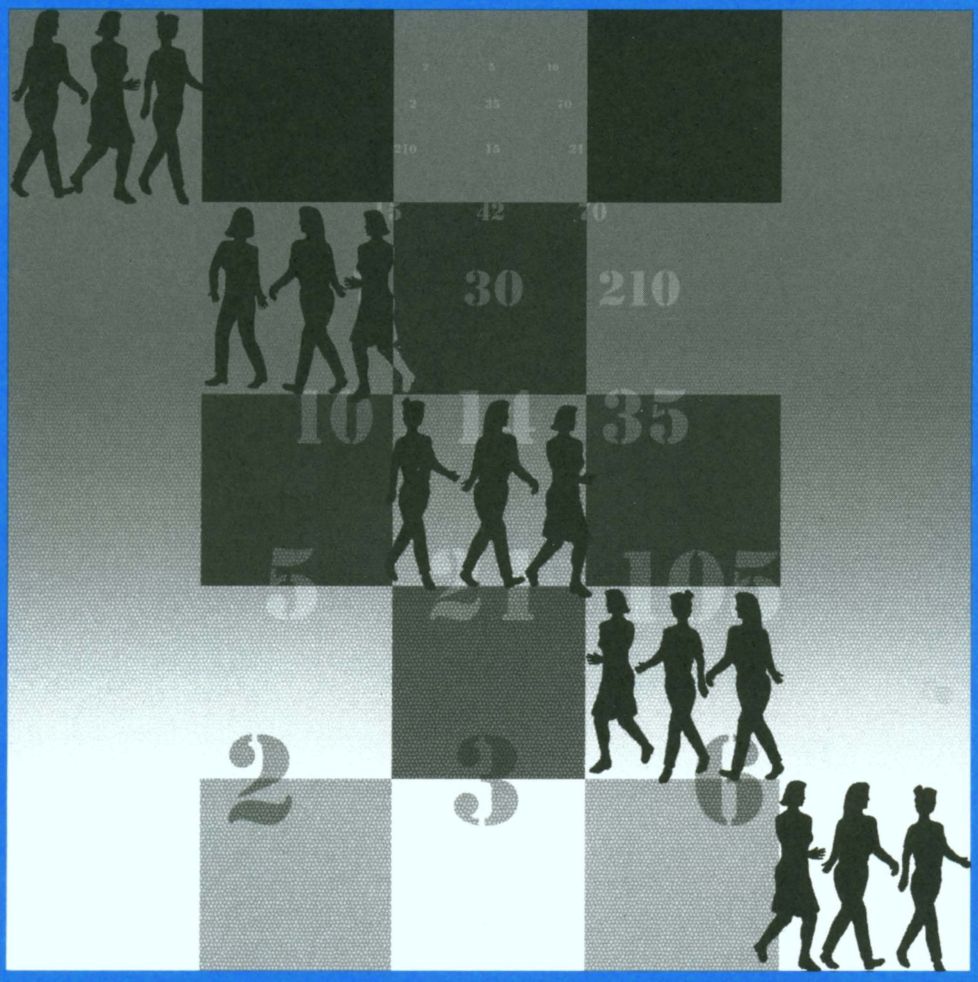
Vol. 82, No. 1, February 2009

MATHEMATICS MAGAZINE



- Kirkman's Schoolgirls Wearing Hats and Walking through Fields of Numbers
- When Euler Met l'Hôpital
- Matroids You Have Known

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

CONTENTS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Please submit new manuscripts by email to Editor-Elect Walter Stromquist at mathmag@maa.org. A brief message with an attached PDF file is preferred. Word-processor and DVI files can also be considered. Alternatively, manuscripts may be mailed to Mathematics Magazine, 132 Bodine Rd., Berwyn, PA 19312-1027. If possible, please include an email address for further correspondence.

Cover image: *Kirkman's Schoolgirls Walking through Fields of Numbers*, by Hunter Cowdery, art student at West Valley College, who is animating his way to San Jose State University, and Jason Challas, who lectures on computer graphics and fine art at West Valley.

AUTHORS

Ezra (Bud) Brown grew up in New Orleans, has degrees from Rice and LSU, and has been at Virginia Tech since 1969, where he is currently Alumni Distinguished Professor of Mathematics. His research interests include number theory and combinatorics, and he particularly enjoys discovering connections between apparently unrelated areas of mathematics and working with students who are engaged in research. In his spare time, Bud enjoys singing everything from grand opera to rock and roll, playing jazz piano, and talking about his granddaughter Phoebe Rose. Under the direction of his wife Jo, he has become a fairly tolerable gardener. He occasionally bakes biscuits for his students, and he once won a karaoke contest.

William Dunham is the Truman Koehler Professor of Mathematics at Muhlenberg College and a card-carrying member of the Euler Fan Club. Over the years, Dunham has enjoyed grazing through Euler's work, finding nuggets of pure gold, and sharing them with the wider community. The present paper is a case in point: here Euler resolves the Basel Problem (i.e., sums the reciprocals of the squares) by using four transcendental functions and three applications of l'Hôpital's rule! It is Uncle Leonhard at his symbol-manipulating best.

Keith E. Mellinger earned his Ph.D. in finite geometry at the University of Delaware. After graduate school he spent two years as a VIGRE postdoc at the University of Illinois at Chicago. He currently lives in Fredericksburg, VA, where he is an associate professor and department chair at the University of Mary Washington. Keith's interests include many areas of discrete mathematics and he regularly lures unsuspecting undergraduates into his projects. This article grew out of some lively conversations between Keith and Bud about whether or not anybody would ever want to read something they would coauthor. In his spare time, Keith enjoys spending time with his wife and two darling children, and plays guitar and mandolin in a local bluegrass band.

David L. Neel earned his Ph.D. at Dartmouth College under Kenneth Bogart, who started him down the path of matroids. He has since strayed occasionally into graph theory as he migrated west to Truman State University in Missouri and then westward again to Seattle University, his current milieu where he serves as an associate professor of mathematics. There he enjoys discrete math, and, discreetly, other arts like literature (Woolf, Gaddis, Wallace, Bernhard), film (Coens, P.T. Anderson, Antonioni), and music (Radiohead, Schoenberg, Monk). He believes the world would benefit from fiction written with mathematical sophistication.

Nancy Ann Neudauer received her Ph.D. from the University of Wisconsin. Her research interests include matroid theory, graph theory, enumeration, and their connections to other areas of mathematics. She has been involved in the MAA for as long as she can remember, at least since she gave a talk at an MAA meeting while still a high school student. She now serves on the Board of Governors, is active in the Pacific Northwest section, is a PNW section NExT officer, and will always be a silver dot. She enjoys finding the hidden matroid in all mathematics talks she attends. When not trying to show her students the beauty of matroids, she races sailboats and travels where ever she can, rarely leaving the house without her passport and a packed bag, just in case.

Vol. 82, No. 1, February 2009

MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

Subscription correspondence and notice of change of address should be sent to the Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to

MAA Advertising
1529 Eighteenth St. NW
Washington DC 20036

Phone: (866) 821-1221
Fax: (202) 387-1208
E-mail: advertising@maa.org

Further advertising information can be found online at www.maa.org

Change of address, missing issue 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

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

Postmaster: Send address changes to Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

Printed in the United States of America

ARTICLES

Kirkman's Schoolgirls Wearing Hats and Walking through Fields of Numbers

EZRA BROWN
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
brown@math.vt.edu

KEITH E. MELLINGER
University of Mary Washington
Fredericksburg, VA 22401
kmelling@umw.edu

Fifteen young ladies at school

Imagine fifteen young ladies at the Emmy Noether Boarding School—Anita, Barb, Carol, Doris, Ellen, Fran, Gail, Helen, Ivy, Julia, Kali, Lori, Mary, Noel, and Olive. Every day, they walk to school in the Official ENBS Formation, namely, in five rows of three each. One of the ENBS rules is that during the walk, a student may only talk with the other students in her row of three. These fifteen are all good friends and like to talk with each other—and they are all mathematically inclined. One day Julia says, "I wonder if it's possible for us to walk to school in the Official Formation in such a way that we all have a chance to talk with each other at least once a week?" "But that means nobody walks with anybody else in a line more than once a week," observes Anita. "I'll bet we can do that," concludes Lori. "Let's get to work." And what they came up with is the schedule in TABLE 1.

