphi(n) = \left(p_1^{a_1-1}(p_1 - 1)\right) \cdots \left(p_t^{a_t-1}(p_t - 1)\right)$$

Note that if n is square−free, then $\varphi(n) = (p_1 - 1) \cdots (p_t - 1)$. Our claim says that n is a cyclic number if and only if it has nilpotent factorization and it is square−free. This is equivalent to saying $p_i \nmid (p_j - 1)$ for all i, j, which is equivalent to saying $\gcd(n, \varphi(n)) = 1$. This yields the elegant result: A positive integer n is a cyclic number if and only if $\gcd(n, \varphi(n)) = 1$.

Remark. The only even numbers with nilpotent factorization are powers of 2. Let $f(n)$ denote the number of groups of order n. If $n = p_1^{a_1} \cdots p_t^{a_t}$ is an abelian number, then $f(n) = 2^{\Sigma(a_i - 1)}$. The problem of determining $f(n)$ is quite hard in general and beyond reach even for the nilpotent numbers. This is because estimating $f(p^k)$ for all primes p and all integers k, is too difficult a problem at this time.

Remark. Using a deep result of J. Thompson's on minimal simple groups **[10]** which ultimately relies on the celebrated Feit−Thompson theorem, it is possible to characterize the solvable numbers as well. We can show that a positive integer n is a solvable number if and only if it is not a multiple of any of the following numbers:

(a) $2^p(2^{2p} - 1)$, p any prime.
(b) $3^p(3^{2p} - 1)/2$, p an odd prime.
(c) $p(p^2 - 1)/2$, p any prime greater than 3 such that $p^2 + 1 \equiv 0 \bmod 5$.
(d) $2^4 \cdot 3^3 \cdot 13$.
(e) $2^{2p}(2^{2p} + 1)(2^p - 1)$, p an odd prime.

As a corollary we see that an integer not divisible by 4 must be a solvable number. In particular, every odd number is a solvable number, as expected.

ACKNOWLEDGMENTS. We thank Jørgen Tornehave for useful discussions and we thank the mathematics department of Aarhus University, Denmark for their hospitality.

REFERENCES

1. L. E. Dickson, Definitions of a group and a field by independent postulates, *Trans. Amer. Math. Soc.* **6** (1905) 198–204.
2. J. A. Gallian and D. Moulton, When is $\mathbb{Z}_n$ the only group of order n?, *Elem. Math.* **48** (1993) 117–119.

3. T. Hungerford, *Algebra*, Graduate Texts in Math. **73**, Springer, New York, 1974.
4. D. Jungnickel, On the uniqueness of the cyclic group of order n, *Amer. Math. Monthly* **99** (1992) 545–547.
5. G. Pazderski, Die Ordnungen, zu denen nur Gruppen mit gegebener Eigenschaft gehören, *Arch. Math.* **10** (1959) 331–343.
6. L. Rédei, Das "schiefe Produkt" in der Gruppentheorie mit Anwendung auf die endlichen nichtkommutativen Gruppen mit lauter kommutativen echten Untergruppen und die Ordnungszahlen, zu denen nur kommutative Gruppen göheren, *Comm. Math. Helv.* **20** (1947) 225–264.
7. D. Robinson, *A course in the theory of groups*, Graduate Texts in Math. **80**, Springer, New York, 1993.
8. T. Szele, Über die endlichen Ordnungszahlen, zu denen nur eine Gruppe gehört, *Comm. Math. Helv.* **20** (1947) 265–267.
9. J. Szép, On finite groups which are necessarily commutative, *Comm. Math. Helv.* **20** (1947) 223–224.
10. J. G. Thompson, Nonsolvable finite groups all of whose local subgroups are solvable, *Bull. Amer. Math. Soc.* **74** (1968) 383–437.

University of Wisconsin, Madison, WI 53706
pakianat@math.wisc.edu

University of Michigan, Ann Arbor, MI 48109
shankar@umich.edu

Triangular Triples from Ceilings to Floors

Tom Jenkyns and Eric Muller

1. Introduction. A *triangular triple* is a sequence of non-negative integers (i, j, k) that gives the lengths of the sides of a triangle. Then each integer is at most the sum of the other two. We restrict our attention to incongruent triangles and therefore to triples where $i \le j \le k$ and $k \le i + j$, since any two triangles with these side-lengths are congruent. The associated triangle has perimeter, $p = i + j + k$. When one of p or k or j is fixed, just how many triangular triples are there?

In fact we shall count four types of triples. Let A denote the set of *all* triangular triples, let B denote the set of all *non-degenerate* triangular triples, let C denote the set of all *scalene* triangular triples, and let D denote the set of all triangular triples that are *both* scalene and non-degenerate. For each of these sets $\mathscr{A}$, let $\mathscr{A}(p)$ denote the subset of triples in $\mathscr{A}$ with sum equal p, and let $T_\alpha(p)$ denote the cardinality of $\mathscr{A}(p)$. The first few values of these functions appear somewhat chaotic:

Table 1

p	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
T_a	1	0	1	1	2	1	3	2	4	3	5	4	7	5	8	7	10	8
T_b	0	0	0	1	0	1	1	2	1	3	2	4	3	5	4	7	5	8
T_c	0	0	0	0	0	0	1	0	1	1	2	1	3	2	4	3	5	4
T_d	0	0	0	0	0	0	0	0	0	1	0	1	1	2	1	3	2	4

The main purpose of this note is to provide formulas for the four functions $T_\alpha(p)$. Past attention has focused on non-degenerate triangles [**1**]–[**5**], though [**2**] and [6] also examine scalene triples. We give a new and very elementary proof of an expression for $T_b(p)$ and from it derive several previously obtained expressions. Theorem 2 settles the conjecture on triangles with longest side k that is given at the end of [2].

2. Relations among the $T_\alpha(p)$. For all values of p and integer triples where $i + j + k = p$

$$A(p) = \{(i, j, k) : 0 \le i \le j \le k \le i + j\}$$

and

$$B(p) = \{(i, j, k) : 1 \le i \le j \le k < i + j\}.$$

Then $B(p) = \phi$ unless $p \ge 1 + 1 + 1 = 3$. If $i = 0$ then $j \le k \le i + j$ implies $j = k$ and the triple must be isosceles. Thus,

$$C(p) = \{(i, j, k) : 1 \le i < j < k \le i + j\}$$

and $C(p) = \phi$ unless $p \ge 1 + 2 + 3 = 6$. Also if $i = 1$ and $j < k$ then $1 + j \le k \le i + j$ implies $k = i + j$ and the triple must be degenerate. Thus,

$$D(p) = \{(i, j, k) : 1 < i < j < k < i + j\}$$

and $D(p) = \phi$ unless $p \ge 2 + 3 + 4 = 9$.

Theorem 1. *For all $p \ge 3$, $T_b(p) = T_a(p - 3)$, $T_c(p) = T_b(p - 3)$, and $T_d(p) = T_c(p - 3)$.*

Proof: If $(i, j, k) \in A(p - 3)$ then $0 \le i \le j \le k \le i + j$. So $1 \le i + 1 \le j + 1 \le k + 1 \le i + j + 1 < (i + 1) + (j + 1)$ and $(i + 1, j + 1, k + 1) \in B(p)$. Therefore, $T_a(p - 3) \le T_b(p)$.

If $(i, j, k) \in B(p)$ then $1 \le i \le j \le k < i + j$. So $0 \le i - 1 \le j - 1 \le k - 1 < i + j - 1$, $k - 1 \le (i - 1) + (j - 1)$ and $(i - 1, j - 1, k - 1) \in A(p - 3)$. Therefore, $T_b(p) \le T_a(p - 3)$. Furthermore, $1 \le i < j + 1 < k + 2 < i + j + 2$ so $k + 2 \le i + (j + 1)$ and $(i, j + 1, k + 2) \in C(p + 3)$. Therefore, $T_b(p) \le T_c(p + 3)$.

If $(i, j, k) \in C(p)$ then $1 \le i < j < k \le i + j$. So $1 \le i \le j - 1 \le k - 2 \le i + j - 2 < i + (j - 1)$ and $(i, j - 1, k - 2) \in B(p - 3)$. Therefore, $T_c(p) \le T_b(p - 3)$. Furthermore, $1 < i + 1 < j + 1 < k + 1 \le i + j + 1 < (i + 1) + (j + 1)$ and $(i + 1, j + 1, k + 1) \in D(p + 3)$. Therefore, $T_c(p) \le T_d(p + 3)$.

If $(i, j, k) \in D(p)$ then $1 < i < j < k < i + j$. So $1 \le i - 1 < j - 1 < k - 1 < i + j - 1$, $k - 1 \le (i - 1) + (j - 1)$ and $(i - 1, j - 1, k - 1) \in C(p - 3)$. Therefore, $T_d(p) \le T_c(p - 3)$. ■

We observe that if (i, j, k) is degenerate, then $k = i + j$, so $p = 2k$. Thus, when p is odd, no triangular triples are degenerate and we have $T_b(p) = T_a(p) = T_b(p + 3)$, where $p + 3$ is even. Therefore we can evaluate all $T_\alpha(p)$ if we have an explicit formula for $T_b(p)$ when p is *even*. We construct such a formula in the final sections.

3. Triangles with largest side k. For each non-negative integer k, let $L_\alpha(k)$ denote the number of triangular triples of each of the four types described in

Section 1, with longest side k. The first few values of these functions are given in Table 2.

Table 2

k	0	1	2	3	4	5	6	7	8	9	10
L_a	1	2	4	6	9	12	16	20	25	30	36
L_b	0	1	2	4	6	9	12	16	20	25	30
L_c	0	0	0	1	2	4	6	9	12	16	20
L_d	0	0	0	0	1	2	4	6	9	12	16

The same arguments used in the proof of Theorem 1 give

Corollary 1. $(i, j, k) \in A$ if and only if $(i+1, j+1, k+1) \in B$;
$(i, j, k) \in B$ if and only if $(i, j+1, k+2) \in C$;
and $(i, j, k) \in C$ if and only if $(i+1, j+1, k+1) \in D$.

From this follows

Theorem 2. *For all* $k \geq 0$, $L_a(k) = L_b(k+1)$, $L_b(k) = L_c(k+2)$, *and* $L_c(k) = L_d(k+1)$.

Therefore we can evaluate all $L_\alpha(k)$ when we have an explicit formula for $L_b(k)$. Such a formula is easy to derive.

In non-degenerate triples $2k < (i+j) + k = p \leq k + k + k = 3k$. Thus for a fixed value of k, $2k + 1 \leq p \leq 3k$; and for a fixed value of p, $p/3 \leq k \leq (p-1)/2$. Because k is an integer

$$\lceil p/3 \rceil \leq k \leq \lfloor (p-1)/2 \rfloor$$

where for any real number x, $\lfloor x \rfloor$, the *floor function* of x, is the largest integer $\leq x$ and $\lceil x \rceil$, the *ceiling function* of x, is the smallest integer $\geq x$. That is; if p is fixed, k ranges from the ceiling-value, $\alpha = \lceil p/3 \rceil$, to the floor-value, $\beta = \lfloor (p-1)/2 \rfloor$.

Let $f(p, k)$ denote the number of triangular triples (degenerate or not) with perimeter p and largest entry k. Clearly the largest value for j is k in which case i takes its smallest possible value, $p - 2k$. Since $2i \leq i + j = p - k$, $i \leq (p-k)/2$; so the largest possible value of i is $\lfloor (p-k)/2 \rfloor$. These triples in lexicographic order are:

$$\begin{array}{lll}
(p-2k & , \quad k & , \ k) \\
(p-2k+1 & , \quad k-1 & , \ k) \\
(p-2k+2 & , \quad k-2 & , \ k) \\
\cdots\cdots & \cdots\cdots & \cdots \\
(\lfloor (p-k)/2 \rfloor & , \ \lceil (p-k)/2 \rceil & , \ k)
\end{array}$$

Hence

$$\begin{aligned}
f(p, k) &= \lfloor (p-k)/2 \rfloor - (p-2k) + 1 = \lfloor (p-k)/2 - (p-2k) + 1 \rfloor \\
&= \lfloor (3k - p + 2)/2 \rfloor .
\end{aligned}$$

Then

$$L_b(k) = \sum_{p=2k+1}^{3k} \lfloor (3k - p + 2)/2 \rfloor = \sum_{m=2}^{k+1} \lfloor m/2 \rfloor$$
$$= \lfloor 2/2 \rfloor + \lfloor 3/2 \rfloor + \lfloor 4/2 \rfloor + \lfloor 5/2 \rfloor + \cdots + \lfloor (k+1)/2 \rfloor$$
$$= 1 + 1 + 2 + 2 + \cdots + \lfloor k/2 \rfloor + \lfloor (k+1)/2 \rfloor.$$

If $k = 2q$, $L_b(k) = 2\{1 + 2 + \cdots + q\} = q(q+1)$. If $k = 2q+1$, $L_b(k) = q(q+1) + (q+1)$. Thus for all k

$$L_b(k) = \lfloor (k+1)/2 \rfloor \times \lceil (k+1)/2 \rceil,$$

Furthermore, when p is *even*, the number of degenerate triples in $A(p)$ is

$$f(p, p/2) = \lfloor (3p/2 - p + 2)/2 \rfloor = \lfloor (p+4)/4 \rfloor$$

and therefore

$$T_b(p+3) = T_a(p) = T_b(p) + \lfloor (p+4)/4 \rfloor.$$

4. Triangles with middle side j. For each non-negative integer j, let $M_\alpha(j)$ denote the number of triangular triples of each of the four types described in Section 1, with middle entry j. The first few values of these functions are given in Table 3.

Table 3

j	0	1	2	3	4	5
M_a	1	3	6	10	15	21
M_b	0	1	3	6	10	15
M_c	0	0	1	3	6	10
M_d	0	0	0	1	3	6

Again, from Corollary 1 we have

Theorem 3. *For all $j \geq 0$, $M_a(j) = M_b(j+1)$, $M_b(j) = M_c(j+1)$, and $M_c(j) = M_d(j+1)$.*

Therefore we can evaluate all $M_\alpha(j)$ when we have an explicit formula for $M_b(j)$. If any positive value j is given, then $(i, j, k) \in B$ if and only if $1 \leq i \leq j$ and $k \in \{j, j+1, \ldots, j+i-1\}$. Thus

$$M_b(j) = \sum_{i=1}^{j} i = (j)(j+1)/2.$$

There is no such formula, in terms of i alone, for the number of triangular triples containing i (as the smallest entry) since

$(i, n, n) \in A$ whenever $0 \leq i \leq n$,
$(i, n, n) \in B$ whenever $1 \leq i \leq n$,
$(i, n, n+1) \in C$ whenever $1 \leq i < n$, and
$(i, n, n+1) \in D$ whenever $1 < i < n$.

5. Evaluation of $T_b(p)$.

Theorem 4. *If $p = 12q + r$ where $0 \leq r < 12$ and r is even, then*

$$T_b(p) = \begin{cases} 3q^2 + (r/2)q & \text{if } r = 0, 2 \text{ or } 4 \\ 3q^2 + (r/2)q + 1 & \text{if } r = 6 \text{ or } 8 \\ 3q^2 + (r/2)q + 2 & \text{if } r = 10. \end{cases}$$

Proof: In Section 3 we saw that for non-degenerate triples $\lceil p/3 \rceil \le k \le \lfloor (p-1)/2 \rfloor$ and that the number of triples with the largest entry k is $\lfloor (3k-p+2)/2 \rfloor$. Letting $\alpha = \lceil p/3 \rceil$, $\beta = \lfloor (p-1)/2 \rfloor$ and $H = 3\alpha - p + 2$ we obtain

$$T_b(p) = \sum_{k=\alpha}^{\beta} \lfloor (3k-p+2)/2 \rfloor = \sum_{i=0}^{\beta-\alpha} \lfloor (H+3i)/2 \rfloor.$$

Let N be any integer. Then $\lfloor N/2 \rfloor + \lfloor (N+3)/2 \rfloor = N+1$ and

$$\begin{aligned}\sum_{i=0}^{2t-1} \lfloor (N+3i)/2 \rfloor &= \sum_{j=0}^{t-1} \{\lfloor (N+3(2j))/2 \rfloor + \lfloor (N+3(2j+1))/2 \rfloor\} \\ &= \sum_{j=0}^{t-1} \{1+N+3(2j)\} = (1+N)t + 6(t-1)t/2 \\ &= 3t^2 + (N-2)t.\end{aligned}$$

One other relevant, general observation is this: if $s \in \{0,2,4\}$ then $\lceil s/3 \rceil = s/2$.