TABLE 1: Walking to school

MON	TUE	WED	THU	FRI	SAT	SUN
a, b, e	a, c, f	a, d, h	a, g, k	a, j, m	a, n, o	a, i, l
c, l, o	b, m, o	b, c, g	b, h, l	b, f, k	b, d, i	b, j, n
d, f, m	d, g, n	e, j, o	c, d, j	c, i, n	c, e, k	c, h, m
g, i, j	e, h, i	f, l, n	e, m, n	d, e, l	f, h, j	d, k, o
h, k, n	j, k, l	i, k, m	f, i, o	g, h, o	g, l, m	e, f, g

TABLE 1 was probably what T. P. Kirkman had in mind when he posed the Fifteen Schoolgirls question in 1850. Appearing in the unlikely-sounding *Lady's and Gentlemen's Diary* [**15**], it reads as follows:

> Fifteen young ladies of a school walk out three abreast for seven days in succession: it is required to arrange them daily so that no two shall walk abreast more than once.

Kirkman's publication of this problem and solution [**15, 16**] is one of the starting points for what has become the vast modern field of combinatorial design theory. Its

poser, Thomas Pennyngton Kirkman (1806–1895), is one of the more intriguing figures in the history of mathematics. He published his first mathematical paper when he was 40, and was the first to describe many structures in discrete mathematics. Among these are block designs, which form the basis for the statistical design of experiments; bipartite graphs, which are essential for such problems as classroom scheduling and medical school admissions; and Hamiltonian circuits, which are at the heart of the famous Traveling Salesman Problem. (Biggs [**2**] gives more details about Kirkman's life and work.) For these achievements, combinatorialists regard him as the "Father of Design Theory"—yet his fame outside the field rests entirely on the Schoolgirls Problem and his solution.

This story is about the very problem that made Kirkman famous. His solution is an example of a *resolvable* $(15, 35, 7, 3, 1)$*-design*, and we begin by explaining what those words and numbers mean. We describe how one of us found such a design by looking in a most unlikely place: the algebraic number field $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$. This proves to be a particularly fertile field in which several other block designs grow. We talk about spreads and packings in finite geometries, how a particular packing in the geometry $PG(3, 2)$ answers Kirkman's question, and how the $PG(3, 2)$ design is really the same as the number field design. Finally, we show how our design is a solution to a certain problem in recreational mathematics called the Fifteen Hats Problem.

We begin by talking about block designs.

Block designs and Kirkman Triple Systems

Design theory began with Euler's studies of Latin squares in the 18th century, interest in which was recently rekindled with the world-wide popularity of Sudoku. Many decades after their invention by Kirkman, block designs appeared in connection with R. A. Fisher's work [**10, 11**] on the statistical design of agricultural experiments, and the first comprehensive mathematical study of the field was due to R. C. Bose [**4**]. More recently, they have found applications in coding theory, cryptography, network design, scheduling, communication theory, and computer science. Finally, designs have always appealed to mathematicians because of their elegance, beauty, high degree of symmetry, and connections with many other fields of mathematics [**5**].

A *balanced incomplete block design* with parameters v, b, r, k, and λ is a collection $\mathcal{B}$ of b subsets (or *blocks*) of a v-element set V of objects (or *varieties*) such that each block contains k varieties, each variety appears in r blocks and each pair of distinct varieties appears together in λ blocks. Such a design is also called a (v, b, r, k, λ)-design. We say a design like this is *incomplete* if $k < v$. From a combinatorial point of view, complete designs are not very interesting. However, statisticians do use them to design experiments.

The five parameters in these designs are not independent. Since there are b blocks, each of size k, there are bk occurrences of varieties in the design. On the other hand, there are v varieties, each occurring in r blocks, and so a total of vr varieties appear in the design. Hence $bk = vr$. A similar counting argument shows that $r(k-1) = \lambda(v-1)$. Hence

$$r = \frac{\lambda(v-1)}{k-1} \quad \text{and} \quad b = \frac{\lambda v(v-1)}{k(k-1)}.$$

Because of these relations, such a design is frequently called a (v, k, λ)-design. (There are more details about block designs in [**5**].)

Given a block design with varieties $x_1, \ldots, x_v$ and blocks $B_1, \ldots, B_b$, an efficient way to represent it is by its *incidence matrix*. This is a $b \times v$ matrix $M = [m_{ij}]$, where $m_{ij} = 1$ if $x_j \in B_i$ and $m_{ij} = 0$ otherwise.

A reading of the Kirkman Schoolgirls Problem reveals that he first asks for an arrangement of 15 schoolgirls into sets of size three such that each pair of girls is present in at most one of these triples. There are five triples for each of seven days, making 35 triples in all. Moreover, each girl appears in just one triple each day, and over seven days, each girl would thus appear with each other girl exactly once. We conclude that Kirkman is asking for a way to arrange the girls into a (15, 3, 1)-design. (The incidence matrix for Kirkman's design will reappear when we ask the schoolgirls to wear hats.)

But there is more: he asks for a way to arrange the $b = 35$ triples into seven days of five triples each, so that each girl appears in exactly one triple each day. Such a design, whose b blocks can be arranged into r *parallel* classes of $n = v/k$ blocks each such that each variety appears exactly once in each class, is called *resolvable*. For such a design to exist, v must be a multiple of k. In Kirkman's honor, a resolvable $(3n, 3, 1)$-design is called a *Kirkman Triple System*. (A $(v, 3, 1)$-design is called a *Steiner Triple System*, despite the fact that Kirkman described them six years before Jakob Steiner's publication on the subject—but that's another story.)

Do Kirkman Triple Systems exist? Yes, they do. The smallest possibility has $v = 3$, with exactly one block and one parallel class, but the smallest nontrivial case has $v = 9$. Construction begins with the magic square of order 3, that familiar arrangement of the numbers 1 through 9 into a 3×3 grid such that the triples of numbers in each row, each column and on the two main diagonals add up to 15. The three rows, three columns, three extended diagonals parallel to the principal diagonal, and three more parallel to the principal contrary diagonal form the four parallel classes of a resolvable (9, 3, 1)-design. The following picture tells the tale, with the magic square on the left and the four parallel classes of the resolvable (9, 3, 1)-design on the right:

8	1	6
3	5	7
4	9	2

{1, 6, 8}	{3, 5, 7}	{2, 4, 9}
{1, 5, 9}	{2, 6, 7}	{3, 4, 8}
{1, 4, 7}	{2, 5, 8}	{3, 6, 9}
{1, 2, 3}	{4, 5, 6}	{7, 8, 9}

The next smallest case has $v = 15$, which is the design Kirkman sought in his query; where do we look? If we could find a structure containing fifteen objects arranged in thirty-five sets, with three objects per set, that would be a place to start. It happens that there are such structures, and we find one of them in the world of algebraic number theory—specifically, in the number field $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$. The field K contains several interesting designs, and we'll talk about them, but first we supply some background about this area of mathematics.

$K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$ and the designs it contains

Évariste Galois (1811–1832) described relations involving the roots of polynomials, number fields, and finite groups, now known as Galois theory. One basic idea is that if $p(x)$ is a polynomial with rational coefficients, then there is a smallest subfield of the complex numbers $\mathbb{C}$ containing all the roots of $p(x)$. This is the *splitting field* of p over $\mathbb{Q}$. If $a, b, \ldots \in \mathbb{C}$, we write $\mathbb{Q}(a, b, \ldots)$ to mean the smallest subfield of $\mathbb{C}$ containing $\mathbb{Q}$ and $a, b, \ldots$. For example, the splitting field of the polynomial $p(x) =$

$(x^2-2)(x^2-3)(x^2-5)$ is the field $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$. Now it is a fact that $\mathbb{Q}(a, b, \ldots)$ is a vector space over $\mathbb{Q}$, and the *degree* of $\mathbb{Q}(a, b, \ldots)$ over $\mathbb{Q}$ is the dimension of this vector space. These splitting fields have a good bit of internal structure, which we illustrate with the field $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$, described in [**5**].

Now by definition, the *biquadratic* (degree-4) field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ contains the two elements $\sqrt{2}$ and $\sqrt{3}$, and since it is a field, it also contains $\sqrt{2}\sqrt{3} = \sqrt{6}$. Hence $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ also contains three *quadratic* (degree-2) subfields: $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{3})$, and $\mathbb{Q}(\sqrt{6})$. A similar argument shows that $\mathbb{Q}(\sqrt{6}, \sqrt{10})$ contains $\sqrt{15} = \sqrt{6}\sqrt{10}/2$, and so it also contains the three quadratic subfields $\mathbb{Q}(\sqrt{6})$, $\mathbb{Q}(\sqrt{10})$, and $\mathbb{Q}(\sqrt{15})$. In the same vein, one can show that $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$ contains seven quadratic subfields $\mathbb{Q}(\sqrt{d})$, for $d = 2$, 3, 5, 6, 10, 15, and 30, and seven *biquadratic* subfields $\mathbb{Q}(\sqrt{d_1}, \sqrt{d_2})$. Not only does each biquadratic subfield contain three quadratic subfields, but each quadratic is contained in three biquadratics, and in [**5**], these subfields of $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$ are shown to form a (7, 7, 3, 3, 1)-design with the biquadratic fields as the blocks and the quadratic fields as the varieties. Such a design, in which $b = v$ and $r = k$, is called a *symmetric* design, and we will encounter some more symmetric designs later in this section.

We now turn to the polynomial $p(x) = (x^2-2)(x^2-3)(x^2-5)(x^2-7)$, whose splitting field is the degree-16 field $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$, the smallest subfield of the complex numbers containing $\mathbb{Q}(\sqrt{d})$ for $d = 2$, 3, 5, and 7. Now let $S = \{2, 3, 5, 6, 7, 10, 14, 15, 21, 30, 35, 42, 70, 105, 210\}$. Then K contains the 15 quadratic subfields $\mathbb{Q}(\sqrt{d})$ for $d \in S$. Moreover, each pair of these quadratics is contained in a unique biquadratic subfield of K, and each biquadratic contains three quadratics. A counting argument shows that K contains 35 biquadratic subfields $\mathbb{Q}(\sqrt{d_1}, \sqrt{d_2})$, and it is straightforward to show that each quadratic is contained in seven biquadratics.

Now consider the block design with the 15 quadratic subfields of K as varieties and the 35 biquadratic subfields of K as blocks. Our work in the previous paragraph shows that these form a block design with $v = 15$, $b = 35$, $r = 7$, $k = 3$, and $\lambda = 1$, that is, a (15, 3, 1)-design, which we call KS for short. But is KS resolvable?

In fact, it is, and TABLE 2 shows the seven columns that are the seven parallel classes. The three numbers in each of the 35 cells in this table determine a block, that is, one of the 35 biquadratic subfields of K. We began by placing the seven biquadratic subfields containing $\mathbb{Q}(\sqrt{2})$ in separate classes across the top row and proceeded, mainly by trial and error, to arrange the thirty-five blocks in seven parallel classes. The end result is a resolvable (15, 3, 1)-design—in short, a solution to Kirkman's Schoolgirls problem.

But that is not all. The field K also contains another resolvable (15, 3, 1) design as well as two other types of designs. We construct the other Kirkman design as follows.

TABLE 2: The Kirkman design in $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$

MON	TUE	WED	THU	FRI	SAT	SUN
2, 3, 6	2, 5, 10	2, 7, 14	2, 15, 30	2, 21, 42	2, 35, 70	2, 105, 210
5, 21, 105	3, 70, 210	3, 5, 15	3, 14, 42	3, 35, 105	3, 7, 21	3, 10, 30
7, 30, 210	6, 14, 21	6, 35, 210	5, 7, 35	5, 6, 30	5, 42, 210	5, 14, 70
10, 14, 35	7, 15, 105	10, 42, 105	6, 70, 105	7, 10, 70	6, 10, 15	6, 7, 42
15, 42, 70	30, 35, 42	21, 30, 70	10, 21, 210	14, 15, 210	14, 30, 105	15, 21, 35

The blocks are the 35 biquadratic subfields of K, and the varieties are the 15 *octic* (degree-8) subfields of K, which we number a through o as in TABLE 3. Notice that a is the subfield $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$, with which we began this section. But in a reversal of the previous construction, a variety (octic field) is a member of those blocks (biquadratic fields) *that it contains as a subfield*. That is, "contains" means "is a subfield of" in this context. Thus, the "block" $\mathbb{Q}(\sqrt{21}, \sqrt{35})$ "contains" the three "varieties" d, o and k, as shown in TABLE 3.

It is straightforward to show that each of the 35 biquadratic subfields of K is a subfield of exactly three of these octic fields, each octic contains seven biquadratic subfields, and each pair of biquadratics are subfields of a unique octic. Thus, we have another (15, 3, 1) -design, which we call KS^*.

Is KS^* resolvable? Yes, it is, and to see this, we look at TABLE 2 again. In it, each biquadratic is designated by the triple of *quadratics it contains*. If we replace each biquadratic in TABLE 2 by the triple of *octics that contain it*, we are led to TABLE 1, the arrangement found by the fifteen ladies at the ENBS.

The field K contains fifteen octic subfields, and each of these contains seven quadratic subfields. It turns out that each quadratic appears in seven octics, and that each pair of quadratics appear together in exactly three octics. This gives us a symmetric (15, 7, 3)-design OQ with the quadratics as varieties and the octics as blocks. Each row of TABLE 3 begins with a letter referring to an octic field, followed by seven numbers $d_1, \ldots, d_7$; these are the values of d for which $Q(\sqrt{d})$ is contained in that octic field. For example, line l refers to the octic field $L = \mathbb{Q}(\sqrt{3}, \sqrt{10}, \sqrt{14})$. It contains the seven quadratic subfields $\mathbb{Q}(\sqrt{r})$ for $r = 3$, 10, 14, 30, 35, 42, and 105.

Now, the elements of the blocks in TABLE 3 can themselves be arranged into block designs. For each of the 15 octic subfields of K contains 7 biquadratic subfields (the

TABLE 3: The (15, 7, 3)-design OQ in $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$

Octic Field	**Contains $\mathbb{Q}(\sqrt{d})$ for these d**
a	2, 3, 5, 6, 10, 15, 30
b	2, 3, 7, 6, 14, 21, 42
c	2, 5, 7, 10, 14, 35, 70
d	3, 5, 7, 15, 21, 35, 105
e	2, 3, 6, 35, 70, 105, 210
f	2, 5, 10, 21, 42, 105, 210
g	2, 7, 14, 15, 30, 105, 210
h	3, 5, 14, 15, 42, 70, 210
i	3, 7, 10, 21, 30, 70, 210
j	5, 6, 7, 30, 35, 42, 210
k	2, 15, 21, 30, 35, 42, 70
l	3, 10, 14, 30, 35, 42, 105
m	5, 6, 14, 21, 30, 70, 105
n	6, 7, 10, 15, 42, 70, 105
o	6, 10, 14, 15, 21, 35, 210

blocks) as well as 7 quadratic subfields (the varieties). Each biquadratic contains 3 quadratics, each quadratic is contained in 3 biquadratics, and each pair of quadratics lie in a unique biquadratic. Thus, each block is a triple of quadratics, and we conclude that K contains fifteen symmetric (7, 3, 1)-designs. Continuing with line l, we list the triples of the (7, 3, 1)-design contained in the octic field L here:

10, 14, 35; 30, 35, 42; 10, 42, 105; 3, 14, 42; 3, 35, 105; 14, 30, 105; 3, 10, 30

As an exercise, find these seven triples in TABLE 2, and observe that they occur in different columns.

Finally, if D is a symmetric design, then the *dual* design D^* of D is obtained from D by a formal exchange of blocks and varieties. Thus, if the variety x belongs to the block B in D, then the variety B belongs to the block x in D^*. In this way, we obtain the dual OQ^* of the (15, 7, 3) symmetric design OQ depicted in TABLE 3, again by a formal exchange of blocks and designs. We note that this construction fails for nonsymmetric designs, that is designs in which $v \neq b$.

Exchanging the roles of blocks and varieties in a block design is analogous to exchanging the roles of points and lines in projective geometry. To see this more clearly, we need to pass to a geometric description of KS. So, let's talk about finite projective geometries and spreads.

Spreads in $PG(3, 2)$ and the geometry of Kirkman

One very elegant way to generate a solution to the Kirkman Schoolgirls problem involves a nice partitioning and packing problem in finite projective geometry. Hundreds of years ago, projective spaces arose as extensions of the familiar real Euclidean spaces. The essential difference between Euclidean and projective spaces is that in projective spaces every pair of lines in a plane must intersect—there is no notion of parallelism. This lack of parallelism provides a nice duality to projective planes: Every two distinct points determine a unique line and every two distinct lines meet in a unique point. Lines can be skew, but this requires them to be noncoplanar. We will see examples of skew lines shortly since we will be mostly interested in finite projective 3-space, a place where lines can indeed be skew. But first let's talk more about finite projective geometry.

Just as with Euclidean geometry, there is a way to assign coordinates to the points of a finite projective space. We do this using a finite field (rather than the more familiar fields $\mathbb{R}$ or $\mathbb{Q}$). The classic example of a finite field is the set of integers $\{0, 1, \ldots, p-1\}$, with all arithmetic performed modulo p. But it can be shown that finite fields exist of any size that is a power of a prime. Typically, we use $q = p^k$ for a power of a prime and we let $GF(q)$ denote the finite field with q elements.