We now return to the task of evaluating $T_b(p)$ when p is even and written as $12q+r$. Then β is $\lfloor (p-1)/2 \rfloor = 6q + (r/2) - 1$. If $r \in \{0,2,4\}$ then

$$\alpha = \lceil (12q+r)/3 \rceil = 4q + \lceil r/3 \rceil = 4q + (r/2); \quad \beta - \alpha = 2q - 1; \text{ and}$$

$$H = 12q + 3(r/2) - 12q - r + 2 = (r/2) + 2.$$

Thus,

$$T_b(p) = \sum_{i=0}^{2q-1} \lfloor (H+3i)/2 \rfloor = 3q^2 + (H-2)q = 3q^2 + (r/2)q.$$

If $r \in \{6,8,10\}$ then

$$\begin{aligned}\alpha &= \lceil (12q+6+r-6)/3 \rceil = 4q+2+\lceil (r-6)/3 \rceil = 4q+2+(r-6)/2 \\ &= 4q + (r/2) - 1; \quad \beta - \alpha = 2q;\end{aligned}$$

and

$$H = 12q + 3(r/2) - 3 - 12q - r + 2 = (r/2) - 1.$$

Thus,

$$\begin{aligned}T_b(p) &= \sum_{i=0}^{2q} \lfloor (H+3i)/2 \rfloor = \lfloor H/2 \rfloor + \sum_{i=0}^{2q-1} \lfloor (H+3+3i)/2 \rfloor \\ &= \lfloor H/2 \rfloor + 3q^2 + (H+3-2)q = 3q^2 + (r/2)q + \lfloor H/2 \rfloor\end{aligned}$$

Since $\lfloor H/2 \rfloor = 1$ if $r = 6$ or $r = 8$ and $\lfloor H/2 \rfloor = 2$ if $r = 10$, we have the theorem. ■

6. A Simple Unified Formula for $T_b(p)$. Various formulas for $T_b(p)$ when p is even have been obtained by other authors using different methods. Most can be derived directly from Theorem 4.

From Table 1 it is clear that $T_b(p) = 3q^2 + (r/2)q + T_b(r)$. Since $r \in \{0,2,4,6,8,10\}$ we can remove $T_b(r)$ from this and obtain $T_b(p) = 3q^2 + (r/2)q +$

$\lfloor r/5 \rfloor$. We can also remove q. Since $(p^2 - r^2)/48 = (144q^2 + 24rq)/48 = 3q^2 + (r/2)q$, $T_b(p) = (p^2 - r^2)/48 + T_b(r)$ as in [**4**]. Finally, we can even remove r.

Table 4

r	$(r^2 + 21)/48$	$T_b(r)$	$r^2/48$
0	21/48	0	0
2	25/48	0	4/48
4	37/48	0	16/48
6	57/48	1	36/48
8	85/48	1	64/48
10	121/48	2	100/48

Since $\lfloor (p^2 + 21)/48 \rfloor = 3q^2 + (r/2)q + \lfloor (r^2 + 21)/48 \rfloor$, $T_b(p) = \lfloor (p^2 + 21)/48 \rfloor$ as in [**6**].

However by far the nicest formula uses $\lfloor x \rceil$, the *nearest integer* to a real number x (when there is one). Since $\lfloor p^2/48 \rceil = \lfloor 3q^2 + (r/2)q + r^2/48 \rceil = 3q^2 + (r/2)q + \lfloor r^2/48 \rceil$, $T_b(p) = \lfloor p^2/48 \rceil$ as in [**3**].

REFERENCES

1. G. Andrews, A note on partitions and triangles with integer sides, *Amer. Math. Monthly* **86** (1979) 477.
2. D. Hale, One thing leads to another, *Mathematics Teaching* **72** (1975) 18–21.
3. R. Honsberger, *Mathematical Gems* III, vol. 9, Dolciana Mathematical Expositions, Mathematical Association of America, Washington, DC, 1985.
4. J. H. Jordan, R. Welch, and R. J, Wisner, Triangles with integer sides, *Amer. Math. Monthly* **86** (1979) 686–689.
5. H. Paddags, Triangular Triples, *Ontario Association of Mathematics Educators Gazette* **35** (1997) 27–29.
6. H. Wellstein, Eine Zahlformel fur Dreiecke, *Elem. Math.* **52** (1997) 169–172.

Brock University, St. Catharines, L28 3A1, Ontario, Canada
tjenkyns@brocku.ca, emuller@spartan.ac.brocku.ca

Normal Numbers Without Measure Theory

Rodney Nillsen

1. Introduction. Consider a given number x in $[0, 1)$. The binary expansion of this number produces a sequence of digits, each of which is zero or one. For $n = 1, 2, 3, \ldots$ let $S_n(x)$ denote the number of ones that appear among the first n digits of the binary expansion of x. Then x is said to be *simply normal to base* 2 if

$$\lim_{n \to \infty} \frac{S_n(x)}{n} = \frac{1}{2}.$$

Thus, a number is simply normal to base 2 if it has an "equal" number of zeros and ones in its binary expansion. A similar definition may be made for a number to be simply normal to other bases. The following result was proved by Émile Borel in 1909.

Borel's Theorem. *There is a subset* $\mathscr{Z}$ *of* $[0, 1)$ *that has measure zero and is such that every number in* $[0, 1)$ *that is not in* $\mathscr{Z}$ *is simply normal to base* 2. *That is, almost every number in* $[0, 1)$ *is simply normal to base* 2.

In [**2**], Marc Kac described a very elegant approach to proving Borel's Theorem, using the *Rademacher functions*. More recently, in a paper in this MONTHLY, Goodman [**1**] has shown how Kac's approach may be extended so as to obtain deeper results relating to normal numbers, including some of those obtained by Mendès France [**4**] using more difficult concepts and techniques. The approach of Kac is elementary, and is quite accesible to undergraduate students, except at one point, where it is necessary to invoke the Beppo Levi Theorem to deduce that a series of non-negative functions converges almost everywhere. The Beppo Levi Theorem is similarly invoked by Goodman [**1**], in his extension of Kac's approach to other aspects of normal numbers.

The main aim of this note is to show how to avoid using Beppo Levi's Theorem (and the associated background in measure and integration), in the approaches of Kac and Goodman, thus making results on normal numbers more accessible to undergraduate students. In fact, the intention is to make Borel's Theorem completely accessible to the student who knows what the integral of a step function is and who is familiar with convergent sequences and series.

2. Avoiding the Measure Theory. Well, measure theory cannot be completely avoided, since Borel's Theorem requires the notion of a set of measure zero. If J is an interval, let $\mu(J)$ denote its length.

Definition. A subset A of $[0, 1)$ is called a set of *measure zero* if for each $\varepsilon > 0$ there is a sequence (J_n) of intervals such that

$$A \subseteq \bigcup_{n=1}^{\infty} J_n \quad \text{and} \quad \sum_{n=1}^{\infty} \mu(J_n) < \varepsilon.$$

This definition of a set of measure zero is quite accessible, and it is easy to show that any countable set has measure zero and that any interval of positive length does not have measure zero. It can also be shown at the undergraduate level that an uncountable set such as the Cantor set is a set of measure zero. If a statement holds for all x except for those x in some set of measure zero, we say that the statement holds for *almost all* x.

Now, let's look at Kac's approach as in [**2**]. For $n = 1, 2, 3, \ldots$ the n^{th} Rademacher function r_n on $[0, 1)$ is

$$r_n(x) = \begin{cases} 1, \text{ if } x \in \left[\dfrac{(k-1)}{2^n}, \dfrac{k}{2^n}\right) & \text{and } k \text{ is odd} \\ -1, \text{ if } x \in \left[\dfrac{(k-1)}{2^n}, \dfrac{k}{2^n}\right) & \text{and } k \text{ is even.} \end{cases}$$

Equivalently, $r_n(x) = 1$ if the n^{th} binary digit of x is 0, while $r_n(x) = -1$ if the n^{th} binary digit is 1 [**2**, p. 3]. This relationship between $r_n(x)$ and the n^{th} binary digit of x implies that Borel's Theorem may be expressed equivalently in terms of the Rademacher functions as follows [**2**, pp. 16–17]:

$$\lim_{n \to \infty} \frac{1}{n}(r_1(x) + r_2(x) + \cdots + r_n(x)) = 0, \quad \text{for almost all } x \in [0, 1).$$

Then from this point, Kac's proof of Borel's Theorem is along the following lines.

(I) Any product of distinct Rademacher functions has integral zero. This property allows a direct calculation, which shows that

$$\int_0^1 \left(\frac{r_1(x) + r_2(x) + \cdots + r_n(x)}{n} \right)^4 dx = \frac{3n-2}{n^3},$$

and we deduce that

$$\sum_{n=1}^{\infty} \left(\int_0^1 \left(\frac{r_1(x) + r_2(x) + \cdots + r_n(x)}{n} \right)^4 dx \right) < \infty. \tag{1}$$

(II) Beppo Levi's Theorem means we can change the order of summation and integration in (1), to deduce that

$$\int_0^1 \left(\sum_{n=1}^{\infty} \left(\frac{r_1(x) + r_2(x) + \cdots + r_n(x)}{n} \right)^4 \right) dx < \infty.$$

It follows that

$$\sum_{n=1}^{\infty} \left(\frac{r_1(x) + r_2(x) + \cdots + r_n(x)}{n} \right)^4 < \infty, \quad \text{for almost all } x \in [0,1). \tag{2}$$

(III) Since the n^{th} term of a convergent series has limit zero, we deduce from (2) that

$$\lim_{n \to \infty} \frac{r_1(x) + r_2(x) + \cdots + r_n(x)}{n} = 0, \quad \text{for almost all } x \in [0,1). \tag{3}$$

Then, as noted above, Borel's Theorem follows.

Step (I) can be carried out at the undergraduate level, but step (II) requires a preliminary course in measure and integration theory. We now show how to replace steps (II) and (III) with a direct argument that leads to the conclusion (3).

Lemma. *Let (a_n) be a sequence of non-negative numbers such that $\sum_{n=1}^{\infty} a_n < \infty$. Then there is a sequence (b_n) of positive numbers such that*

$$\lim_{n \to \infty} b_n = \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} a_n b_n < \infty.$$

Proof: As $\sum_{n=1}^{\infty} a_n < \infty$, there is a sequence $k_1 < k_2 < k_3 < \cdots$ such that $k_1 = 0$ and

$$\sum_{n=k_j+1}^{\infty} a_n < 2^{-j}, \quad \text{for all } j = 2, 3, \ldots.$$

We define the sequence (b_n) as follows: if $n \in \mathbb{N}$, there is a unique $j \in \mathbb{N}$ with $n \in \{k_j + 1, k_j + 2, \ldots, k_{j+1}\}$, in which case we put $b_n = j$. Now, for all $j = 2, 3, 4, \ldots,$

$$\sum_{n=k_j+1}^{k_{j+1}} a_n b_n = \sum_{n=k_j+1}^{k_{j+1}} j a_n \leq j \sum_{n=k_j+1}^{\infty} a_n < j 2^{-j}.$$

Hence,

$$\sum_{n=1}^{\infty} a_n b_n = \sum_{n=1}^{k_2} a_n + \sum_{j=2}^{\infty} \left(\sum_{n=k_j+1}^{k_{j+1}} a_n b_n \right) < \sum_{n=1}^{k_2} a_n + \sum_{j=2}^{\infty} j 2^{-j} < \infty.$$

Theorem. *Let (ϕ_n) be a sequence of real or complex valued step functions on* $[0,1)$ *such that*

$$\sum_{n=1}^{\infty}\left(\int_0^1 |\phi_n(x)|\,dx\right) < \infty.$$

Then,

$$\lim_{n\to\infty} \phi_n(x) = 0, \quad \text{for almost all } x \in [0,1).$$

Proof: By the Lemma, there is a sequence (b_n) of positive numbers such that

$$\lim_{n\to\infty} b_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} b_n\left(\int_0^1 |\phi_n(x)|\,dx\right) < \infty. \tag{4}$$

Now let $\mathscr{Z}$ denote the set of all points x in $[0,1)$ such that the sequence $(\phi_n(x))$ does *not* converge to 0, and let $x \in \mathscr{Z}$. Then, by the definition of a convergent sequence, there is some $\eta > 0$ such that $|\phi_n(x)| > \eta$ for an infinite number of $n \in \mathbb{N}$. Since $\lim_{n\to\infty} b_n = \infty$, $b_n^{-1} < \eta$ for all sufficiently large n. It follows that $|\phi_n(x)| > b_n^{-1}$ for infinitely many n.

Define, for $n = 1, 2, 3, \ldots,$

$$\mathscr{A}_n = \{y : y \in [0,1) \text{ and } |\phi_n(y)| > b_n^{-1}\}.$$

We have

$$\begin{aligned} x \in \mathscr{Z} &\Rightarrow |\phi_n(x)| > b_n^{-1} \quad \text{for infinitely many } n, \\ &\Rightarrow x \in \mathscr{A}_n \quad \text{for infinitely many } n \\ &\Rightarrow x \in \bigcup_{k=n}^{\infty} \mathscr{A}_k, \quad \text{for all } n \in \mathbb{N}. \end{aligned}$$

Hence,

$$\mathscr{Z} \subseteq \bigcup_{k=n}^{\infty} \mathscr{A}_k, \quad \text{for all } n \in \mathbb{N}. \tag{5}$$

Now note that because each function ϕ_n is a step function, the set $\mathscr{A}_n$ is a finite union of disjoint intervals, $J_1, J_2, \ldots, J_r$ say. Then, let us put $\mu(\mathscr{A}_n) = \sum_{j=1}^{r} \mu(J_j)$. Also, it follows from the definition of $\mathscr{A}_n$ that

$$b_n|\phi_n(x)| \geq 1, \quad \text{for all } x \in \mathscr{A}_n.$$

Consequently,

$$\mu(\mathscr{A}_n) = \int_{\mathscr{A}_n} 1\,dx \leq \int_{\mathscr{A}_n} b_n|\phi_n(x)|\,dx \leq b_n\int_0^1 |\phi_n(x)|\,dx.$$

It now follows from (4) that

$$\sum_{n=1}^{\infty} \mu(\mathscr{A}_n) \leq \sum_{n=1}^{\infty}\left(b_n\int_0^1 |\phi_n(x)|\,dx\right) < \infty,$$

so that

$$\lim_{n\to\infty} \sum_{k=n}^{\infty} \mu(\mathscr{A}_k) = 0. \tag{6}$$

It follows from (5) and (6) that $\mathscr{Z}$ is a set of measure zero. But as $\mathscr{Z}$ is by definition the set of all points x in $[0,1)$ such that the sequence $(\phi_n(x))$ does not converge to 0, we have

$$\lim_{n\to\infty} \phi_n(x) = 0, \quad \text{for almost all } x.$$

This Theorem implies Borel's Theorem, as we can see by taking ϕ_n to be the step function $((r_1 + r_2 + \cdots + r_n)/n)^4$ and using (1) to deduce from the Theorem that

$$\lim_{n\to\infty} \frac{1}{n}(r_1(x) + r_2(x) + \cdots + r_n(x)) = 0, \quad \text{for almost all } x.$$

3. Concluding Remarks. Let $|A|$ denote the number of elements in a finite set A. Then, a number x in $[0, 1)$ is *normal to base* 2, as distinct from *simply normal to base* 2, if it has the following property: if $d_1(x), d_2(x), \ldots$ denotes the sequence of zeros and ones in the binary expansion of x, and if $b_1, b_2, \ldots, b_r$ is a finite sequence of zeros and ones, then

$$\lim_{n\to\infty} \frac{1}{n}\left|\{j : 1 \leq j \leq n \quad \text{and} \quad d_j(x) = b_1, d_{j+1}(x) = b_2, \ldots, d_{j+r-1}(x) = b_r\}\right| = \frac{1}{2^r}.$$

It is known that *almost every number in* $[0, 1)$ *is normal to base* 2, a result known as the *Normal Numbers Theorem for base* 2 [**1**]. It was Mendès France [**4**] who made a connnection between the numbers normal to base 2 and the *Walsh functions*, which are formed by taking products of the Rademacher functions. In [**1**], Goodman shows that the argument that Kac used on the Rademacher functions to prove Borel's Theorem can be used in a like manner on the Walsh functions, to deduce the Normal Numbers Theorem for base 2. Goodman's argument uses the Beppo Levi Theorem. But, just as the approach in this paper shows how to avoid measure theory in Kac's approach to Borel's Theorem, so too this approach avoids the use of measure theory in Goodman's approach to the Normal Numbers Theorem to base 2. In this sense, the Normal Numbers Theorem to base 2 is as accessible to students as Borel's Theorem. For bases other than 2, Goodman uses complex valued functions that correspond to the Rademacher functions, but the Theorem of Section 2 still applies, and the discussion in [**1**] for more general bases thus may proceed independently of measure theory.

The Theorem in Section 2 is stated in terms of step functions to emphasise that the present approach to Borel's Theorem requires a knowledge of integration extending only to step functions. In [**6**, p. 345] Weyl proves a similar result to the Theorem of Section 2, but where the functions are continuous. Weyl's proof uses measure theory, and does in fact apply to step functions, in which case his argument simplifies somewhat, becoming independent of measure theory and providing a different proof of the Theorem in Section 2. However, the proof of the Theorem in Section 2 can be modified so as to apply to continuous functions, without using measure theory. Then, for students who know about integrals of continuous functions, and who know that a 2π-periodic continuous function may be uniformly approximated by trigonometric polynomials, Weyl's criterion for uniformly distributed sequences may be proved. This means that measure theory may be avoided in discussing results on the uniform distribution of sequences where Beppo Levi's Theorem is routinely invoked, such as in [**3**, pp. 32–33]. Note also that the distinction between the concept of a set of measure zero and the general theory of measure is discussed by F. Riesz in [**5**], especially pp. 363–365. Riesz also outlines a proof of Borel's Theorem [**5**, pp. 369–370], which the interested reader may care to compare with both Kac's approach and the ideas presented here.

THE EVOLUTION OF...

Edited by **Abe Shenitzer**

Mathematics and Statistics, York University, Toronto, Ontario M3J 1P3, Canada

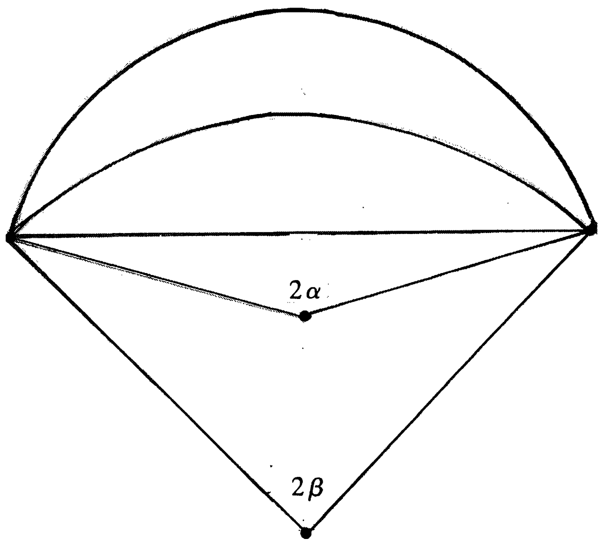
The Problem of Squarable Lunes

M. M. Postnikov

translated from the Russian by Abe Shenitzer

Translator's note. Hippocrates of Chios (second half of the fifth century BCE) seems to have been the first mathematician to square a figure with curved boundary. The figure in question was a *lune*. (For details about Hippocrates and his work see pp. 131–136 in B. L. van der Waerden, *Science Awakening*, P. Noordhoff, Groningen, 1954.)

A lune is a figure bounded by two circular arcs with a common chord. Figure 1 shows a concave-convex lune and Figure 2 shows a convex lune.



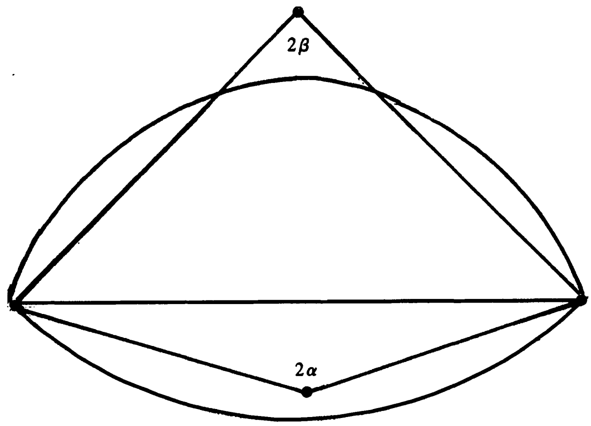
concave-convex lune

Figure 1

The problem of squarable lunes is usually referred to as *The problem of the lunes of Hippocrates*. The lunes in this problem are *constructible*, i.e., they can be constructed by ruler and compass, and *squarable*, i.e., one can construct for such a lune (by ruler and compass) a square of the same area. Using these terms we can state the problem of the lunes of Hippocrates as follows:

Find all constructible squarable lunes.

Over the ages, the problem of squarable lunes was dealt with by many mathematicians (see the index of *The Dictionary of Scientific Biography* under Quadra-



convex lune

Figure 2

ture, of lunes), but was solved only in the 20th century by two Russian algebraists. The two were Chebotaryov, who came close to the complete solution in 1934, and his student Dorodnov, who completed his teacher's work in 1947. Their principal tool was Galois theory.

The solution of the problem of squarable lunes presented in the following essay is taken from M. M. Postnikov's book *Galois Theory*, published by Fizmatgiz, Moscow, in 1963. This is an enlarged version of Postnikov's earlier book with the same title. The earlier version was translated by L. F. Boron and published by Noordhoff in 1962. It does not contain the material on lunes.

* * *

Note. In what follows, the remarks in brackets [ ] are due to the translator.)

[Figures 1 and 2 make it clear that] A lune is given by the length of the common chord that subtends its bounding arcs and by the central angles 2α and 2β that are their measures (we assume for definiteness that $\alpha > \beta$). We limit ourselves to lunes with *commensurable* angles α and β, i.e., angles for which there exists an angle θ and coprime (positive) integers m and n ($m > n$) such that $\alpha = m\theta$ and $\beta = n\theta$. With these restrictions, the construction of a lune reduces to the construction of the angle θ.

We make use of the following result: *If* $\theta \neq 0$, *then* θ *and* $\sin\theta$ *cannot both be algebraic*. [For a proof see p. 131 in I. Niven, *Irrational Numbers*, Carus Monograph No. 11, MAA, 1956.]

We consider a squarable lune with angles α and β. Without restriction of generality we may assume that the length of the chord that subtends the bounding arcs of the lune is 1. Then it is not difficult to see that the area of the lune is given by the formula

$$S = \frac{\alpha}{\sin^2\alpha} - \frac{\beta}{\sin^2\beta} + \frac{\cot\beta}{4} - \frac{\cot\alpha}{4}$$

if both bounding arcs are on the same side of the common chord (this is a so-called concave–convex lune) and by the formula

$$S = \frac{\alpha}{\sin^2\alpha} + \frac{\beta}{\sin^2\beta} + \frac{\cot\beta}{4} + \frac{\cot\alpha}{4} \tag{1}$$

otherwise (a convex lune). We consider the latter case first. Since $\alpha = m\theta$ and $\beta = n\theta$, it follows from (1) that

$$\theta = \frac{\left(S - \frac{\cot n\theta}{4} - \frac{\cot m\theta}{4}\right)\sin^2 m\theta \sin^2 n\theta}{m\sin^2 n\theta + n\sin^2 m\theta}. \qquad (*)$$

[We are dealing with constructible lunes of common chord length 1. The vertices of the central angles in Figures 1 and 2 are assumed to be constructible. Hence their distances from the common chord are algebraic. Specifically, these distances are $(\cot\alpha)/2$ and $(\cot\beta)/2$. It follows that $\sin\alpha = \sin m\theta, \sin\beta, \cot\alpha$, and $\cot\beta$ are algebraic. But then θ in $(*)$ is algebraic. Note also that if $\sin m\theta$ is algebraic then $\sin\theta$ is also algebraic.] Thus in this case θ and $\sin\theta$ are both algebraic. But this, we know, can happen only if $\theta = 0$. Hence

there are no squarable convex lunes.

A similar argument shows that

a squarable concave–convex lune with angles $\alpha = m\theta$ and $\beta = n\theta$ can exist only if

$$n\sin^2 m\theta - m\sin^2 n\theta = 0. \qquad (2)$$

Thus

the construction of squarable lunes (with commensurable angles α and β) reduces to the construction of an angle θ satisfying (2).

If in (2) we express $\sin m\theta$ and $\sin n\theta$ in terms of $\cos\theta$, then we obtain for $\cos\theta$ an equation that we must investigate. A lune with angles $m\theta$ and $n\theta$ can be constructed with ruler and compass if and only if this equation has a real root whose absolute value does not exceed 1 and whose computation reduces to the solution of a chain of quadratic equations.

Incidentally, from a computational standpoint it is more convenient to consider not the number $\cos\theta$ but the number $\xi = \cos 2\theta + i\sin 2\theta$. Of course, this change is of no major significance (ξ is constructible if and only if $\cos\theta$ is constructible). Since

$$\sin k\theta = \frac{(\xi^k - 1)^2}{4\xi^k}, \qquad k = m, n,$$

the equation for ξ has the form

$$n(x^m - 1)^2 - mx^{m-n}(x^n - 1)^2 = 0. \qquad (3)$$

Thus we have arrived at the following, purely algebraic, problem:

for what coprime positive integers m and n (with $m > n$) does the solution of (3) *reduce to the solution of quadratic equations*?

After finding all equations (3) that are reducible to quadratic equations, we can then find among them the equations that correspond to "real" lunes, i.e., equations with root ξ with absolute value 1.

It is possible to prove that

if m is composite, then, except for the case $m = 9, n = 1$, the solution of (3) *cannot be reduced to the solution of quadratic equations.*

The proof of this statement is beyond the level of this book and we leave it out.

Now let m be a prime. It turns out that

if the solution of (3) *for a prime $m = p$ reduces to the solution of quadratic equations, then $p = 2$ or p is a Fermat prime.*

To prove this proposition we make the substitution $x = y + 1$. We obtain the equation

$$n\left(\frac{(y+1)^p - 1}{y}\right)^2 = p(y+1)^{p-n}\left(\frac{(y+1)^n - 1}{n}\right)^2,$$

i.e., the equation

$$\begin{aligned} &n\left(y^{p-1} + \binom{1}{p}y^{p-2} + \binom{2}{p}y^{p-3} + \cdots + \binom{p-1}{p}\right)^2 \\ &\quad - p(y+1)^{p-n}\left(y^{n-1} + \binom{1}{n}y^{n-2} + \binom{2}{n}y^{n-3} + \cdots + \binom{n-1}{n}\right)^2 = 0. \end{aligned} \tag{4}$$

This yields the equation

$$ny^{2(p-1)} + a_1 y^{2p-3} + \cdots + a_{2p-3}y + a_{2p-2} = 0,$$

where $a_1, \dots, a_{2p-2}$ are certain integers.

Now we note that

If $k \neq 0, p$ then the binomial coefficient $\binom{k}{p}$ is divisible by p.

Indeed,

$$\binom{k}{p} = \frac{p(p-1)\cdots(p-k+1)}{1\cdot 2\cdots k},$$

and the prime p in the numerator cannot be reduced (for all factors in the denominator are less than p).

It follows that we can write (4) as

$$n\left(y^{p-1} + pf_1(y)\right)^2 + pf_2(y) = 0,$$

where $f_1(y)$ and $f_2(y)$ are certain polynomials with integer coefficients, and thus as

$$ny^{2(p-1)} + pf(y) = 0, \tag{5}$$

where $f(y)$ is a polynomial with integer coefficients. This means that all the coefficients $a_1, \dots, a_{2p-2}$ are divisible by p. Since the leading coefficient n is not divisible by p and the constant term a_{2p-2} (which, clearly, is equal to $np^2 - pn = pn(n-p)$) is not divisible by p^2, (5) satisfies the Eisenstein criterion [see A. Clark, *Elements of Abstract Algebra*, Wadsworth, 1971, pp. 86–87] and thus is irreducible (over Q). It follows that its solution is reducible to quadratic equations only if its degree $2(p-1)$ is equal to two or is a Fermat prime. This completes the proof.

It turns out that the necessary condition just established is not sufficient. Specifically, it can be shown that

if $p > 5$ then (3) (*for $m = p$*) *cannot be reduced to the solution of quadratic equations.*

We will not prove this result.

It remains to consider the cases $m = 2, m = 3, m = 5$, and the case $m = 9$ and $n = 1$.

Let $m = 2$. Then, after division by $(x-1)^2$, (3) becomes

$$x^2 + 1 = 0.$$

Hence in this case $2\theta = 90^\circ$ and we see that

A lune with angles $2\alpha = 180°$ and $2\beta = 180°$ is squarable.

This is the well-known lune of Hippocrates.

Let $m = 3$. Then $n = 1$ or $n = 2$. In the first case, after division by $(x-1)^2$, (3) becomes

$$(x^2 + x + 1)^2 - 3x^2 = 0.$$

This equation has two real roots (which are of no interest to us) and two complex roots

$$\frac{\sqrt{3} - 1}{2} \pm \sqrt{-\frac{\sqrt{3}}{2}}\,.$$

Hence $\cos 2\theta = (\sqrt{3} - 1)/2$, whence $2\theta \approx 68°.5$. It follows that

a lune with angles $\arccos(\sqrt{3} - 1)/2 \approx 68°.5$ *and* $3\arccos(\sqrt{3} - 1)/2 \approx 205°.6$ *is squarable.*

For $n = 2$ we obtain the equation

$$2(x^2 + x + 1)^2 - 3x(x+1)^2 = 0.$$

Putting $x = y^2$ we obtain an equation that splits over the field $Q(\sqrt{2}, \sqrt{3})$ into two reciprocal quartic equations that reduce to quadratic equations. [An equation $a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$ is said to be *reciprocal* if $a_i = a_{n-i}$ for all $i = 0, 1, \ldots, n$.] After carrying out the necessary computations we find that $\cos 2\theta = (\sqrt{33} - 1)/8$, whence $2\theta \approx 53°.6$. It follows that

a lune whose angles are $\arccos(\sqrt{33} - 1)/8 \approx 107°.2$ *and* $3\arccos(\sqrt{33} - 1)/8 \approx 160°.9$ *is squarable.*

These two lunes were also constructed by Hippocrates.