The technique for coordinatizing projective spaces is fairly easy and is a straightforward extension of the standard linear algebra techniques that we learn using real numbers. To construct a 3-dimensional projective space, we start with a 4-dimensional vector space over the finite field with q elements, $GF(q)$. The lattice of subspaces then gives us the geometry. That is, 1-dimensional subspaces represent points, 2-dimensional subspaces represent lines, and so on. This is the unique finite projective space of dimension 3 and *order* q, denoted by $PG(3, q)$. Notice that we have a representation problem for points: Since points are defined as 1-dimensional subspaces, all nonzero vectors in a particular 1-dimensional subspace represent the same projective point. This leads to the concept of *homogeneous coordinates* for projective spaces: When we use the nonzero vector (w, x, y, z) to represent a projective point, it is understood that any nonzero scalar multiple of this vector represents the same projective point. (The formalities involve equivalence classes of vectors.)

Now that we have a finite set and a nice representation, we can use standard counting techniques to determine some properties of our space. There are $q^4 - 1$ nonzero vectors in the entire vector space, and any *nonzero* scalar multiple of a nonzero vector gives the same projective point. Hence, the total number of points of $PG(3, q)$ is given by $(q^4 - 1)/(q - 1) = q^3 + q^2 + q + 1$. Similarly counting the number of 1-dimensional subspaces contained in a 2-dimensional subspace, we see that every line contains $q + 1$ points. Now consider the case when $q = 2$. Here the finite projective space $PG(3, 2)$ contains 15 points and every line contains 3 points. Sound familiar?

A solution to Kirkman's famous problem could be obtained with lines of $PG(3, 2)$. A solution would go something like this. First you would have to partition the projective space into lines. Such a partition of the points of $PG(3, q)$ into lines is called a *spread* by finite geometers. A spread in our setting would contain 5 disjoint lines (each containing 3 points). The points of our projective space would correspond to the girls, and the lines of our spread would correspond to the groups of girls walking together on the first day. To find the groups for the second day would require us to find a second spread such that no line from the first spread gets reused in the second spread. Then we continue in this fashion until we get 7 pairwise disjoint spreads (or 7 days' worth of partitions). Seems possible, I suppose. But are we satisfying the condition that no two girls walk together more than once? If this were not the case, then we would have two points of the projective space lying on two different lines. Recall that this violates the axiom for projective geometry requiring that every two distinct points determine a unique line. So, the geometric model actually guarantees us the desired property.

To solve Kirkman's problem, we would need 7 pairwise disjoint spreads (no two sharing a common line). Hence, we would need to use 35 different lines of the projective space. Do we have enough? Just as we counted points, we can easily count lines. Any two independent vectors would determine a 2-dimensional subspace. There are $q^4 - 1$ choices for a first vector and then $q^4 - q$ choices for a second vector that is independent from the first. Any particular 2-dimensional subspace will be counted by any pair of independent vectors in that subspace. Hence, the total number of 2-dimensional subspaces (that is, the number of lines of $PG(3, q)$) is

$$\frac{(q^4 - 1)(q^4 - q)}{(q^2 - 1)(q^2 - q)} = (q^2 + 1)(q^2 + q + 1).$$

Plugging in $q = 2$ gives us 35, precisely the number of lines we need.

Now, let's get back to the Kirkman solution in $PG(3, 2)$. In geometric terms, we are trying to partition the lines of $PG(3, 2)$ into 7 disjoint spreads. Such a partition of lines into spreads is known as a *packing*. It is fairly well-known that spreads and packings exist. Hirschfeld [**13**] gives details about how to actually construct such packings and even shows that they exist for projective 3-spaces of *any* order (that is, any value of q). The trick is to model $PG(3, q)$ not using a vector space, but rather using the finite field $GF(q^4)$. Then, subfields isomorphic to $GF(q^2)$ correspond to lines and some algebra can be used to show the existence of the spreads that we need. In general, the projective space $PG(3, q)$ contains $(q^2 + 1)(q^2 + q + 1)$ lines and a packing of $PG(3, q)$ is comprised of $q^2 + q + 1$ spreads, each of size $q^2 + 1$. Hence, packings of $PG(3, q)$ actually give a solution to a generalized Kirkman Schoolgirls problem:

> If $(q^2 + 1)(q + 1)$ schoolgirls go walking each day in $q^2 + 1$ rows of $q + 1$, they can walk for $q^2 + q + 1$ days so that each girl has walked in the same row as has every other girl and hence with no girl twice.

Incidentally, finite geometry provides a wealth of examples of designs, and Kirkman designs are no exception. By generalizing the spreads and packings described above,

one can construct resolvable $(3n, 3, 1)$ designs for many values of n simply by varying the dimension of the space you work in. A very thorough, albeit technical, description of these methods can be found in the book by Hirschfeld [**13**].

Let us represent the nonzero 4-bit strings (the projective points of $PG(3, 2)$) by the decimal integers they represent: $1 = 0001, 2 = 0010, \ldots, 10 = 1010, \ldots, 15 = 1111$. Then TABLE 4 shows a packing of the lines of $PG(3, 2)$ into 7 disjoint spreads, a solution to Kirkman's Schoolgirls Problem.

TABLE 4: The Kirkman design as a spread in $PG(3, 2)$

MON	TUE	WED	THU	FRI	SAT	SUN
1, 2, 3	1, 4, 5	2, 4, 6	1, 6, 7	3, 4, 7	3, 5, 6	2, 5, 7
4, 10, 14	2, 13, 15	1, 8, 9	2, 9, 11	2, 12, 14	2, 8, 10	1, 14, 15
7, 8, 15	3, 9, 10	3, 12, 15	4, 8, 12	1, 10, 11	4, 11, 15	4, 9, 13
5, 9, 12	6, 8, 14	5, 11, 14	3, 13, 14	5, 8, 13	1, 12, 13	3, 8, 11
6, 11, 13	7, 11, 12	7, 10, 13	5, 10, 15	6, 9, 15	7, 9, 14	6, 10, 12

Notice that the seven blocks in the first row make up a $(7, 3, 1)$-design. This is no coincidence. Since lines of a spread cannot intersect, and every pair of lines in a projective plane must intersect, it follows that the set of lines of a $PG(2, 2)$ inside our $PG(3, 2)$ must all lie in different spreads (that is, different columns of our table). The points across the top lie in the projective plane (isomorphic to $PG(2, 2)$) that is obtained by looking at all projective points of $PG(3, 2)$ whose first homogeneous coordinate is 0. You can verify that the set of such vectors forms a 3-dimensional vector space over $GF(2)$ and therefore serves as a model for the projective plane $PG(2, 2)$. As an exercise, consider the projective plane (isomorphic to $PG(2, 2)$) obtained from the points whose *last* homogeneous coordinate is 0. These points are represented by the even integers. Verify that each of the lines contained in this plane lies in a different column of our table. In other words, each column contains exactly one entry composed of all even integers.

Now, we have two ways to describe Kirkman's Fifteen Young Ladies: the spreads in $PG(3, 2)$ and the subfields of an algebraic number field. By a $KS(15)$, we mean a resolvable $(15, 3, 1)$ block design. As we have seen, the quadratic (varieties) and biquadratic (blocks) subfields of the degree-16 algebraic number field $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$ form a $KS(15)$.

But wait—there's more. The points in $PG(3, 2)$ are 4-digit bit strings. There is a one-to-one correspondence (prove it!) between the points in $PG(3, 2)$ and the quadratic subfields of K, defined by mapping the nonzero bit string $b_1b_2b_3b_4$ to the quadratic subfield $\mathbb{Q}(\sqrt{2^{b_1}3^{b_2}5^{b_3}7^{b_4}})$. Let's see how this correspondence acts on a KS(15) design.

A set of three points in $PG(3, 2)$, such as $\{0110, 1101, 1011\}$, are collinear provided their vector sum over the 2-element field $GF(2)$ is the zero vector. This follows from the fact that projective lines are defined as 2-dimensional subspaces. Recall that a "line" or "block" in the extension-field version of the $KS(15)$ design is a set of three quadratic subfields $\{\mathbb{Q}(\sqrt{p}), \mathbb{Q}(\sqrt{q}), \mathbb{Q}(\sqrt{r})\}$ that belong to the same biquadratic subfield of K. We would like to find conditions on p, q, and r that force this to occur. Necessarily, we would require $\sqrt{r} \in \mathbb{Q}(\sqrt{p}, \sqrt{q})$, $\sqrt{p} \in \mathbb{Q}(\sqrt{q}, \sqrt{r})$, and $\sqrt{q} \in \mathbb{Q}(\sqrt{p}, \sqrt{r})$. In other words, the biquadratic subfields determined by any two of $\{p, q, r\}$ are all the same. This boils down to a fairly simple condition on the variables. The condition that the subfields $\{\mathbb{Q}(\sqrt{p}), \mathbb{Q}(\sqrt{q}), \mathbb{Q}(\sqrt{r})\}$ lie in a common biquadratic subfield is that p, q, and r are nonsquare integers whose product

is a square. For instance, the three quadratic subfields $\{\mathbb{Q}(\sqrt{15}), \mathbb{Q}(\sqrt{42}), \mathbb{Q}(\sqrt{70})\}$ lie in a unique biquadratic subfield since $15 \cdot 42 \cdot 70 = 44100 = 210^2$ is a square. A little algebra shows that three 4-dimensional vectors $\{a, b, c\}$ representing points in $PG(3, 2)$ sum to zero mod 2 if and only if their corresponding quadratic subfields $\{\mathbb{Q}(\sqrt{p}), \mathbb{Q}(\sqrt{q}), \mathbb{Q}(\sqrt{r})\}$ have the property that pqr is a square. To see this, note that a vector sum of zero for $a \oplus b \oplus c$ translates to an even number of 1s summed in each of the four coordinate positions. In the field model, this means that each exponent of $2^{b_1}3^{b_2}5^{b_3}7^{b_4}$ is even and so the entire product is a square. Thus, the correspondence preserves lines, and so is an isomorphism between the set of 35 lines of a $KS(15)$ in $PG(3, 2)$ and its corresponding $KS(15)$ in K.

At this point you probably will not be surprised that Kirkman's famous design arises in two more seemingly unrelated areas of mathematics, namely recreational mathematics and the very applied area of coding theory. So, let's talk about Kirkman's design and how it relates to a certain guessing game involving fifteen players wearing hats.

Fifteen schoolgirls, fifteen hats

Here is a famous problem in recreational mathematics that we'll call the Three Hats Game. Three players enter a room and a maroon or orange hat is placed on each person's head. The color of each hat is determined by a coin toss, with the outcome of one coin toss having no effect on the others. Each person can see the other players' hats but not his own.

No communication of any sort is allowed, except for an initial strategy session before the game begins. Once players have had a chance to look at the other hats, they must simultaneously guess the color of their own hats or pass. The group shares a hypothetical $3 million prize if at least one player guesses correctly and no players guess incorrectly. The problem is to find a strategy whereby the group's chance of winning exceeds 50%.

Mathematicians credit the Three Hats Game to Todd Ebert, a computer scientist, who introduced it in his Ph.D. thesis in 1998 [**9**]. The problem was then popularized by an April 2001 article in the *New York Times* [**18**].

The winning strategy is as follows. Each player looks at the other two hats. A player who sees two of the same color guesses the *missing* color. A player who sees two different colors passes. Now there are eight ways of distributing hats of two colors among three distinct players. In six of these ways, two players see hats of different colors and they pass; the third player sees two hats of the same color, guesses the missing color—and that turns out to be a win. In the other two cases, all hats are the same color; each player guesses the missing color, and all three are wrong. Hence, the strategy works in six of eight cases, and so the three players will win 3/4 of the time. This comes as a surprise to most readers.

We will see how this technique generalizes, with increasingly better odds, to any number of players of the form $2^n - 1$ for $n \geq 3$. In particular, it generalizes to a situation where there are $2^4 - 1 = 15$ players—maybe even fifteen schoolgirls—and the analysis involves a mathematical middle-man known as a Hamming Code. So, before we describe the general technique, let's talk about error-correcting codes.

Some coding theory Mathematical schemes to deal with signal errors first appeared in the 1940s in the work of several researchers, including Claude Shannon, Richard Hamming, and Marcel Golay. These researchers saw the need for something that would automatically detect and correct errors in signal transmissions across noisy channels. What they came up with was a new branch of mathematics called *coding*

theory—specifically, the study of error-detecting and error-correcting codes. They modeled these signals as sets of n-long strings called *blocks*, to be taken from a fixed alphabet of size q; a particular set of such blocks, or *codewords*, is called a *q-ary code of length n*. If q is a prime number, then a q-ary code of length n is called *linear* if the code words form a subspace of $\mathbb{Z}_q^n$, the n-dimensional vector space over $\mathbb{Z}_q$, the integers mod q. A basis for such a linear code is called a *generating set* for the code. One way to describe such a set is with a *generator matrix*, which is a q-ary matrix of n columns whose rows generate the code.

To *detect* errors means to determine that a codeword was incorrectly received; to *correct* errors means to determine the right codeword in case it *was* incorrectly received. Just how this correction happens will vary from code to code.

The fact that d errors in transmission change d characters in a block gives rise to the idea of distance between blocks. If v and w are n-blocks, then the *(Hamming) distance* $D(v, w)$ is the number of positions in which v and w differ. Thus, $D(11001, 10101) = 2$ and $D(1101000, 0011010) = 4$. If I send the block v and you receive the block w, then $D(v, w)$ errors occurred while sending v.

It follows that if the words in a code are all "far apart" in the Hamming distance sense, they can detect errors. Even better, if we assume that only a few errors occur, then we can sometimes change the received block to the correct codeword. Let us now look at an example of an error-correction scheme.

One way to transmit bits is to send each bit three times, so that our only codewords are 000 and 111. If you receive 010, then it is most likely that I sent 000 and so the intended message was 0; this is the triplication or majority-vote code and will successfully correct a single error. Thus, a codeword of length n contains a certain number k of *message bits*, and the other $n - k$ *check bits* are used for error detection and correction. Such a code is called an (n, k) code: the triplication code is a $(3, 1)$ code.

The *minimum distance* of a code is the smallest distance between its codewords. This minimum distance determines the code's error detection and correction features. (Exercise: Show that a code with minimum distance 5 will detect up to 4 errors and correct up to 2. You can then show that a code with minimum distance d will detect up to $d - 1$ errors and correct up to $\lfloor (d - 1)/2 \rfloor$ errors.) For an (n, k) code to be efficient, the ratio k/n should be as large as possible, consistent with its error detection and correction capabilities. Maximum efficiency in an (n, k) m-error correcting code occurs when it can correct up to m errors, and no others. Such a code is called *perfect*.

Hamming's first error correcting scheme was a perfect 1-error correcting code of length seven with four message bits, three check bits, and minimum distance 3; hence, it could correct all errors in which a single bit was received incorrectly. Golay extended Hamming's work and constructed a family of $(2^n - 1, 2^n - 1 - n)$ linear binary perfect 1-error correcting codes of minimum distance 3 for all $n \geq 2$. These are now known as the Hamming codes, and they include both Hamming's original $(7, 4)$ code and the $(3, 1)$ triplication code.

Here is the connection between the Kirkman Schoolgirls Problem and Hamming codes. As we have seen, the 35 triples in TABLE 4 are the 35 blocks of a resolvable $(15, 3, 1)$-design, and the numbers $1, \ldots, 15$ are the varieties. The incidence matrix M for this design is a 35×15 matrix of zeros and ones. It is straightforward to show that the *row space* of M—that is, the vector space generated by the rows of M—is an 11-dimensional subspace of $\mathbb{Z}_2^{15}$, and that this subspace is a $(15, 11)$ Hamming code.

We now show how Hamming codes are the keys to understanding the winning strategy for the Three Hats Game, and how the Kirkman Schoolgirls problem is linked to the Fifteen Hats Game.

Fifteen schoolgirls, fifteen hats: A solution. Go back and look at the Three Hats Game again. Notice that the triplication code contains two codewords and six blocks with errors. The six erroneous blocks correspond to the six winning hat placements for the three players, and the two codewords correspond to the two losing hat placements. As we see in what follows, that is not an accident.

Here is how a solution to the Kirkman Schoolgirls problem leads to a solution to the Fifteen Hats Game in which the probability of winning is much greater than 50%: in fact, it is well over 90%.

First, we number the girls from 1 to 15 in the same way that they are labeled in TABLE 4. We think of these as 4-digit nonzero binary numbers.

Now, suppose that the girls enter the room, each obtaining a hat, and circle up in order 1 through 15. Each player now does the following. She looks at the numbers corresponding to each girl wearing a maroon hat, and she computes the corresponding vector sum. For example, if girls 1, 3, 5, 8, 10, 12, and 14 are wearing maroon hats, then girl 4 will compute $1 \oplus 3 \oplus 5 \oplus 8 \oplus 10 \oplus 12 \oplus 14$. As a mod-2 vector sum, this is

$$0001 \oplus 0011 \oplus 0101 \oplus 1000 \oplus 1010 \oplus 1100 \oplus 1110 = 0111, \text{ or } 7.$$

1. If that sum is equal to her number, she guesses that her color is orange.
2. If that sum is equal to zero, she guesses that her color is maroon.
3. If neither of these two situations occurs, she passes.

Let's analyze what happens. First suppose that the sequence of all maroon hats corresponds to a vector sum of 0. Then every schoolgirl falls into one of the first two cases. All of them will guess incorrectly, and the team loses. More precisely, if a particular girl has on a maroon hat, the corresponding sum that she computes will be equal to her number. So, she will fall into case 1 above and will therefore guess that her hat is orange. Wrong! A similar mistake occurs if the girl is wearing orange.