Lastly, let $m = 5$. Then $n = 1$, 2, 3, or 4. For $n = 1$ we obtain the equation

$$(x^4 + x^3 + x^2 + x + 1)^2 - 5x^4 = 0,$$

which splits into two reciprocal quartic equations

$$x^4 + x^3 + (1 \pm \sqrt{5})x^2 + x + 1 = 0,$$

and hence is reducible to quadratic equations. In this case $\cos 2\theta = (\sqrt{33} - 1)/8$, whence $2\theta \approx 46°.9$. It follows that

a lune with angles $\arccos(\sqrt{5 + 4\sqrt{5}} - 1)/4 \approx 46°.9$ *and* $5\arccos(\sqrt{5 + 4\sqrt{5}} - 1)/4 \approx 234°.4$ *is squarable.*

This lune was found in 1840 by Clausen.

For $n = 2$ we obtain the equation

$$2(x^4 + x^3 + x^2 + x + 1)^2 - 5x^3(x+1)^2 = 0.$$

Putting $x = y^2$ we reduce it to two reciprocal equations of degree eight

$$y^8 + y^6 + y^4 + y^2 + 1 \pm \sqrt{\frac{5}{2}}\, y^3(y^2 + 1) = 0.$$

Solving these equations by the well-known method (using the substitution $y + y^{-1} = z$) we obtain two quartic equations

$$4z^4 - 3z^2 \pm \sqrt{\frac{5}{2}}\, z + 1 = 0. \tag{6}$$

When we solve these equations by Ferrari's method [see A. Clark, op. cit., pp. 141–143] we obtain for the auxiliary unknown u the cubic equation

$$16u^3 - 24u^2 + 20u - 5 = 0. \tag{7}$$

It is easy to see that this equation is irreducible (it has no rational roots). Hence the initial equation is not reducible to quadratic equations (for the roots of (7) are rationally expressible in terms of the roots of (6)). It follows that

there is no squarable lune corresponding to the values $m = 5$ *and* $n = 2$.

For $n = 3$ we obtain the equation

$$3(x^4 + x^3 + x^2 + x + 1)^2 - 5x^2(x^2 + x + 1)^2 = 0,$$

which splits over the field $Q(\sqrt{3}, \sqrt{5})$ into two reciprocal quartic equations that reduce to quadratic equations. After carrying out the necessary computations we find that $\cos 2\theta = t$, where

$$t = \frac{\sqrt{\frac{5}{3}} - 1 + \sqrt{\frac{20}{3} + \sqrt{\frac{20}{3}}}}{4},$$

whence $2\theta \approx 33°.6$. It follows that

a lune with angles $3 \arccos t \approx 100°.8$ *and* $5 \arccos t \approx 168°.0$ *is squarable*.

This lune was also found by Clausen.

For $n = 4$ we obtain the equation

$$4(x^4 + x^3 + x^2 + x + 1)^2 - 5x(x^3 + x^2 + x + 1)^2 = 0.$$

Just as in the case $n = 2$, it is easy to show that the solution of this equation is not reducible to the solution of quadratic equations. It follows that

there is no squarable lune corresponding to the values $m = 5$ *and* $n = 4$.

The last possibility to be considered is that of $m = 9$ and $n = 1$. In this case we obtain the reducible equation

$$(x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1)^2 - 9x^8 = 0,$$

which splits into two reciprocal equations

$$\begin{cases} x^8 + x^7 + x^6 + x^5 - 2x^4 + x^3 + x^2 + x + 1 = 0, \\ x^8 + x^7 + x^6 + x^5 + 4x^4 + x^3 + x^2 + x + 1 = 0. \end{cases} \tag{8}$$

Now the substitution $x + x^{-1} = y$ yields the two quartic equations

$$y^4 + y^3 - 3y^2 - 2y - 2 = 0,$$

$$y^4 - y^3 - 3y^2 - 2y - 4 = 0.$$

The solution of the first of these equations does not reduce to the solution of quadratic equations, for the cubic equation obtained by Ferrari's method is irreducible. On the other hand, the second equation splits over $Q(i\sqrt{3})$ into the two quadratic equations

$$\begin{cases} y^2 + \eta y - 2 = 0, \\ y^2 + \bar{\eta} y - 2 = 0. \end{cases} \tag{9}$$

where $\eta = (1 + i\sqrt{3})/2$, so that its solution reduces to the solution of quadratic equations. Hence

for $m = 9$ *and* $n = 1$ *the solution of* (3) (*more correctly, of one of its irreducible factors of degree eight*) *reduces to the solution of quadratic equations.*

We will try to obtain the corresponding lune. It is easy to see that the equations in (9) have no real roots. It follows that for any root ξ of (3) that is of interest to us (i.e., for any root of the second one of the equations in (8)), the quantity $\xi + \xi^{-1}$ is not real. If a lune with angle θ would correspond to this root, then, since $\xi = \cos 2\theta + i\sin 2\theta$, the quantity $\xi + \xi^{-1} = 2\cos 2\theta$ would be real. This contradiction shows that

there is no squarable lune corresponding to the values $m = 9$ *and* $n = 1$.

The final result of our investigation can be stated in the form of the following theorem:

Squarable lunes exist only in the following five cases:

$$\begin{array}{ccccc} m = 2, & m = 3, & m = 3, & m = 5, & m = 5 \\ n = 1, & n = 1, & n = 2, & n = 1, & n = 3. \end{array}$$

The harmonic series $\sum_{n=1}^{\infty} 1/n$ is a popular example of a divergent series whose terms tend to zero. Another example is $\sum_{n=1}^{\infty} \log(1 + 1/n)$, whose partial sums are unbounded because they telescope to $\log n$:

$$\sum_{k=1}^{n-1} \log\left(1 + \frac{1}{k}\right) = \sum_{k=1}^{n-1} \log\left(\frac{k+1}{k}\right)$$
$$= \sum_{k=1}^{n-1} (\log(k+1) - \log k) = \log n.$$

This telescoping series can be used to show that the harmonic series diverges. Start with the classical inequality $x \geq \log(1 + x)$, which is valid for all $x > -1$ (compare the derivatives of both sides and integrate from 0 to x). Now put $x = 1/k$ with $k = 1, 2, 3, \ldots, n - 1$ and add the resulting inequalities, obtaining

$$\sum_{k=1}^{n-1} \frac{1}{k} \geq \sum_{k=1}^{n-1} \log\left(1 + \frac{1}{k}\right) = \log n.$$

Contributed by David M. Bradley, University of Maine, Orono, ME

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Mario Benedicty, Paul Bracken, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Glenn G. Chappell, Randall Dougherty, Roger B. Eggleton, Kevin Ford, Zachary Franco, Ira M. Gessel, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Kiran S. Kedlaya, Murray S. Klamkin, Fred Kochman, Frederick W. Luttmann, Vania Mascioni, Frank B. Miles, Richard Pfiefer, Cecil C. Rousseau, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before February 28, 2001; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10816. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.* A *Motzkin path* of length n is a lattice path from $(0,0)$ to $(n,0)$ with steps $(1,1)$, $(1,0)$, and $(1,-1)$ that never goes below the x-axis. For $n \ge 2$, show that the number of Motzkin paths of length n with no $(1,0)$ steps on the x-axis is equal to the number of Motzkin paths of length $n-1$ with at least one $(1,0)$ step on the x-axis.

10817. *Proposed by Walther Janous, Ursulinengymnasium, Innsbruck, Austria.* For $n \ge 2$, let $x_1, x_2, \ldots, x_n$ be nonnegative real numbers summing to 1. Choose $j \in \{1, 2, \ldots, n-1\}$ and a real number $\alpha \ge 1$. Prove that

$$\sum_{k=1}^{n} \frac{(x_k+1)^{2\alpha}}{(x_{k+j}+1)^{\alpha}} \ge n^{1-\alpha}(n+1)^{\alpha},$$

where subscripts are taken modulo n, and determine conditions for equality.

10818. *Proposed by Cezar Joita, State University of New York, Buffalo, NY.*
(a) Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous function such that $\lim_{x\to\infty} g(x) - x = \infty$ and such that the set $\{x : g(x) = x\}$ is finite and nonempty. Prove that if $f\colon \mathbb{R} \to \mathbb{R}$ is continuous and $f \circ g = f$, then f is constant.
(b) Suppose that $g\colon \mathbb{R} \to \mathbb{R}$ is a quadratic function such that $\{x\colon g(x) = x\}$ is empty. Find a nonconstant continuous function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f \circ g = f$.

10819. *Proposed by Olaf Krafft, Rheinisch-Westfällische Technische Hochschule, Aachen, Germany.* Let m and n be integers with $m \ge 2$ and $n \ge 1$. Show that

$$\binom{mn}{n} \ge \frac{m^{m(n-1)+1}}{(m-1)^{(m-1)(n-1)}} n^{-1/2}.$$

10820. *Proposed by M. Mirzavaziri, Ferdowsi University, Mashhad, Iran.* Let $f(m)$ be the least natural number with exactly m divisors. Find a formula for $f(m)$ in terms of $f(m/p)$, where p is the least prime divisor of m.

10821. *Proposed by Gerard J. Foschini, Bell Laboratories, Holmdel, NJ.* Find a sequence of functions $f_1, f_2, \ldots$ in $L_2[0,1]$ that satisfies the following conditions.
(1) For all $\epsilon \in (0,1)$, the space spanned by $\{f_{1,\epsilon}, f_{2,\epsilon}, \ldots\}$ is $L_2[\epsilon, 1]$, where $f_{n,\epsilon}$ is the restriction of f_n to $[\epsilon, 1]$.
(2) The space spanned by $\{f_1, f_2, \ldots\}$ has an infinite-dimensional orthogonal complement in $L_2[0,1]$.

10822. *Proposed by Jeffrey Lagarias, AT&T Laboratories, Florham Park, NJ, and Jade Vinson, Princeton University, Princeton, NJ.*
(a) Let $f(z) = 1/(2 - z^2)$. Prove that all periodic points of f are real.
(b) More generally, set $f_\lambda(z) = 1/(\lambda - z^2)$. For which positive real values of λ does f_λ have only real periodic points?

SOLUTIONS

The Asymptotics of the Birthday Problem

10665 [1998, 464]. *Proposed by Jerrold R. Griggs, University of South Carolina, Columbia, SC.* For positive integers s and t, let $P(s,t)$ denote the probability that a random function $f\colon S \to T$ is injective, where S, T are sets with $|S| = s$, $|T| = t$, and, for each $x \in S$, $f(x)$ is chosen uniformly and independently from T. For example, $P(n, 365)$ approximates the probability that, in a class with n students, no two students have the same birthday.
(a) Show that $P(s,t) \to 0$ as $s \to \infty$ if $t \sim ks$ for some constant $k > 1$.
(b) What happens to $P(s,t)$ as $s \to \infty$ if $t \sim cs^2$ for some constant $c > 0$?

Solution I by Darryl K. Nester, Bluffton College, Bluffton, OH. Since $P(s,t) = \prod_{i=0}^{s-1}(t-i)/t$, we have $-\ln P(s,t) = -\sum_{i=1}^{s-1} \ln(1 - i/t)$.
(a) Since $-\ln(1-x) > x$ for $x \in (0,1)$, we have

$$-\ln P(s,t) > \sum_{i=1}^{s-1} \frac{i}{t} = \frac{(s-1)s}{2t} \sim \frac{s-1}{2k} \qquad \text{as } s \to \infty.$$

Thus $-\ln P(s,t) \to \infty$ as $s \to \infty$, which yields $P(s,t) \to 0$.
(b) We show that $P(s,t) \to e^{-1/(2c)}$ as $s \to \infty$. Note that $-\ln(1-x) < x + x^2$ for $x \in (0, 1/2)$. Since $t \sim cs^2$, for all sufficiently large s we have $(s-1)/t < 1/2$, and thus

$$\frac{(s-1)s}{2t} = \sum_{i=1}^{s-1} \frac{i}{t} < -\ln P(s,t) < \sum_{i=1}^{s-1} \left(\frac{i}{t} + \frac{i^2}{t^2}\right) = \frac{(s-1)s}{2t} + \frac{(s-1)s(2s-1)}{6t^2}.$$

For $t \sim cs^2$, both bounds are asymptotic to $1/(2c)$.

Solution II by Sung Soo Kim, Hanyang University, Ansan, Kyunggi, Korea. If $b > a$, then $(x-b)(x+b) < (x-a)(x+a)$. Thus,

$$\left(1 - \frac{s-1}{t}\right)^{s/2} \le P(s,t) = \frac{t(t-1)\cdots(t-s+1)}{t^s} \le \left(1 - \frac{s-1}{2t}\right)^s.$$

(a) If $t \sim ks$ for some constant $k > 1$, then $0 \le P(s,t) \le (1 - 1/(4k))^s$ for all sufficiently large s, and $P(s,t) \to 0$.
(b) If $t \sim cs^2$ for some constant $c > 0$, then for all sufficiently large s,

$$\left(1 - \frac{1}{\big(c - \varepsilon(s)\big)s}\right)^{s/2} \le P(s,t) \le \left(1 - \frac{1}{2\big(c + \varepsilon(s)\big)s}\right)^s,$$

where $\varepsilon(s)$ is positive and tends to 0. Both bounds tend to $e^{-1/(2c)}$.

Editorial comment. It is proved in G. A. Heuer, Estimation in a certain probability problem, this MONTHLY **66** (1959) 704–706 that if $P(s,t)$ is fixed, then $\lim_{t\to\infty}(s^2/t) = -2\ln P(s,t)$.

As observed by John Lawrence, Michael P. Cohen, and Frank B. Miles, the bounds obtained in solving part **(b)** also imply that $P(s,t) \to 0$ when $t \sim cs^n$ with $1 < n < 2$, and $P(s,t) \to 1$ when $t \sim cs^n$ with $n > 2$.

Solved also by M. Boase (U. K.), R. J. Chapman (U. K.), M. P. Cohen, R. Ehrenborg, E. Hertz, G. A. Heuer, J. Lawrence, J. H. Lindsey II, J. Marengo, R. Martin, F. B. Miles, K. Schilling, P. Schuette, P. Simeonov, N. C. Singer, A. Stadler (Switzerland), E. I. Verriest, Anchorage Math Solutions Group, GCHQ Problems Group (U. K.), Laurel and Hardy Problems Group, NSA Problems Group, and the proposer.

A Powerful Property

10674 [1998, 560]. *Proposed by Marius Cavachi, Constanţa, Romania.* Let a, b be natural numbers greater than 1. Suppose that $b^n - 1$ divides $a^n - 1$ for every positive integer n. Prove that $a = b^k$ for some positive integer k.

Solution by the proposer. Define a sequence of polynomials with $\deg(Q_k) \le k$ by $Q_0 = -1$ and

$$Q_{k+1}(x) = b^{k+1}(x-1)Q_k(bx) - a(b^{k+1}x - 1)Q_k(x).$$

for $k \ge 0$. Observe that $Q_{k+1}(0) = (a - b^{k+1})Q_k(0)$. Iterating this and employing $Q_0 = -1$ leads to $Q_k(0) = -(a-b^k)(a-b^{k-1})\cdots(a-b)$. Assume that $a \ne b^j$ for every $j \in \mathbb{N}$, so that $Q_k(0) \ne 0$. We obtain a contradiction by identifying a k such that Q_k is identically 0.

Let $r_{0,n} = (a^n-1)/(b^n-1)$ for $n \ge 1$. By assumption, $r_{0,n}$ is an integer. For $k \ge 0$, define $r_{k+1,n}$ recursively by $r_{k+1,n} = b^{k+1}r_{k,n+1} - ar_{k,n}$. Let $p_0 = 1$, and let $p_{k+1} = a(1-b^{k+1})p_k$ for $k \ge 0$. By induction on k, it follows for $n \ge 1$ and $k \ge 0$ that

$$r_{k,n} = \frac{p_k a^n + Q_k(b^n)}{(b^{n+k}-1)(b^{n+k-1}-1)\cdots(b^n-1)}.$$

Now fix k so that $b^k < a < b^{k+1}$. Since

$$(b^n-1)(b^{n+1}-1)\cdots(b^{n+k}-1) = b^{n(k+1)}(1-b^{-n})(b-b^{-n})\cdots(b^k-b^{-n}) \ge \frac{b^{n(k+1)}}{2},$$

we see that

$$|r_n^{(k)}| \le \frac{|p_k a^n + Q_k(b^n)|}{b^{n(k+1)}/2} \le 2\left(|p_k|\left(\frac{a}{b^{k+1}}\right)^n + \frac{|Q_k(b^n)|}{(b^n)^{k+1}}\right). \qquad (*)$$

Since $a < b^{k+1}$ and $\deg(Q_k) < k+1$, the rightmost expression in $(*)$ is bounded above by 1 when n is sufficiently large. Thus $r_n^{(k)} = 0$ for all large n, since $r_n^{(k)}$ is an integer. For all large n, this yields $p_k a^n + Q_k(b^n) = 0$, and thus

$$p_k\left(\frac{a}{b^k}\right)^n + \frac{Q_k(b^n)}{(b^n)^k} = 0.$$

This forces $p_k = 0$, since otherwise the left side is unbounded as $n \to \infty$. We now conclude that $Q_k(b^n) = 0$ for all $n \in \mathbb{N}$, and thus Q_k is the zero polynomial. This contradicts $Q_k(0) \ne 0$, so there must exist some $j \in \mathbb{N}$ such that $a = b^j$.

A Characterization of Min or Max

10677 [1998, 666]. *Proposed by Harm Derksen, Massachusetts Institute of Technology, Cambridge, MA.* Let $*$ be a binary operation on the set $\mathbb{Q}$ of rational numbers that is associative and commutative and satisfies $0 * 0 = 0$ and $(a+c)*(b+c) = a*b+c$ for all $a, b, c \in \mathbb{Q}$. Prove that either $a * b = \max(a,b)$ for all $a, b \in \mathbb{Q}$, or $a * b = \min(a,b)$ for all $a, b \in \mathbb{Q}$.

Solution by Jeremy Dawson, Australian National University, Canberra, Australia. Let $r(x) = 0 * x$. We first compute

$$x * (x + y) = (0 + x) * (y + x) = 0 * y + x = x + r(y). \tag{1}$$

With $y = 0$, (1) becomes $x * x = x + r(0) = x + 0 * 0 = x$.

We next prove the *multiplicative property*: If m and n are positive integers, then $r(mx/n) = mr(x)/n$. To do this, we first show by induction on n that $0 * y * 2y * \cdots * ny = nr(y)$. This is trivial for $n = 0$, and for $n = 1$ it is the definition of $r(y)$. Setting $x = (k-1)y$ in (1) yields $(k-1)y * ky = (k-1)y + r(y)$. For the induction step, we can now compute

$$\begin{aligned}
0 * y * 2y * \cdots * ny &= 0 * (y * y) * (2y * 2y) * \cdots * \big((n-1)y * (n-1)y\big) * ny \\
&= (0 * y) * (y * 2y) * \cdots * \big((n-1)y * ny\big) \\
&= \big(0 + r(y)\big) * \big(y + r(y)\big) * \big(2y + r(y)\big) * \cdots * \big((n-1)y + r(y)\big) \\
&= \big(0 * y * 2y * \cdots * (n-1)y\big) + r(y), \\
&= (n-1)r(y) + r(y) = nr(y), \qquad (2)
\end{aligned}$$

using in the last two lines the defining distributive-like property and the induction hypothesis.

The associative, commutative, and distributive-like properties of $*$ and (1) and (2) imply that

$$\begin{aligned}
0 * y * 2y * \cdots * 2ny &= (0 * ny) * \big(y * (n+1)y\big) * \big(2y * (n+2)y\big) * \cdots * (ny * 2ny) \\
&= r(ny) * \big(y * (y + ny)\big) * \big(2y * (2y + ny)\big) * \cdots * \big(ny * (ny + ny)\big) \\
&= \big(0 + r(ny)\big) * \big(y + r(ny)\big) * \big(2y + r(ny)\big) * \cdots * \big(ny + r(ny)\big) \\
&= \big(0 * y * 2y * \cdots * ny\big) + r(ny) \\
&= nr(y) + r(ny). \qquad (3)
\end{aligned}$$

From (2) and (3), we obtain $2nr(y) = nr(y) + r(ny)$, and we conclude that $nr(y) = r(ny)$. Setting $y = x/n$ yields $nr(x/n) = r(x)$. Hence $r(mx/n) = mr(x/n) = mr(x)/n$, which completes the proof of the multiplicative property.

Next, note that $r(z) = 0 * z = (0 * 0) * z = 0 * (0 * z) = r(r(z))$, so r is the identity function on its image. Also,

$$r(z) = 0 * z = (-z + z) * (0 + z) = (-z) * 0 + z = r(-z) + z.$$

Hence if $z \neq 0$, then $r(z)$ and $r(-z)$ are distinct numbers (hence not both zero) on which r is the identity. Let y be a nonzero fixed point of r.

If $y > 0$, then we can write each positive rational x as $x = ny/m$ for some positive integers m, n. By the multiplicative property, $mx/n = y = r(y) = r(mx/n) = mr(x)/n$, and thus $r(x) = x$ for all positive x. With $r(x) = r(-x) + x$, this implies $r(z) = 0$ for all $z < 0$. Thus, $r(x) = \max(0, x)$, from which it follows that

$$x * w = x * \big(x + (w - x)\big) = x + r(w - x) = x + \max(0, w - x) = \max(x, w).$$

If $y < 0$, then an analogous argument yields $r(x) = \min(0, x)$ and $x * w = \min(x, w)$.

Editorial comment. Keith Kearnes noted that the result follows easily from Theorem 2.14 in M. Maroti, Semilattices with a group of automorphisms, *Algebra Universalis* **38** (1997) 238–265. Duane Broline showed that the result is true over the integers and the rationals but found a counterexample over $\mathbb{Q}(\sqrt{2})$; the function r in this example is not continuous. Roy Barbara showed that the result is true over the reals provided that the function r is continuous and that $r(1)$ is rational.

Solved also by R. Barbara (Lebanon), M. Bhattacharyya (India), D. M. Broline, R. J. Chapman (U. K.), D. Goldstein, K. A. Kearnes, S. S. Kim (Korea), N. Komanda, O. P. Lossers (The Netherlands), R. Martin, R. Newcomb, A. Nijenhuis, D. Niles, C. Rupert, K. Schilling, P. Sindelarova (Czech Republic), C. Stevenson, T. Trimble, GCHQ Problems Group (U. K.), and the proposer.

10686 [1998, 768]. *Proposed by C. R. Pranesachar, Indian Institute of Science, Bangalore, India.* A *cevian* of a triangle is a line segment that joins a vertex to the line containing the opposite side. An *equicevian* point of a triangle ABC is a point P (not necessarily inside the triangle) such that the cevians on the lines AP, BP, and CP have equal length. Let SBC be an equilateral triangle, and let A be chosen in the interior of SBC on the altitude dropped from S.
(**a**) Show that ABC has two equicevian points.
(**b**) Show that the common length of the cevians through either of the equicevian points is constant, independent of the choice of A.
(**c**) Show that the equicevian points divide the cevian through A in a constant ratio, independent of the choice of A.
(**d**) Find the locus of the equicevian points as A varies.
(**e**) Let S' be the reflection of S in the line BC. Show that (**a**), (**b**), and (**c**) hold if A moves on any ellipse with S and S' as its foci. Find the locus of the equicevian points as A varies on the ellipse.

Composite solution by the proposer and the editors. We solve part (**e**). Choose a coordinate system so that $B = (-1, 0)$, $C = (1, 0)$, $S = (0, \sqrt{3})$, and $S' = (0, -\sqrt{3})$. Let the ellipse have vertices $(0, \pm k)$ with $k > \sqrt{3}$, so its equation is $x^2/(k^2-3) + y^2/k^2 = 1$. If we write $r = \sqrt{k^2-3}$, then the ellipse is parameterized by $A = (r\cos\theta, k\sin\theta)$. When $k = \sqrt{3}$ and $r = 0$, this is a parameterization for the line segment SS', so the argument that follows proves parts (**a**) through (**d**) as a special case.

Let $A = (r\cos\theta, k\sin\theta)$ with θ not a multiple of π. Let

$$\begin{aligned} D &= \big((k+r)\cos\theta, 0\big),\\ E &= \left(\frac{(rk+2k^2-3)\cos\theta + k - r}{2k+r+3\cos\theta}, \frac{k(k+2r)\sin\theta}{2k+r+3\cos\theta}\right),\\ F &= \left(\frac{(rk+2k^2-3)\cos\theta - k + r}{2k+r-3\cos\theta}, \frac{k(k+2r)\sin\theta}{2k+r-3\cos\theta}\right),\\ \text{and}\quad P &= \left(\frac{2k+r}{3}\cos\theta, \frac{k+2r}{3}\sin\theta\right). \end{aligned}$$

The denominators are never zero, since $r \geq 0$ and $k \geq \sqrt{3}$. Using $r^2 = k^2 - 3$, we compute

$$\begin{aligned} P &= \frac{k+2r}{3k}A + \frac{2k-2r}{3k}D = \frac{k-r-3\cos\theta}{3k}B + \frac{2k+r+3\cos\theta}{3k}E\\ &= \frac{k-r+3\cos\theta}{3k}C + \frac{2k+r-3\cos\theta}{3k}F. \end{aligned}$$

Also, D is on line BC, E is on line CA, F is on line AB, and $AD = BE = CF = k$. So P is an equicevian point, and $AP/PD = (2k-2r)/(k+2r)$ is independent of A.

Similarly, let

$$\begin{aligned} D' &= \big((r-k)\cos\theta, 0\big),\\ E' &= \left(\frac{(2k^2-rk-3)\cos\theta - k - r}{r-2k+3\cos\theta}, \frac{k(2r-k)\sin\theta}{r-2k+3\cos\theta}\right),\\ F' &= \left(\frac{(2k^2-rk-3)\cos\theta + k + r}{r-2k-3\cos\theta}, \frac{k(2r-k)\sin\theta}{r-2k-3\cos\theta}\right), \qquad (*)\\ \text{and}\quad P' &= \left(\frac{r-2k}{3}\cos\theta, \frac{k-2r}{3}\sin\theta\right). \end{aligned}$$

The denominators $r - 2k \pm 3\cos\theta$ vanish only when $k = 2$, $r = 1$, and θ is a multiple of π, but this was ruled out. It may be seen that P' is an equicevian point, $AD' = BE' = CF' = k$, and $AP'/P'D' = (2k + 2r)/(k - 2r)$, which is also independent of A. When $r \neq 0$, the two points P and P' are distinct, since $\sin\theta \neq 0$. When $r = 0$, the two points P and P' are distinct provided that $\cos\theta \neq 0$; this occurs only when $A = S$, which is ruled out by the problem statement. Since $\left((r \pm 2k)/3\right)^2 - \left((k \pm 2r)/3\right)^2 = 1$, the points P and P' trace confocal ellipses with foci at B and C.

Editorial comment. Are the points P and P' the only equicevian points? In one sense, no. There can be more than two equicevian points if we allow an equicevian point to be on a side of the triangle, in which case two of the cevians through this point coincide. For example, if P is on line BC such that $AP = BC$, then P may be considered to be an equicevian point. Possibly as many as 6 points of this type exist. In fact, when $k = 2$ and $r = 1$, formulas $(*)$ yield $E' = C$, $F' = B$, and $D' = P'$, so P' lies on BC and is of this degenerate type. An additional calculation shows, however, that there are no nondegenerate equicevian points other than P and P'.

The GCHQ Problems Group (U. K.) also solved parts (a) through (d).

When Quadrilaterals Have Inscribed Circles

10698 [1998, 995]. *Proposed by Wu Wei Chao, Guang Zhou Normal University, Guang Zhou City, China.* Let P be the intersection of the two diagonals of a convex quadrilateral $ABCD$. Let the radii of the circles inscribed in the four triangles APB, BPC, CPD, and DPA be r_1, r_2, r_3, and r_4, respectively. Show that $ABCD$ has an inscribed circle if and only if

$$\frac{1}{r_1} + \frac{1}{r_3} = \frac{1}{r_2} + \frac{1}{r_4}.$$

Solution by Plamen Simeonov, Ohio State University, Columbus, OH. Let $a = AB, b = BC$, $c = CD$, $d = DA$, $p_1 = PA$, $p_2 = PB$, $p_3 = PC$, $p_4 = PD$, and $\varphi = \angle APB$. We use the fact that $ABCD$ has an inscribed circle if and only if $a + c = b + d$.

The area K of triangle APB can be expressed in two ways:

$$\frac{1}{2}p_1 p_1 \sin\varphi = K = \frac{1}{2}(a + p_1 + p_2)r_1.$$

Therefore $(1/r_1)\sin\varphi = a/(p_1p_2) + 1/p_1 + 1/p_2$. Similarly, we deduce $(1/r_2)\sin\varphi = b/(p_2p_3) + 1/p_2 + 1/p_3$, $(1/r_3)\sin\varphi = c/(p_3p_4) + 1/p_3 + 1/p_4$, and $(1/r_4)\sin\varphi = d/(p_4p_1) + 1/p_4 + 1/p_1$. Thus,

$$\left(\frac{1}{r_1} + \frac{1}{r_3} - \frac{1}{r_2} - \frac{1}{r_4}\right)\sin\varphi = \frac{ap_3p_4 + cp_1p_2 - bp_4p_1 - dp_2p_3}{p_1p_2p_3p_4}$$
$$= \frac{p_1p_2p_3p_4\left(\left(\frac{a}{p_1p_2} + \frac{c}{p_3p_4}\right)^2 - \left(\frac{b}{p_2p_3} + \frac{d}{p_4p_1}\right)^2\right)}{ap_3p_4 + cp_1p_2 + bp_4p_1 + dp_2p_3}.$$

Write $s = 1/p_1^2 + 1/p_2^2 + 1/p_3^2 + 1/p_4^2$ and apply the law of cosines to each triangle to obtain

$$p_1p_2p_3p_4\left(\left(\frac{a}{p_1p_2} + \frac{c}{p_3p_4}\right)^2 - \left(\frac{b}{p_2p_3} + \frac{d}{p_4p_1}\right)^2\right)$$
$$= p_1p_2p_3p_4\left(s - 2\left(\frac{1}{p_1p_2} + \frac{1}{p_3p_4}\right)\cos\varphi + \frac{2ac}{p_1p_2p_3p_4}\right.$$
$$\left. - s - 2\left(\frac{1}{p_2p_3} + \frac{1}{p_4p_1}\right)\cos\varphi - \frac{2bd}{p_1p_2p_3p_4}\right)$$

$$= \left(p_1^2 + p_2^2 + p_3^2 + p_4^2 - 2p_3p_4\cos\varphi - 2p_1p_2\cos\varphi + 2ac\right) \\ - \left(p_1^2 + p_2^2 + p_3^2 + p_4^2 + 2p_4p_1\cos\varphi + 2p_2p_3\cos\varphi + 2bd\right) \\ = (a+c)^2 - (b+d)^2.$$

Therefore

$$\left(\frac{1}{r_1} + \frac{1}{r_3} - \frac{1}{r_2} - \frac{1}{r_4}\right)\sin\varphi = \frac{(a+c)^2 - (b+d)^2}{ap_3p_4 + cp_1p_2 + bp_4p_1 + dp_2p_3}.$$

But $\sin\varphi \neq 0$, so $a + c = b + d$ if and only if $1/r_1 + 1/r_3 = 1/r_2 + 1/r_4$.

Solved also by J. Anglesio (France), R. J. Chapman (U. K.), M. Benedicty, M. S. Klamkin (Canada), N. Komanda, J. Manoharmayum (U. K.), M. Reid, T. Trif (Romania), and the proposer.