Next, suppose that the sequence of all maroon hats corresponds to a vector sum of $n \neq 0$. Girl k sees a vector sum of $n \oplus k$ or n, according as she is wearing maroon or orange, respectively. If $k \neq n$, then what Girl k sees is neither her own number nor zero, so she passes. Girl n, however, sees 0 if she is wearing maroon and sees n if she is wearing orange; in both of these cases, she guesses correctly and the team wins.

In the previous example, in which the sequence of all maroon hats corresponds to a vector sum of $7 \neq 0$, the only girl to see either 0 or her own number is Girl 7, who sees 7. That is her own number so she correctly guesses that her hat is orange, and the team wins.

As an exercise, suppose that girls 1, 4, 6, 8, 9, 10, and 12 are wearing maroon hats, and the others are wearing orange hats. Is this a winning configuration, and if so, which girl makes the correct guess? The solution is at the end of the next section.

Why does this work? This is where Hamming codes point the way. The reason is that the configurations of maroon hats with vector sums of 0 are in one-to-one correspondence with the binary vectors of length 15 in the row space of M, the incidence matrix of the Kirkman (15, 3, 1)-design, and as previously mentioned, this row space forms a (15, 11) Hamming code. Recall that the Hamming codes are perfect codes with minimum distance 3. This means that every vector in the entire vector space $\mathbb{Z}_2^{15}$ either (a) is a codeword, or (b) differs in one coordinate from a unique Hamming codeword. That is, changing just one special coordinate position of a vector that is *not* a codeword will leave us with a codeword. Thus, in an arrangement of hats not corresponding to a codeword, the only one who can detect this is the girl who occupies that

special coordinate position. She can tell what her hat color should be in order to make the entire configuration a codeword—and so she guesses the opposite color.

As for the probability of winning with this strategy, it is 15/16, and here is why: We have seen that the triples corresponding to the Kirkman Schoolgirls problem generate a vector space, the row space of the incidence matrix M, that corresponds to the (15, 11) Hamming code. The incorrect guesses will occur exactly when the arrangements of maroon hats correspond to a vector in the Hamming code. Hence, the probability that the players lose the game is given by the size of the Hamming code divided by the total number of $\mathbb{Z}_2$-vectors of length 15. This gives us $2^{11}/2^{15} = 1/16$. So the chances of winning are actually $1 - 1/16$, or 15/16. We hope you find this as surprising as we do. By increasing the number of players, you actually *increase* your chances of winning.

As for the Three Hats Game, the triplication code is a (3, 1) Hamming code. Its generator matrix is [1 1 1], the set of codewords is {000, 111} and there are 8 binary vectors of length 3. Hence, the probability of a win is $1 - 1/4$, or 3/4.

With that, we leave Thomas Kirkman and his fifteen schoolgirls, whose simple arrangement question has led us into many varied areas of mathematics. Hats off to all fifteen of you!

Questions

Where can I find out more about Kirkman designs and block designs in general? One of the best places to begin is Chapter 6 of Kenneth Bogart's beautifully written book [**3**], which will take you a fair way into the subject. Two others are Marshall Hall's classic [**12**] and the more technical book by Beth, Jungnickel, and Lenz [**1**], both of which are excellent and will take you as far as you want to go.

Is the Kirkman design found in $PG(3, 2)$ the only solution to Kirkman's Schoolgirls Problem? We say that two block designs are *isomorphic* if there is a 1-1 correspondence between the two sets of varieties that is also a 1-1 correspondence between the two sets of blocks. It was known for a long time that there are eighty nonisomorphic (15, 3, 1)-designs. In 1922, F. N. Cole [**7**] proved that only four of these eighty designs are resolvable. Cole also proved that three of these have two nonisomorphic resolutions, while the fourth has only one. (Exercise: Determine whether the (15, 3, 1)-design presented in this paper has a resolution not isomorphic to the one in TABLE 2.)

For which values of v do resolvable $(v, 3, 1)$-designs exist? This question dates back to Kirkman himself [**15, 16**] and was open for over a hundred years. Finally, in 1971 D. K. Ray-Choudhury and R. M. Wilson proved that resolvable $(v, 3, 1)$-designs exist if and only if $v \equiv 3 \bmod 6$ [**17**] .

Are there Kirkman designs in number fields other than the degree-16 field described above? Yes. Let $n > 3$ and let $p_1, p_2, \ldots, p_n$ be distinct primes. The field $L_n = \mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_n})$ is an extension of degree 2^n over the rational numbers $\mathbb{Q}$. In such fields, the quadratic and the biquadratic subfields of L_n are the varieties and blocks, respectively, of a resolvable $(2^n - 1, 3, 1)$-design. As an exercise, show that such a design contains $b = (2^n - 1)(2^{n-1} - 1)/3$ blocks, and each variety appears in $r = 2^{n-1} - 1$ blocks. A more challenging exercise is to show that these designs are resolvable.

The set of 15 Schoolgirls contains 455 3-element subsets, or trios. Suppose the school term is 13 weeks long. What if the Schoolgirls wanted to arrange 13 weeks' worth of walks so that each trio of girls can walk together exactly once during the term? They can do it. Note that this amounts to partitioning the 455 trios into 13

distinct Kirkman (15, 3, 1)-designs. Evidently, in 1850 Cayley referred to Kirkman's original problem as well as to Sylvester's extension to 13 walks. In 1974, R. H. F. Denniston briefly discussed the problem's history, and then presented a solution [**8**]. As an exercise, find a partition of the 84 3-element subsets of $\{1, \ldots, 9\}$ into seven resolvable (9, 3, 1)-designs. Happy walking!

The Schoolgirl Problem connects block designs, finite projective geometries, algebraic number fields, error-correcting codes, and recreational mathematics. Are there any other connections? Yes, there is at least one more connection. The set $G_K = (\mathbb{Z}/2\mathbb{Z})^4 = \{(a, b, c, d) : a, b, c, d \in 0, 1\}$ is a group under the operation of coordinate-wise addition mod 2. This group, $(\mathbb{Z}/2\mathbb{Z})^4$, has 15 subgroups of order 2, 35 subgroups of order 4 and 15 subgroups of order 8; each order-2 subgroup is contained in three order-4 subgroups and seven order-8 subgroups. (Does this sound familiar?) In fact, G_K is what is known as the *Galois group* of the degree-16 field $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7})$. It is the group of isomorphisms of K to itself that leaves $\mathbb{Q}$ fixed. There is a one-to-one, order-reversing correspondence between the subfields of K and the subgroups of G_K, and the details of this correspondence are laid out in the Fundamental Theorem of Galois Theory, one of the most beautiful theorems in mathematics.

What about the solution to that exercise? We know that girls 1, 4, 6, 8, 9, 10, and 12 are wearing maroon hats, and the others are wearing orange hats. The sequence of all maroon hats yields the vector sum $1 \oplus 4 \oplus 6 \oplus 8 \oplus 9 \oplus 10 \oplus 12$, that is,

$$0001 \oplus 0100 \oplus 0110 \oplus 1000 \oplus 1001 \oplus 1010 \oplus 1100 = 0100, \text{ or } 4.$$

Girl k sees $k \oplus 4$, and the only one with a winning view is Girl 4, who sees the all-zeros vector. Therefore, she guesses maroon, nobody else guesses, and the team wins.