Beyond the De Longchamps Point

10734 [1999, 470]. *Proposed by Floor van Lamoen, Goes, The Netherlands.* Let ABC be a triangle with orthocenter H, incenter I, and circumcenter O. Let $[P, r]$ denote the circle with center P and radius r. Show that the radical center of $[A, CA + AB]$, $[B, AB + BC]$, and $[C, BC + CA]$ is the point obtained by reflecting H through O and then reflecting the result through I.

Solution by J. C. Binz, University of Bern, Bern, Switzerland. We may select the coordinate system so that $A = (0, 0)$, $B = (\cos\alpha, \sin\alpha)$, and $C = (b, 0)$, with $0 < \alpha < \pi$ and $b > 0$. As usual, let a, b, c denote the lengths of the sides opposite A, B, C, respectively. Then $c = 1$, $a^2 = b^2 + 1 - 2b\cos\alpha$, and the semiperimeter $s = (a + b + 1)/2$. Hence $I = \big(s - a, b\sin\alpha/(2s)\big)$, $O = \big(b/2, (1 - b\cos\alpha)/(2\sin\alpha)\big)$, and $H = \big(\cos\alpha, (b - \cos\alpha)\cos\alpha/\sin\alpha\big)$.

The image of H after the first reflection is $2O - H$, the De Longchamps point. The image H' after the second reflection is $H' = 2I - 2O + H$. Then

$$H' = (x', y') = \big(1 - a + \cos\alpha, (b\sin\alpha)/s + (2b\cos\alpha - \cos^2\alpha - 1)/\sin\alpha\big).$$

The three circles are given by the equations

$$x^2 + y^2 = (b + 1)^2, \tag{1}$$

$$(x - b)^2 + y^2 = (a + b)^2, \tag{2}$$

and

$$(x - \cos\alpha)^2 + (y - \sin\alpha)^2 = (a + 1)^2. \tag{3}$$

Subtracting (2) from (1) and (3) from (1) yields two of the radical axes, namely

$$2bx - b^2 = 2b + 1 - a^2 - 2ab = 2b - 2ab + 2b\cos\alpha - b^2$$

and

$$2x\cos\alpha - \cos^2\alpha + 2y\sin\alpha - \sin^2\alpha = b^2 + 2b - a^2 - 2a.$$

These equations simplify to $x = 1 - a + \cos\alpha$ and $x\cos\alpha + y\sin\alpha = b\cos\alpha + b - a$. The radical center H'' of the three circles is the intersection of these two lines, so $H'' = (x'', y'') = \big(1 - a + \cos\alpha, \big(b - a + (b + a - 1)\cos\alpha - \cos^2\alpha\big)/\sin\alpha\big)$.

Clearly $x' = x''$. Since

$$\frac{b\sin\alpha}{s} = \frac{2b\sin\alpha}{a+b+1} = \frac{2b(b+1-a)\sin\alpha}{(b+1)^2 - a^2} = \frac{(b+1-a)\sin\alpha}{1+\cos\alpha} \\ = \frac{(b+1-a)(1-\cos\alpha)}{\sin\alpha},$$

we have

$$y' = \frac{(b+1-a)(1-\cos\alpha) + 2b\cos\alpha - \cos^2\alpha - 1}{\sin\alpha}$$
$$= \frac{b - a + (b+a-1)\cos\alpha - \cos^2\alpha}{\sin\alpha} = y''.$$

Thus $H' = H''$.

Solved also by J. Anglesio (France), M. Bataille (France), S. B. Ekhad, J. Fukuta (Japan), N. Lakshamanan, V. Schindler (Germany), P. Y. Woo (China), R. L. Young, GCHQ Problems Group (U. K.), and the proposer.

Moments of the Poisson Distribution

10738 [1999, 471]. *Proposed by Radu Theodorescu, Université Laval, Sainte-Foy, PQ, Canada.* For $t > 0$, let $m_n(t) = \sum_{k=0}^{\infty} k^n e^{-t} t^k / k!$ be the nth moment of a Poisson distribution with parameter t. Let $c_n(t) = m_n(t)/n!$. A sequence $a_0, a_1, \ldots$ is *log-convex* if $a_{n+1}^2 \le a_n a_{n+2}$ for all $n > 0$ and is *log-concave* if $a_{n+1}^2 \ge a_n a_{n+2}$ for all $n > 0$.
(**a**) Show that $m_0(t), m_1(t), \ldots$ is log-convex.
(**b**) Show that $c_0(t), c_1(t), \ldots$ is not log-concave when $t < 1$.
(**c**) Show that $c_0(t), c_1(t), \ldots$ is log-concave when t is sufficiently large.
(**d**)* Is $c_0(t), c_1(t), \ldots$ is log-concave when $t \ge 1$?

Solution by Jonathan M. Borwein, Simon Fraser University, Burnaby, BC, Canada.
(**a**) After we cancel a factor e^{-2t}, the log-convexity condition becomes

$$\sum_{k,j\ge 0} \frac{t^{k+j}(jk)^{n+1}}{k!j!} \le \sum_{k,j\ge 0} \frac{t^{k+j}(jk)^n}{k!j!} k^2 = \sum_{k,j\ge 0} \frac{t^{k+j}(jk)^n}{k!j!} \frac{k^2+j^2}{2},$$

and this follows from $2jk \le k^2 + j^2$.
(**b**) A computation reveals that $m_0(t) = 1$, $m_1(t) = t$, and $m_2(t) = t(1+t)$. Since $t < 1$, we have $c_0(t)c_2(t) = (t(1+t))/2 < t^2 = c_1(t)^2$.
(**c, d**) The generating function of $c_n(t)$ is

$$\sum_{n\ge 0} c_n(t)\, u^n = \sum_{n\ge 0}\sum_{k\ge 0} \frac{k^n e^{-t} t^k u^n}{n!k!} = e^{-t} \sum_{k\ge 0} \frac{t^k}{k!} \sum_{n\ge 0} \frac{(uk)^n}{n!} = e^{t(e^u - 1)}.$$

Note that $t(e^u - 1) = \sum_{j\ge 1} b_j u^j / j$, where $b_j = t/(j-1)!$ for $j \ge 1$. Because $t \ge 1$, the sequence $1, b_1, b_2, \ldots$ is log-concave. The log-concavity of $c_n(t)$ now follows from a theorem of E. Rodney Canfield, Engel's Inequality for Bell Numbers, *J. Comb. Theory*, Series A, **72** (1995) 184–187: If a sequence $1, b_1, b_2, \ldots$ is nonnegative and log-concave, then so is the sequence $1, c_1, c_2, \ldots$ determined by the generating function equation

$$\sum_{n\ge 0} c_n u^n = \exp\left(\sum_{j\ge 1} \frac{b_j u^j}{j}\right).$$

Editorial comment. Borwein speculates that the case $n = 1$ used in (**b**) may be the *only* violation of log-concavity of the sequence $c_n(t)$ for $t > 0$.

Parts (**a**), (**b**), and (**c**) were solved also by the proposer.

A Definite Integral

10746 [1999, 685]. *Proposed by Stepan Tersian, University of Rousse, Rousse, Bulgaria.* Prove that

$$\int_0^{\infty} \left(e^{-y\sqrt{(s/x)^2+1}} - e^{-x\sqrt{(s/y)^2+1}}\right) \cos s\, ds = 0.$$

for all positive real numbers x and y.

Solution by Richard Bagby, New Mexico State University, Las Cruces, NM. Let $I(x, y) = \int_0^\infty e^{-y\sqrt{(s/x)^2+1}} \cos s\, ds$. Comparison with $\int_0^\infty e^{-ys/x}\, ds$ shows that the integral converges.

Using $\cos s = (e^{is} + e^{-is})/2$, we obtain

$$I(x, y) = (1/2) \int_{-\infty}^{\infty} \exp\big(-y\sqrt{(s/x)^2 + 1} + is\big)\, ds.$$

The substitution $s = x(r - r^{-1})/2$ yields

$$I(x, y) = \frac{x}{4} \int_0^\infty \exp\Big(-y(r + r^{-1})/2 + ix(r - r^{-1})/2\Big)(1 + r^{-2})\, dr. \qquad (1)$$

Let $F(z) = \exp\big(-y(z + z^{-1})/2 + ix(z - z^{-1})/2\big)$ for complex z. Let $0 < r_1 < r_2 < \infty$, and let C be the closed contour in the complex plane consisting of the line segment from r_1 to r_2, the quarter circle centered at 0 from r_2 to ir_2, the line segment from ir_2 to ir_1, and the quarter circle centered at 0 from ir_1 to r_1. The function $F(z)(1 + z^{-2})$ is analytic everywhere in the complex plane except at $z = 0$, so

$$\oint_C F(z)(1 + z^{-2})\, dz = 0. \qquad (2)$$

For $r > 0$ and $0 \le \theta \le \pi/2$ we have $|F(re^{i\theta})| = \exp\big(-(x \sin\theta + y\cos\theta)(r + r^{-1})/2\big)$. If $c = \min\{x/\sqrt{2}, y/\sqrt{2}\}$, then $c > 0$ and $x\cos\theta + y\sin\theta \ge c$ for $0 \le \theta \le \pi/2$, so that $|F(re^{i\theta})| \le \exp\big(-c(r + r^{-1})/2\big)$. Therefore

$$\int_0^{\pi/2} F(re^{i\theta})\left(1 + r^{-2}e^{-2i\theta}\right) ire^{i\theta}\, d\theta \to 0$$

both as $r \to 0+$ and as $r \to \infty$. When we let $r_1 \to 0+$ and $r_2 \to \infty$ in (2), we get $\int_0^\infty F(r)(1 + r^{-2})\, dr = i\int_0^\infty F(ir)(1 - r^{-2})\, dr$, and thus

$$I(x, y) = \frac{ix}{4} \int_0^\infty F(ir)\left(1 - r^{-2}\right) dr.$$

Next we use what amounts to an integration by parts. Note that

$$F(ir) = \exp\Big(-iy(r - r^{-1})/2 - x(r + r^{-1})/2\Big),$$

and so we obtain $(d/dr)F(ir) = -(iy/2)F(ir)(1 + r^{-2}) - (x/2)F(ir)(1 - r^{-2})$. Since $\lim_{r\to 0+} F(ir) = 0 = \lim_{r\to\infty} F(ir)$, we have

$$\frac{x}{2}\int_0^\infty F(ir)\left(1 - r^{-2}\right) dr = \frac{-iy}{2}\int_0^\infty F(ir)\left(1 + r^{-2}\right) dr.$$

Hence

$$\begin{aligned} I(x, y) &= \frac{i}{2}\left(\frac{-iy}{2}\int_0^\infty F(ir)\left(1 + r^{-2}\right) dr\right) \\ &= \frac{y}{4}\int_0^\infty \exp\Big(-iy(r - r^{-1})/2 - x(r + r^{-1})/2\Big)\left(1 + r^{-2}\right) dr. \end{aligned}$$

This is the integral in (1) with x and y interchanged, so $I(x, y) = I(y, x)$, and the required result follows.

Editorial comment. Several solvers point out that the result may be deduced easily from results found in standard tables of integrals.

Solved also by J. Anglesio (France), H. J. Barten, D. Beckwith, O. Ciaurri (Spain), M. L. Glasser, G. L. Isaacs, O. P. Lossers (The Netherlands), A. Stadler (Switzerland), N. S. Thornber, GCHQ Problems Group (U. K.), and the proposer.

REVIEWS

Edited by **Gerald B. Folland**

Mathematics Department, University of Washington, Seattle, WA 98195-4350

Life by the Numbers. By Keith Devlin. Wiley, 1998, ix + 214 pp., $29.95.

Reviewed by **Colm Mulcahy**

Are there any mathematicians in the USA who haven't been asked what they do, only to have the answer greeted with: "Oh, math was always my worst subject in school"? A highly educated society whose members can still take pride in trotting out that old cliché is going to be somewhat resistant to attempts to communicate mathematics to it. The problem may be the huge gap between the way we mathematics teachers see the subject and the way it is applied in the real world. Indeed, not only is the general public more isolated than ever from the mathematics we know and love, but many of us are arguably less and less aware of the reach of applied mathematics today. Keith Devlin's *Life by the Numbers*, like the seven-part Public Television series that it echoes (produced by the appositely named WQED Pittsburgh), provides a fresh look at the amazingly pervasive role of mathematics in the world around us, without getting bogged down in equations or abstract concepts.

"What is mathematics and what is it good for?" is a simple enough question, and one that MONTHLY readers can no doubt answer to their own satisfaction with little hesitation, but what does the general public think? I put this question to a heavily biased sample of people in a thoroughly unscientific email survey. Here is a highly non-random selection of the responses.

- "Mathematics is language everyone can understand (with help from a dedicated interpreter). It quantifies and specifies things. It helps me get a good deal at the grocery store. It helps with gas mileage... wow! I can use it in real life! Things like the satellite spin-out remind me of the complicated uses of math, uses I will never demonstrate personally, but uses I can respect." (Colleague in English Department)
- "Mathematics is choreographic notation for the dance of nature. It helps us understand what happened and predict what may happen." (Campus IT director, Albany, NY)
- "Mathematics allows us to describe the world with total accuracy." (Orthodontist, Ireland)
- "Mathematics is a beautiful, logical, precise, analytical mental activity that tries to understand the beautiful, logical, precise, analytical principles governing the universe. It is good for training the mind to think clearly, and is essential to all the scientific progress we take for granted." (Law librarian, Philadelphia, PA)

One respondent cut right to the chase with:

- "I used to have a math teacher who introduced us to Calculus by saying 'If this stuff didn't work or wasn't useful, Neil Armstrong would never have set foot on the moon.'" (Communications Engineer, Houston, TX)

Needless to say, the correspondents' knowledge that I am a mathematician (whatever that is) probably colored these answers. People whose jobs bring them into regular contact with mathematics had views closer to those of the profession itself:

- "Mathematics is a language designed mainly for efficient expression of logic and precision." (Electrical engineer, Lawrence Livermore National Laboratory, CA)
- "The study of patterns and structures and good for training the mind to think." (Senior Mathematics Editor at publishing house, NJ)
- "Mathematics is the use of simple arithmetic (addition, subtraction, division and multiplication) in a more advanced form adding symbols and words to do computations that prove or disprove any numerical (or any because you usually have to count something in one way or another) theory." (Mathematics Department Administrative Assistant, Atlanta)
- "Mathematics is a set of completely theoretic definitions which is completely abstract from reality. Save for the simplest arithmetic you cannot check with reality to see if your theory is right or not. Nevertheless, despite this complete abstractness, mathematics has proven to be very useful to solve real-world problems in various fields." (Programmer, Sweden)

While these responses would likely please many of us if they came from our students, the same cannot be said for this one, from somebody who clearly has a less rosy view of the subject:

- "What is it?—self indulgence??? What is it good for?—self delusion??? Read Dennet's *Darwin's Dangerous Idea* for an exposé of poor theoretical mathematical thinking about the overall universal picture or supposed mathematical limits on understanding human development." (Academic environmental scientist, Ireland)

Life by the Numbers is about how mathematics is used all around us—often in conjunction with physics, engineering, and computer science—in science, biology, medicine, industry, commerce, sports, art, and entertainment. Author Keith Devlin was a consultant and contributor to the television series, and here he adopts the same fresh approach to a myriad of manifestations of applied mathematics often unfamiliar to the very people who are entrusted with getting students to learn mathematics: those of us who teach the subject. That's a good enough reason for us to want to watch the series (available on video) and to read the book; we mathematicians may be quite convinced that if anybody knows what mathematics is and what it's good for, we do, but perhaps we should not be so sure!