REFERENCES

1. T. Beth, D. Jungnickel and H. Lenz, *Design Theory*, 2nd ed., Cambridge University Press, 1999.
2. N. L. Biggs, T. P. Kirkman, mathematician, *Bull. London Math. Soc.* **13** (1981) 97–120.
3. K. P. Bogart, *Introductory Combinatorics*, 3rd ed., Harcourt Academic Press, 2000.
4. R. C. Bose, On the construction of balanced incomplete block designs, *Ann. Eugenics* **9** (1939) 353–399.
5. E. Brown, The many names of (7, 3, 1), this MAGAZINE **75** (2002) 83–94.
6. A. Cayley, On the triadic arrangements of seven and fifteen things, *Philos. Mag.* **37**(3) (1850) 50–53.
7. F. N. Cole, Kirkman parades, *Bull. Amer. Math. Soc.* **28** (1922) 435–437.
8. R. H. F. Denniston, Sylvester's problem of the 15 schoolgirls, *Discrete Math.* **9** (1974) 229–233.
9. T. Ebert, Applications of recursive operators to randomness and complexity. Ph.D. Thesis, University of California at Santa Barbara, 1998.
10. R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.
11. R. A. Fisher, An examination of the different possible solutions of a problem in incomplete blocks, *Ann. Eugenics* **10** (1940), 52–57.
12. M. Hall, Jr., *Combinatorial Theory*, 9th ed., Blaisdell Publishing Company, 1967.
13. J. W. P. Hirschfeld, *Projective Geometries over Finite Fields*, Oxford University Press, 1998.
14. T. A. Kirkman, On a problem in combinations, *Camb. Dublin Math. J.* **2** (1847) 191–204.
15. ———, Query 6, *Lady's and Gentlemen's Diary* (1850) 48.
16. ———, Note on an unanswered prize question, *Camb. Dublin Math. J.* **5** (1850) 255–262.
17. D. K. Ray-Chaudhury and R. M. Wilson, Solution of Kirkman's schoolgirl problem, in *Combinatorics*, Proc. Sympos. Pure Math., vol. 19, AMS, 1971, pp. 187–203.
18. S. Robinson, Why mathematicians now care about their hat color, *The New York Times*, 10 April 2001.
19. J. J. Sylvester, Note on the historical origin of the unsymmetrical six-valued function of six letters, *Philos. Mag.* **21**(4) (1861) 369–377.
20. J. J. Sylvester, Note on a nine schoolgirls problem, *Messenger of Mathematics* **22** (1893) 159–160.
21. W. S. B. Woolhouse, Prize question 1733, *Lady's and Gentleman's Diary*, 1844.

When Euler Met l'Hôpital

WILLIAM DUNHAM
Muhlenberg College
Allentown, PA 18104
wdunham@muhlenberg.edu

We begin with a disclaimer: Leonhard Euler, born in Basel in 1707, never actually met the multiply-named Guillaume François Antoine, Marquis de l'Hôpital, who had died in Paris three years before. The title of this article refers to a meeting of minds, not of mathematicians.

Figure 1 Portraits of l'Hôpital and Euler

Chronological details notwithstanding, the connection between these two individuals is real. For one thing, l'Hôpital published the first text on differential calculus, his *Analyse des infiniment petits*, in 1696. This was justifiably regarded as the standard exposition, a status not relinquished until Euler's treatment of the same topic in his masterful *Institutiones calculi differentialis* of 1755, shown in FIGURE 2.

More germane is Euler's discussion, in that 1755 work, of indeterminate forms. There, he presented various methods to attack the 0/0 problem, from using simple algebra and trigonometry, to introducing infinitely small quantities, to applying that most sophisticated of techniques, l'Hôpital's rule. Because it had originally appeared in l'Hôpital's text, the rule now carries his name, but, as is widely known, it had been discovered as early as 1694 by Johann Bernoulli (1667–1748). At the time, Johann was employed by the Marquis to provide lectures on the emerging subject of calculus. In a letter to Bernoulli, l'Hôpital described their financial arrangement [**11**]:

> I shall give you with pleasure a pension of three hundred livres ... I ask you to give me occasionally some hours of your time to work on what I shall ask you—and also to communicate to me your discoveries, with the request not to mention them to others.

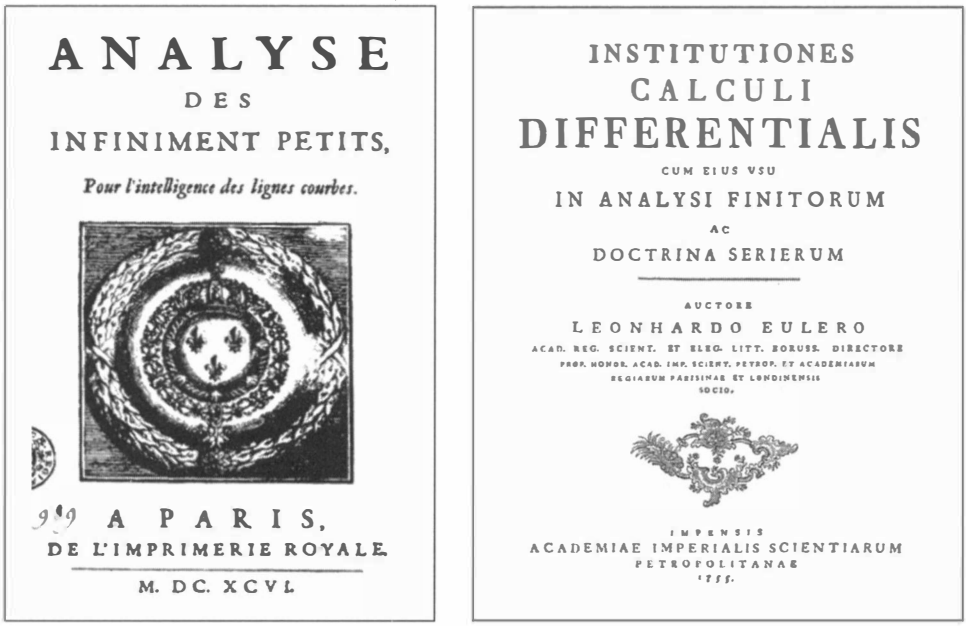


Figure 2 Frontispieces of l'Hôpital and Euler Texts

L'Hôpital acknowledged his debt in the introduction to the *Analyse des infiniment petits* by saying he had made "free use" of the discoveries of others and would happily "return to them whatever they please to claim as their own" [7]. Upon l'Hôpital's death, Bernoulli did indeed claim the rule, but somehow posterity never returned it. To those who bemoan the injustice of this situation, the historian of mathematics Dirk Struik had a tart response: "Let the good Marquis keep his elegant rule; he paid for it" [11].

Euler addressed these topics in Chapter 15 of his differential calculus text, a chapter titled, "On the values of functions which in certain cases seem indeterminate" [6]. The purpose of this article is to examine the highlights of that discussion.

We begin with Euler's statement and proof of the result. Then, after a few well-chosen examples, we consider his clever use of the rule to derive the summation formula for the first n whole numbers:

$$1 + 2 + 3 + \cdots + n = n(n + 1)/2.$$

Even more improbably, he applied l'Hôpital's rule to evaluate the *infinite* series

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{n^2} + \cdots.$$

This, of course, is the famous "Basel Problem." Euler had earned mathematical immortality in the mid-1730s by determining the sum to be $\pi^2/6$. Over the course of his career, he gave at least two other derivations of the result [3], but his evaluation via l'Hôpital's rule seems to be less well known. We hope it will provide a fitting conclusion to this paper.

The rule and its proof

Euler began by introducing the quotient $y = P/Q$, where P and Q were functions of x. For the value $x = a$, Euler assumed that both numerator and denominator vanished,

thereby reducing the expression to the form 0/0. In such a case, he acknowledged that the fraction would "seem indeterminate," but, in fact, its value might yet be found.

To illustrate, he considered $y = (a^2 - x^2)/(a - x)$, which for $x = a$ obviously took the form 0/0. Euler divided numerator by denominator to get $y = a + x$, which, for $x = a$, became $y = 2a$. "In this case," he wrote, "the fraction 0/0 is equal to the quantity $2a$."

Here the indeterminate was, in fact, determined, but this example was so simple that calculus was unnecessary. Euler intended to set his sights on more complicated quotients and thus would need the heavy mathematical artillery embodied in l'Hôpital's rule.

With no explicit mention of l'Hôpital (nor of Bernoulli!) Euler introduced the rule as follows: if $y = P/Q$ where $P(a) = Q(a) = 0$, then, for $x = a$, "...the fraction dP/dQ takes the value of the fraction P/Q in question."

Before considering his proof, we emphasize that it pre-dated the modern era. Euler's analysis was not our analysis, as will be immediately clear. Indeed, today's reader might be uneasy about the very *statement* of the rule. At its heart were the differentials dP and dQ rather than the corresponding derivatives—a reminder that in the 18th century, the notion of *differential* calculus was taken literally. More significantly, neither the statement nor proof mentioned limits, a 19th century innovation. Where we now use l'Hôpital's rule to evaluate

$$\lim_{x \to a} \frac{P(x)}{Q(x)},$$

Euler instead wanted to find the "value" of the quotient $P(x)/Q(x)$ when $x = a$.

So, to determine $y = (a^2 - x^2)/(a - x)$ when $x = a$, our predecessors would find differentials top and bottom to get $(-2x\,dx)/(-dx) = 2x$ and conclude that the fraction $(a^2 - x^2)/(a - x)$ takes the "value" of $2a$ when $x = a$. (Of course, in all cases the differential dx will cancel as it did here, leaving us with $P'(x)/Q'(x)$, the familiar quotient of derivatives.)

As was characteristic of the time, Euler's proof of l'Hôpital's rule rested upon the notion of the infinitely small. Here and elsewhere, he used infinitely small quantities without hesitation and had an uncanny—although perhaps "insightful" is a better word—understanding of how to exploit them in mathematical arguments. Euler had discussed these quantities at some length in the third chapter of his differential calculus text, where he explained why "...a finite quantity can neither be increased nor decreased by adding or subtracting an infinitely small quantity" [4].