Putting aside any reservations about the quality of some of the applied mathematics depicted here, or feelings that applications not founded on firm theoretical bases are suspect by definition, one cannot ignore the main message: mathematics permeates virtually all of our lives, and people can be motivated to learn mathematics via something they find interesting, be that special effects in movies and videos, the miracle of the Internet, the history of art, the wonders of cosmology, the pitfalls of gambling, sports analysis, map making, flight simulation, national surveys, wearable computers, modeling international economies, DNA, life insurance, playing chess on the surface of a doughnut, or estimating the chances of being attacked by giant locusts! The inescapable conclusion, which is obvious to those with training in the field but which still surprises most otherwise well-educated people in the world, is that *mathematics is not dead*! For instance, it is the invisible fuel that drives much of this information age: without recent advances in the application of mathematics, there would be no fax, no television, no compact discs, no fast cheap telephone network encircling the globe, no email, and no MAA Online.

While some real live mathematicians (Tom Banchoff, Nate Dean, Bill Massey, Bob Osserman, De Witt Summers, Jeff Weeks) make impressive cameos here, and a few long dead mathematicians are quoted ("It is true that a mathematician who is not somewhat of a poet will never be a perfect mathematician", Weierstrass in 1902), overall Devlin lets an array of today's applied practitioners take center stage. As in the television series, the "geek factor" is downplayed, and the human element and excitement quotient are successfully brought to the fore.

Tucked away at the end of the sixth program of the television series is this wonderful observation from Devlin for those who fear complexity in mathematics: "To most people, mathematics makes the world more complicated, *it doesn't do that*! Math makes the world more simple! Mathematicians are simplistic creatures. We look at the world in the simplest possible way. We look at it in such a simple way that the only way of capturing the simplicity is with symbols, lines, nodes, edges of graphs. We strip away the complexity!" Similar sentiments are expressed early on in the book, including this: "The essence of the mathematician's approach, and the key to its incredible power and success, lies in the extreme simplicity and highly abstract nature of the patterns singled out for study. Mathematicians take a highly simplistic view of the world, avoiding the complexities of nature and of life. The mathematician views the world in terms of perfectly straight lines, perfect circles . . . " (p. 18).

The most tongue-in-check email response to my informal survey came from a Grammy-winning composer (and one-time lawyer) in Ireland, who unwittingly came close to agreeing with Devlin and inadvertently echoed the call letters of the TV station that made the series on which the book is based: "Mathematics = adding, Mathematics = subtracting. Therefore, Adding = subtracting. QED." Mathematics adds meaning and clarity to the world by subtracting the inessentials, a lesson our students can learn a lot from. Devlin's book adds meaning and clarity to the world of applied mathematics by subtracting the inessentials, and we can all learn a lot from it.

Spelman College, Atlanta, GA 30314
colm@spelman.edu

Small Worlds: The Dynamics of Networks between Order and Randomness. By Duncan J. Watts. Princeton University Press, 1999, xvi + 262 pp., $39.50.

Reviewed by **Jerrold W. Grossman**

Networks are in! From the most obvious example that has forever changed our access to information—the Internet—to the pop cultural folklore that you can reach anyone in the world through a short string of acquaintances ("six degrees of separation"), networks have been forced into our thoughts, often without our realizing that they involve nontrivial mathematics. Even casual users of personal computers understand the tree structure in the organization of their files, folders, and directories; and the ease of clicking to new sites on the World Wide Web or hitting the "Back" button to retreat has provided a gentler introduction to discrete mathematics for the masses than we might have imagined possible.

Clearly the study of networks has assumed a new importance. In mathematics, of course, graph theory has been around for at least 264 years, ever since Leonhard Euler showed that the citizens of Königsberg couldn't traverse their seven bridges without backtracking. (We use the terms *graph* and *network* interchangeably.) Some early progress in graph enumeration grew out of chemists' desires to classify molecular structures, and biologists use coupled dynamical systems to study everything from neural networks to the synchronous chirping of crickets. Computer scientists and communications researchers have tackled complex problems in graph theory for decades, and discrete mathematical models have become firmly entrenched in the social sciences, for example in mapping the spread of infectious diseases. In this climate it is not surprising, therefore, that we find the book under review, which grew out of the author's recent PhD dissertation in Theoretical and Applied Mechanics at Cornell University, pitched to a lay audience. Glimpses of the ideas contained here have previously appeared in the journal *Nature* ([**4**], [**7**]) and attracted considerable media attention ([**2**], [**6**]).

Let's try to model the most basic of social structures—the acquaintanceship relation among people—with a graph: vertices represent people, and two vertices are joined by an edge if the two people know each other, say on a first-name basis. This network is quite sparse; that is, each of the $n \approx 6 \times 10^9$ vertices is joined to just a handful of others, say k on the average. Here k is called the *average degree* of each vertex and is surely orders of magnitude smaller than n. Note that of the $C(n, 2) = n(n-1)/2$ possible edges in this graph, only $kn/2$ are present (we have to divide by 2 because each edge has two endpoints).

At least two aspects of our network deserve attention: clustering, the propensity of the edges to occur in dense subgraphs, which is a local property; and the lengths of paths between pairs of vertices, which is a global property. Because of the nature of social interactions, one should expect this graph to be rather highly clustered—many of your friends tend to know one another. One might also expect most paths in this graph to be fairly long. Psychologist Stanley Milgram in the 1960s studied path lengths by having random people in Kansas try to get a package delivered to random people in Massachusetts by sending it through a sequence of acquaintances. Surprisingly, Milgram found that the median path length was just six or seven edges.

Now let's see whether we can construct a graph abstractly that will reflect this situation: fairly high clustering and remarkably short paths. Suppose that we want to build a graph with $n = 1000$ vertices and 5000 edges (no loops or parallel edges are allowed); on average, then, every vertex is adjacent to $k = 10$ others (these are the parameters that Watts uses in most of his simulations). We could arrange the vertices in a circle and put an edge between every vertex and its five nearest neighbors on each side. The result, which graph theorists would call a *power of a cycle* and denote by C_{1000}^5, is highly structured. Because of the "local" nature of this construction, the vertices adjacent to a given vertex have many edges connecting them; in fact it's easy to count that of the $C(10,2) = 45$ possible edges that could join the vertices in each such neighborhood, 30 of them are in our graph. Thus the *clustering coefficient* γ is $2/3$. (This isn't the only way to achieve high clustering, of course, and Watts gives several other constructions, in which γ is even closer to 1.) On the other hand, vertices in our graph are rather far apart, about a quarter of the circle away on average ($246\frac{10}{37}$ edges to be precise); thus what Watts calls the *(characteristic) path length* L is very large. This model achieves the high γ we were seeking, but it fails to give the desired low L.

Well, what about a random graph? We can start with the cycle C_{1000}—a circle of 1000 vertices, each joined to its nearest neighbor on each side—and then add 4000 more edges at random, taking care not to double up any edges. The result is a "random" graph with the desired parameters. (We start with the cycle for technical reasons, simply to ensure that the resulting random graph is connected.) Random graphs have been extensively studied for decades, and it is known that processes such as this produce graphs with very short paths. Both theory and Watts' empirical work show that almost surely we get an average path length of $L \approx 3$ in this case. On the other hand, we added only 4000 of the 498,500 possible new edges, so each edge has only a 0.008 chance of occurring. Thus we'd expect to find almost none of the edges joining vertices in the neighborhood of a typical vertex, so our graph has a very low average clustering coefficient. Thus this graph, too, poorly models acquaintanceship.

These failed attempts could be considered two extremes in a continuum of possibilities, from the highly structured to the highly random. Watts developed a mechanism—several mechanisms, actually—to interpolate between them. Here is his simplest one. Let $0 \leq \beta \leq 1$. Start with the highly structured $C_n^{k/2}$ from our first model. Then go through the edges one by one in a systematic way around the circle and, with probability β, "rewire" each edge by reassigning one of its endpoints to a randomly chosen vertex anywhere in the graph; with probability $1 - \beta$ the edge is left unchanged. Clearly if $\beta = 0$ then the result is the original graph, but if $\beta = 1$ then we have in effect a random graph with the same n and k. It is not surprising that once a few edges have been rewired, the path length is lowered dramatically, since these new edges provide significant shortcuts. Thus even small nonzero values of β should help achieve our goal. (As the author points out, this basic idea was anticipated in a 1988 paper **[3]** by Béla Bollobás and Fan Chung, which analyzed the expected maximum path length for a cycle plus a random matching, giving each vertex degree 3.) On the other hand, if β isn't very large, then the few rewirings do not lower γ significantly. Watts found through extensive simulations that setting $\beta \approx 0.1$ brings L down to about 5 while keeping γ fairly high. This hybrid model does the trick.

So there do exist networks that have the low characteristic path lengths of random graphs but much larger clustering coefficients than a random graph would have. Watts dubs them *small-world graphs*, the subject of his book. He writes with the zeal of one who believes that he has discovered a key model that may unlock the secrets of relationships in every realm of human endeavor. Chapters 2–4 take us through the graph theoretical discussions and analyses of several variations on this model, both algebraically and empirically via computer simulations. In Chapter 5 Watts turns to some interesting real world situations.

Mathematicians have entertained themselves with a real small-world graph since the 1950s: the collaboration graph C. Here the vertices are mathematicians, and two vertices are joined by an edge if the mathematicians have done research together, resulting in a joint publication (for simplicity, we'll say with or without other coauthors). Paul Erdős (1913–1996) is usually viewed as the "center" of C, because he has over 500 collaborators, far more than any other mathematician. A person's *Erdős number* is just the length of a shortest path from that person to Erdős in C. Thus Paul's collaborators have Erdős number 1, their other collaborators (of which there are about 6000) have Erdős number 2, and so on. I have been studying research collaboration for the past five years, mostly by hand so far in the absence of accessible, reliable, and usable data in electronic form, and our Erdős number web site **[5]** contains lists of people with Erdős numbers less than 3, statistics on Erdős's neighborhood in C, preprints, miscellaneous mathematical and biographical information from the fanciful to the serious, and numerous links.

The available electronic evidence certainly leads one to conclude that the large connected component of C is a small-world graph. This component seems to have about $n \approx 200{,}000$ vertices and average degree about $k \approx 5$. (Another 100,000 vertices are in components of very much smaller sizes, including of course isolated vertices representing mathematicians with no joint papers.) We haven't yet been able to calculate average path lengths, but it appears that the typical mathematician is within about four steps of Erdős, so L can't be much more than 8. Since people tend to write papers more and more in larger and larger groups (fewer than half of all new papers listed in *Mathematical Reviews* have just one author) or with others working in their subspecialties, it is reasonable to expect that the average γ is significantly positive, although we haven't yet been able to calculate it, again for lack of reliable electronic data. (For the neighborhood of Paul Erdős, $\gamma \approx 0.012$, but this is not typical since his degree is so large. A more representative example might be my own neighborhood, for which $\gamma = 15/C(19,2) \approx 0.088$. Someone with only one jointly authored paper, which is a three-author work, would have $\gamma = 1$. Watts seems to ignore the problem of vertices of degree 1, for which γ would be undefined.)

The author acknowledges this primordial example of a small-world graph but needs to get his hands on actual data in electronic form to test his theories in the real world. The book contains three fascinating examples. The first is the collaboration graph of film actors, in which an edge joins two actors if they have appeared in a movie together. The producers of the "Kevin Bacon Game" web site (see **[5]** for a link to it) provided the author with the data, and Watts found that $n = 225{,}226$, $k \approx 61$, $L \approx 3.65$, and $\gamma \approx 0.79$, a small-world graph if there ever was one. His next example is based on the network used for electricity transmission in the western United States. For this power grid, $n = 4941$, $k \approx 2.67$, $L \approx 18.7$, and $\gamma \approx 0.08$. Again, L and γ fall into the small-world range. The third example comes from biology: the (almost) complete neural network of the worm *C. elegans*. Once again, it's a small world: $n = 282$, $k \approx 14$, $L \approx 2.65$, and $\gamma \approx 0.28$. The author encourages others to cast the net further and report back on the results: try science citations, word associations, organizational structures, or, of course, the World Wide Web. (A recent study of the Web reported in *Nature* [**1**] estimates the average path length at 19.)

The last third of *Small Worlds*—Part II—deals with something quite different. Until now we have looked at networks as static: the edges represent certain connections or relationships among the vertices, but we don't expect the network to *do* anything. But graphical models also provide ways to analyze dynamical systems, such as the spread of a disease in a population or the changing strategies of players in a game in which other players' strategies change as well. Watts has fewer results and more speculations here than in Part I, but the potential impact of his model may be far greater.

As an example, consider this density classification problem in the field of cellular automata. We are given a network containing a large number of processors, each of which can be either "on" or "off" at time t, for $t = 0, 1, 2, \ldots$. Each processor decides what state to enter at time $t + 1$ based on the states of its neighbors and itself at time t. The problem is to design the processors so that if the system is started at time 0 with more processors on than off [respectively, more off than on], then at some future time all the processors will have entered and forever remain in the on [respectively, off] state. Thus we want to solve a global computational problem with a locally connected system. Research on this problem and similar ones has used tools such as genetic algorithms. Watts provides some experimental evidence that a more naive approach works as well if the underlying network is a small-world graph.

Should you read this book, or is reading this review and scanning the references cited here enough? Certainly if you are interested in the details or subtleties of the various graph constructions (we only brush the surface here), or full discussions of the numerous examples, you need to go to the source. As a graph theorist, I found *Small Worlds* quite interesting for its blend of theoretical and experimental graph theory and its treatment of such diverse applications. Although the book could be read with comprehension by someone without a lot of advanced mathematical training, some of it will come off as either too dry or too technical. For example, I counted 120 half-page plots, showing the behavior of the parameters mentioned here and several more, and Chapter 4 contains some tedious elementary algebraic derivations of the expectations associated with the various constructions. The book suffers just a bit from the annoying common errors that scientists and engineers make when writing about our subject: occasionally sloppy mathematics, too many definitions and subscripted symbols, imprecision when precision is called for, the overuse of what appear to be technical terms in nontechnical ways, and misuse of some standard mathematical conventions (here big-O notation, for example). But these are minor gripes, and the book is basically well written.

Watts' book seems to be an uneasy compromise between two other books that could have been written: a serious mathematical treatise, in collaboration with an expert in random graph theory, about the properties of these fascinating graphs that interpolate between order and randomness; and a lush account of the applications for a general audience, with the technicalities drastically pared down. Nevertheless, the compromise seems to work on its own terms. There is a nice bibliography: the author has clearly done his homework, and the apparent breadth and depth of his knowledge in so many different disciplines is impressive. (Since leaving Cornell, Duncan Watts has worked with sociologists and economists, among others, in postdoctoral positions at the Center for Social Sciences at Columbia University, the Santa Fe Institute, and the Laboratory for Financial Engineering in the MIT Sloan School of Management.) We should be thankful for the author's enthusiasm for his work and the publicity that such a book generates. One thing is clear: the subject has only begun to be studied, and it is an exciting and fruitful area for research at all levels. On the one hand, it's something undergraduates can sink their teeth into, and on the other hand, it is full of profound open questions for specialists in mathematics, physics, biology, computer science, and myriad other fields.

REFERENCES

1. Réka Albert, Hawoong Jeong, and Albert-László Barabási, Diameter of the World-Wide Web, *Nature* **401** (9 September 1999) 130–131.
2. Sandra Blakeslee, Mathematicians show how big world becomes a small one, *The New York Times*, 16 June 1998, F3.
3. B. Bollobás and F. R. K. Chung, The diameter of a cycle plus a random matching, *SIAM J. on Discrete Math.* **1** (1988) 328–333.
4. James J. Collins and Carson C. Chow, It's a small world, *Nature* **393** (4 June 1998) 409–410.
5. Jerrold W. Grossman, The Erdős Number Project, World Wide Web site, January, 2000, `http://www.oakland.edu/~grossman/erdoshp.html`.
6. Ivars Peterson, Close connections: It's a small world of crickets, nerve cells, computers, and people, *Science News* **154** (22 August 1998) 124–126.
7. Duncan J. Watts and Steven H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (4 June 1998) 440–442.

Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309-4485
grossman@oakland.edu

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**

with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P. *African Americans in Mathematics II.* Eds: Nathaniel Dean, Cassandra M. McZeal, Pamela J. Williams. Contemp. Math., V. 252. AMS, 1999, xii + 168 pp, $35 (P). [ISBN 0-8218-1195-9] Proceedings of the Fourth Conference for African–American Researchers in the Mathematical Sciences held at Rice University in 1998. Three parts: research talks; poster presentations; historical articles.

General, S. *The Math Chat Book.* Frank Morgan. MAA, 2000, xiv + 113 pp, $19.95 (P). [ISBN 0-88385-530-5] Gathers together the dialogue the author has guided first through a column in *The Christian Science Monitor*, then on a live call-in TV show. Good evidence that mathematics correctly presented is of interest to general audiences. AWR

Logic, P. *An Essay in Universal Semantics.* Achille C. Varzi. Topoi Lib., V. 1. Kluwer Academic, 1999, xi + 146 pp, $72. [ISBN 0-7923-5629-2] Semantic theories are usually based on the selection of different languages, different classes of models, and admissibility relations between them. This book develops a general theory of valuations from languages to models that do not necessarily yield perfect homomorphisms, to give useful semantic information even in the presence of incompleteness and inconsistency. RM

Number Theory, T(18: 1, 2), P. *Cohomology of Number Fields.* Jürgen Neukirch, Alexander Schmidt, Kay Wingberg. Ser. of Comprehensive Stud. in Math., V. 323. Springer-Verlag, 2000, xv + 699 pp, $109. [ISBN 3-540-66671-0] Applications of cohomology to number theory, in particular cohomology on the set of all finite extensions of a field. As an example, reciprocity laws become duality statements. DB

Number Theory, P. *Integral Quadratic Forms and Lattices.* Eds: Myung-Hwan Kim, *et al.* Contemp. Math., V. 249. AMS, 1999, ix + 302 pp, $95 (P). [ISBN 0-8218-1949-6] Proceedings of a 1998 conference at Seoul National University. Includes both expository articles and presentations of new research results.

Linear Algebra, T*(14: 1). *Elementary Linear Algebra: A Matrix Approach.* Lawrence E. Spence, Arnold J. Insel, Stephen H. Friedberg. Prentice Hall, 2000, xiv + 477 pp. [ISBN 0-13-716722-9] Carefully organized, well-paced, matrix-oriented presentation of standard topics with numerous applications (many motivate definitions and concepts). Early introduction of linear combinations, linear independence and spanning sets in context of Euclidean 2- and 3-space; abstract vector spaces introduced in last chapter. Exercises of varying difficulty, some requiring technology (no specific software or calculator presumed). JNC

Group Theory, P. *Groups, Languages and Geometry.* Ed: Robert H. Gilman. Contemp. Math., V. 250. AMS, 1999, ix + 136 pp, $35 (P). [ISBN 0-8218-1053-7] Proceedings of a 1998 AMS–IMS–SIAM Joint Summer Research Conference.

Real Analysis, T(16–17: 2). *Fourier and Wavelet Analysis.* George Bachman, Lawrence Narici, Edward Beckenstein. Universitext. Springer-Verlag, 2000, ix + 505 pp, $59.95.

[ISBN 0-387-98899-8] A book for students who know nothing about Fourier series when they start. Introduces the standard material plus fast transforms and wavelets. Draws a little on Lebesgue integrals and very minimally on Hilbert space ideas. AWR

Dynamical Systems, T(18: 1), P. *Matrix Diagonal Stability in Systems and Computation.* Eugenius Kaszkurewicz, Amit Bhaya. Birkhäuser Boston, 2000, xiv + 267 pp, $69.95. [ISBN 0-8176-4088-6] Emphasizes the possibility of studying stability in widespread classes of dynamical systems using diagonal-type Liapunov functions. Theoretical, but includes diverse applications (e.g., electrical power systems, trophic chains in ecology). DK

Dynamical Systems, T*(17: 2). *Analysis and Simulation of Chaotic Systems, Second Edition.* Frank C. Hoppensteadt. Appl. Math. Sci., V. 94. Springer-Verlag, 2000, xx + 315 pp, $69.95. [ISBN 0-387-98943-9] A rich, theoretical introduction to the analysis of dynamical systems organized around the themes of perturbations, iterations, chaos, oscillations, stability, and computer simulation. (*First Edition*, TR, December 1993.) DK

Dynamical Systems, T(16: 1), C. *Mechanics and Dynamical Systems with Mathematica.* Nicola Bellomo, Luigi Preziosi, Antonio Romano. Model. & Simul. in Sci., Eng., & Tech. Birkhäuser Boston, 2000, xiii + 417 pp, $69.95. [ISBN 0-8176-4007-X] A general introduction to modeling with differential equations with a strong emphasis on classical mechanics. Symbolic and numerical solutions, analysis, and visualization. Presupposes some previous exposure to differential equations. DK

Numerical Analysis, P. *Error Control and Adaptivity in Science Computing.* Eds: Haydar Bulgak, Christoph Zenger. NATO Sci. Ser. C., V. 536. Kluwer Academic, 1999, xvi + 354 pp, $180. [ISBN 0-7923-5808-2] Proceedings of a 1998 NATO Advanced Study Institute held in Antalya, Turkey. Main topics: error estimates and error control in numerical linear algebra algorithms, interval arithmetic, and adaptivity for continuous models.

Analysis, P. *The Functional and Harmonic Analysis of Wavelets and Frames.* Eds: Lawrence Wasson Baggett, David Royal Larson. Contemp. Math., V. 247. AMS, 1999, x + 306 pp, $75 (P). [ISBN 0-8218-1957-7] Proceedings of an AMS Special Session held in San Antonio, Texas in 1999.

Topology, T(18), P. *Lectures on the Topology of 3-Manifolds: An Introduction to the Casson Invariant.* Nikolai Saveliev. Walter de Gruyter, 1999, ix + 199 pp, $59 (P). [ISBN 3-11-016271-7] Introduces readers who have only a basic knowledge of algebraic topology to the Casson λ-invariant. Begins with Heegaard splittings and Dehn surgery, moves on to topics in 4-manifolds and knot theory, then ends with Casson's invariant and its applications. No exercises. JD

Topology, P. *Pseudoperiodic Topology.* Eds: Vladimir Arnold, Maxim Kontsevich, Anton Zorich. AMS Transl., Ser. 2, V. 197. AMS, 1999, xii + 178 pp, $85. [ISBN 0-8218-2094-X] 5 survey papers on topics at the boundary between the ergodic theory of dynamical systems, topology, and number theory.

Topology, P. *Geometry and Topology in Dynamics.* Eds: Marcy Barge, Krystyna Kuperberg. Contemp. Math., V. 246. AMS, 1999, ix + 252 pp, $59 (P). [ISBN 0-8218-1958-5] Proceedings of the 1998 AMS Special Session on Topology in Dynamics held in Winston–Salem, North Carolina, and the 1999 AMS–AWM Special Session on Geometry in Dynamics held in San Antonio, Texas.

Optimization, P. *Modern Optimisation Techniques in Power Systems.* Ed: Yong-Hua Song. Intern. Ser. on Microprocessor-Based & Intelligent Systems Eng., V. 20. Kluwer Academic, 1999, xi + 275 pp, $141. [ISBN 0-7923-5697-7] 11 papers provide an overview of the most important optimization methods used in managing electrical power systems: simulated annealing, tabu search, genetic algorithms, neural networks, fuzzy programming, Lagrangian relaxation, interior point methods, ant colony search, and hybrid methods.

Optimization, S(14–16). *Linear Optimization in Applications.* S.L. Tang. Hong Kong Univ Pr, 1999, vii + 161 pp, $17.50 (P). [ISBN 962-209-483-X] Example-driven introduction and overview of linear, integer, and goal programming. Examples include transportation, shipment, blending, assignment, knapsack, set covering and packing, project scheduling, traveling salesperson, etc. RM

Optimization, P. *Progress in Optimization: Contributions from Australasia.* Andrew Eberhard, *et al.* Appl. Optimiz., V. 30. Kluwer Academic, 1999, xxi + 301 pp, $150. [ISBN 0-7923-5733-7] 14 survey papers on topics in nonsmooth analysis, generalized convexity, algorithms for nonsmooth programming, global optimization, and control methodologies.

Optimization, P. *Advances in Steiner Trees.* Eds: Ding-Zhu Du, J.M. Smith, J.H. Rubinstein.

Comb. Optim., V. 6. Kluwer Academic, 2000, xii + 323 pp, $159. [ISBN 0-7923-6110-5] In two parts: 6 papers on the general geometric Steiner tree problem; 7 papers on the Steiner problem on graphs.

Mathematical Modeling, C, P, L. *Mathematical Modeling and Optimization: An Essay for the Design of Computer-Based Modeling Tools.* Tony Hürlimann. Appl. Optimiz., V. 31. Kluwer Academic, 1999, xxii + 313 pp, $149. [ISBN 0-7923-5927-5] Articulate presentation of a general framework for specifying and solving mathematical models of diverse types. Intended for designers of modeling tools. Presents many philosophical and logical principles of modeling as well as a concrete computer implementation of the framework. Even model builders can learn something here. DK

Mathematical Modeling, T(15: 1), S. *Mathematical Modelling: Concepts and Case Studies.* J. Caldwell, Y.M. Ram. Math. Model.: Theory & Applic., V. 6. Kluwer Academic, 1999, xii + 285 pp, $126. [ISBN 0-7923-5820-1] Introduction to modeling (both continuous and discrete) and numerical techniques. Includes case studies in dynamics, biological treatment systems, contagious diseases, cam design, force analysis in mechanics. RM

Mathematical Modeling, T(17–18: 1, 2), S, P. *Risk Modeling, Assessment, and Management.* Yacov Y. Haimes. Ser. in Systems Eng. Wiley, 1998, xviii + 726 pp, $105. [ISBN 0-471-24005-2] Basic and advanced risk analysis methods. Emphasis on modeling complex systems using a variety of techniques as a prelude to assessing and managing risks. Assumes facility with standard undergraduate engineering mathematics. No exercises, but many example problems are worked in detail. LB

Optimal Control, T(18: 1), P. *Lecture Notes in Control and Information Sciences–248: Iterative Learning Control.* Yangquan Chen, Changyun Wen. Springer-Verlag, 1999, xii + 199 pp, $69.80 (P). [ISBN 1-85233-190-9] Covers the mathematics of Iterative Learning Control, improving control systems for repetitive tasks by updating the feedback control function after each repetition. Presents both theory and simulation studies. DK

Probability, S(13), L. *Taking Chances: Winning with Probability.* John Haigh. Oxford Univ Pr, 1999, xiv + 330 pp, $35. [ISBN 0-19-850292-3] A delightful introduction to probability concepts and calculations in the context of games: lotteries, football pools, dice, snooker, roulette and poker, bookmaking, and so on. Intended for the math-phobic reader, but doesn't pander or hand-wave. Problems and solutions. DK

Stochastic Processes, T(15–16: 1, 2), L. *An Introduction to Stochastic Modeling, Third Edition.* Howard M. Taylor, Samuel Karlin. Academic Pr, 1998, xi + 631 pp. [ISBN 0-12-684887-4] New in this edition: short, optional section on martingales; chapter on Brownian motion and applications, including the Black–Scholes formula for stock option pricing. Includes open-ended "Computer Challenges" requiring combination of simulation, numerical exploration, and paper-and-pencil work. (*Second Edition*, TR, May 1994.) LB

Statistics, P. *Maximum Entropy and Bayesian Methods.* Eds: Wolfgang von der Linden, *et al.* Fund. Theories of Physics, V. 105. Kluwer Academic, 1999, xvi + 360 pp, $177. [ISBN 0-7923-5766-3] Proceedings of the 18th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis held in Garching, Germany in 1998.

Applications (Information Theory), P. *Information Systems Analysis and Modeling: An Informational Macrodynamics Approach.* Vladimir S. Lerner. Ser. in Eng. & Comp. Sci. Kluwer Academic, 2000, xxvi + 303 pp, $139.50. [ISBN 0-7923-8683-3-] Information systems are characterized by attributes of randomness, generation of information (things which have "measure"), dynamic processes which exhibit regularities and create orderliness, and emergent structures which bind information in some way. Theoretical development of these ideas, with goal of analysis and interdisciplinary applications. RM

Applications, P. *Practical Applications of Fuzzy Technologies.* Ed: Hans-Jürgen Zimmermann. Handbooks of Fuzzy Sets Ser. Kluwer Academic, 1999, xxviii + 667 pp, $229.95. [ISBN 0-7923-8628-0] 18 survey papers on applications on fuzzy methods in engineering and natural sciences, medicine, management, and behavioral, cognitive, and social sciences. A final paper discusses some of the software tools available for building fuzzy systems.

Reviewers

LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; JNC: Judith N. Cederberg, St. Olaf; JD: Jill Dietz, St. Olaf; DK: Danny Kaplan, Macalester; RM: Richard Molnar, Macalester; AWR: A. Wayne Roberts, Macalester.

LESTER R. FORD AWARDS FOR 1999

The Lester R. Ford Awards, established in 1964, are made annually to authors of outstanding expository papers in the MONTHLY. The awards are named for Lester R. Ford, Sr., a distinguished mathematician, editor of the MONTHLY (1942–1946), and President of the Mathematical Association of America (1947–1948).

Winners of the Lester R. Ford Awards for expository papers appearing in Volume 106 (1999) of the MONTHLY are:

P. J. McKenna, University of Connecticut, Large Torsional Oscillations in Suspension Bridges Revisited: Fixing an Old Approximation, pp. 1–18.

Linear approximations and, in particular, the approximations $\sin\theta = \theta$, $\cos\theta = 1$, for small θ, are pervasive in applications of differential equations. Recently, increased computing power has made it possible to model a range of nonlinear phenomena with very complicated behavior. McKenna's article, which is based on what is arguably the most famous disaster of twentieth-century engineering—the 1940 collapse of the Tacoma Narrows suspension bridge, presents the relevant differential equations in detail and gives many computer-generated solutions for special values of parameters. It is masterfully written, with the right amount and kind of technical detail, moving easily and authoritatively from motivation to modeling to numerical experiments to discussion of limitations. There are ample references for readers who wish to pursue the story, including some of the controversy that has greeted earlier work of the author and co-workers on the subject. The article would make an excellent supplement to courses in differential equations or modeling.

William J. Terrell, Virginia Commonwealth University, Some Fundamental Control Theory I and II, pp. 705–719 and 812–828.

It is standard to convert a linear n^{th} order differential equation into an equivalent linear system of first order equations in n variables to analyze it. The author addresses the converse problem when it is possible to convert a linear system into an equivalent n^{th} order equation, and uses this as an entree to the range of ideas involved in control theory, including controllability, feedback, observability, and duality.

His paper provides a clear progression of ideas, illuminated by examples and a judicious use of proofs. A reader familiar with the basic theory of ordinary linear differential equations and linear algebra will be able to enjoy the paper. Extensions of the theory are indicated, along with references for those wishing to delve more deeply into control theory.

Vilmos Totik, Bolyai Institute, Szeged, Hungary, and University of South Florida, A Tale of Two Integrals, pp. 227–240.

The article opens with a simply stated problem in analysis: If two functions on an interval have integrals equal to the length of the interval, there must be a subinterval on which the two functions have integrals equal to half the length of the original interval. The problem is more difficult than one might expect, observes the author. He then proceeds to offer several solutions, first giving an equivalent combinatorial formulation of the problem. The remarkable feature of this article is that major, seemingly unrelated, results from several areas of mathematics come to bear on the solution of this "simple" problem. These include the Borsuk–Ulam theorem, the chord theorem, the chess king-moving theorem, the winding number theorem, and the Jordan curve theorem.