Moreover, if $y = y(x)$ and we augment a by the infinitely small amount dx, then y will change by the infinitely small amount dy where $dy = y(a + dx) - y(a)$ as illustrated in FIGURE 3. It follows that

$$y(a + dx) = y(a) + dy. \tag{1}$$

We now consider Euler's proof of l'Hôpital's rule for evaluating $P(x)/Q(x)$ when $P(a) = Q(a) = 0$. As he had noted, the replacement of a by $a + dx$ changed nothing, and so he considered the fraction $P(a + dx)/Q(a + dx)$ and applied (1) to conclude that

$$\frac{P(a + dx)}{Q(a + dx)} = \frac{P(a) + dP}{Q(a) + dQ} = \frac{0 + dP}{0 + dQ} = \frac{dP}{dQ}.$$

The 0/0 form was thereby reduced to the ratio of two differentials, and this established l'Hôpital's rule.

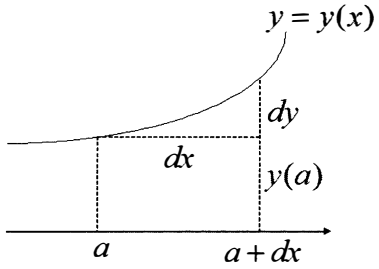


Figure 3 Differentials

Mathematicians of today can lodge a host of objections to this sort of derivation. A modern proof of l'Hôpital's rule, with all the epsilons and deltas intact, is surprisingly complicated (see, for instance, Boas [**1**] or Ross [**10**]) and, in any case, lay well beyond the capability of 18th century analysis.

We should note, however, that demonstrations like Euler's, resting upon infinitesimals, can be recast within the context of hyperreal numbers. In the process, some of the mysterious goings-on surrounding infinitely small quantities turn out to be not quite so mysterious. A paper by McKinzie and Tuckey [**8**] nicely addresses these points.

But the focus of our article is not so much on how Euler proved the rule as on how he used it. In this regard we shall discuss, in turn, some basic examples, the sums of integers, and the Basel Problem.

Four Eulerian examples

As would any textbook author, Euler sought to show l'Hôpital's rule in action. To this end, he included a few routine problems that, if converted to "limit" terminology, would be appropriate in any modern calculus book. For instance, he derived:

$$\frac{a^n - x^n}{\ln a - \ln x} = na^n \text{ when } x = a; \text{ and } \frac{e^x - e^{-x}}{\ln(1+x)} = 2, \text{ when } x = 0.$$

But other examples in the chapter better showcased his mathematical skills. We consider four of these below.

EXAMPLE (A). "Find the value of $(1 - \sin x + \cos x)/(\sin x + \cos x - 1)$ when $x = \pi/2$."

The expression is indeterminate, and an application of l'Hôpital's rule led Euler to $(-\cos x - \sin x)/(\cos x - \sin x)$ which became 1 for $x = \pi/2$. Nothing remarkable here.

But, like the effective expositor that he was, Euler attacked the problem in an entirely different manner. To do so, he invoked the trigonometric identity

$$\cos x = \sqrt{\cos^2 x} = \sqrt{1 - \sin^2 x} = \sqrt{1 + \sin x} \cdot \sqrt{1 - \sin x},$$

which transformed the initial fraction into

$$\begin{aligned}\frac{1 - \sin x + \cos x}{\sin x + \cos x - 1} &= \frac{(1 - \sin x) + \cos x}{\cos x - (1 - \sin x)} \\ &= \frac{\sqrt{1 - \sin x} \cdot \sqrt{1 - \sin x} + \sqrt{1 + \sin x} \cdot \sqrt{1 - \sin x}}{\sqrt{1 + \sin x} \cdot \sqrt{1 - \sin x} - \sqrt{1 - \sin x} \cdot \sqrt{1 - \sin x}} \\ &= \frac{\sqrt{1 - \sin x} + \sqrt{1 + \sin x}}{\sqrt{1 + \sin x} - \sqrt{1 - \sin x}}.\end{aligned}$$

For $x = \pi/2$, the fraction reduced to $(\sqrt{1-1}+\sqrt{1+1})/(\sqrt{1+1}-\sqrt{1-1}) = \sqrt{2}/\sqrt{2} = 1$. This second approach, of course, resulted in the same answer but required only the tools of trigonometry.

EXAMPLE (B). "Find the value of $(e^x - 1 - \ln(1+x))/x^2$ in the case where we place $x = 0$."

For this indeterminate form, Euler's first application of l'Hôpital's rule gave $(e^x - 1/(x+1))/2x$, which was also indeterminate for $x = 0$. A second application was necessary, yielding $(e^x + 1/(x+1)^2)/2$. If $x = 0$, the fraction is 1.

Again, Euler provided an alternative derivation, this one involving differentials along with the series expansions

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad \text{and} \quad \ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots.$$

Substituting the infinitely small dx into the original fraction and using the series above, Euler reasoned that the indeterminate form was equal to

$$\frac{e^{dx} - 1 - \ln(1+dx)}{(dx)^2}$$

$$= \frac{\left[1 + dx + \frac{(dx)^2}{2} + \frac{(dx)^3}{6} + \frac{(dx)^4}{24} + \cdots\right] - 1 - \left[dx - \frac{(dx)^2}{2} + \frac{(dx)^3}{3} - \frac{(dx)^4}{4} + \cdots\right]}{(dx)^2}$$

$$= \frac{(dx)^2 - \frac{(dx)^3}{6} + \frac{7(dx)^4}{24} - \cdots}{(dx)^2} = 1 - \frac{dx}{6} + \frac{7(dx)^2}{24} - \cdots = 1.$$

The last step followed because dx was infinitely small, and so all terms after the first were insignificant compared to the finite quantity, 1. As before, this was the answer obtained via l'Hôpital's rule.

EXAMPLE (C). "Find the value of the expression $(x^x - x)/(1 - x + \ln(x))$ when we place $x = 1$."

What makes this indeterminate form noteworthy—and a splendid example for a modern calculus course—is the appearance of logarithmic differentiation in the solution. Applying l'Hôpital's rule to the quotient, Euler got

$$\frac{(x^x(1+\ln x) - 1)}{(-1 + 1/x)}.$$

This remained indeterminate, so a second application gave

$$\frac{(x^{x-1} + x^x(1+\ln x)^2)}{(-1/x^2)} = -2$$

for $x = 1$. (For most students, an answer of -2 emerging from the expression $(x^x - x)/(1 - x + \ln(x))$ is utterly unexpected.)

EXAMPLE (D). "Find the value of the fraction $x^n/e^{-1/x}$ in the case that $x = 0$."

In modern terms, Euler wanted

$$\lim_{x \to 0^+} \frac{x^n}{e^{-1/x}}$$

for n a whole number. He began by noting that "both the numerator and denominator vanish" when $x = 0$. For algebraic reasons, he introduced $s = x^n/e^{-1/x}$. L'Hôpital's

rule led him to evaluate

$$s = nx^{n-1}/(e^{-1/x} \cdot x^{-2}) = nx^{n+1}/e^{-1/x} \text{ for } x = 0.$$

Here, rather than reducing the degree of the numerator, the process had increased it, a phenomenon Euler characterized as "a misfortune." It seemed to lead away from, not towards, a solution. He had to devise a new strategy.

From $s = x^n/e^{-1/x}$, Euler deduced that

$$x^n = s \cdot e^{-1/x} \text{ and so } (x^n)^{n+1} = s^{n+1} \cdot e^{-(n+1)/x}.$$

In like manner, $s = nx^{n+1}/e^{-1/x}$ implied that

$$x^{n+1} = (s \cdot e^{-1/x})/n \text{ and so } (x^{n+1})^n = (s^n \cdot e^{-n/x})/n^n.$$

Equating these two expressions for $x^{n(n+1)}$, Euler concluded that

$$s^{n+1} \cdot e^{-(n+1)/x} = (s^n \cdot e^{-n/x})/n^n, \text{ and so } s = 1/(n^n e^{-1/x}).$$

Only then did he let $x = 0$ to answer the original question: $s = \infty$.

The answer is correct, although the method certainly seems peculiar to the modern eye. But Euler was Euler, and nobody flung symbols across the page with such zest.

Integer sums

The preceding examples, although intriguing, were concocted to show l'Hôpital's rule at work. More challenging was to apply the rule to establish a result of independent significance. This Euler did when he summed the first n whole numbers via l'Hôpital.

THEOREM. If n is a whole number, then $1 + 2 + 3 + \cdots + n = n(n + 1)/2$.

Proof. To begin, Euler recalled the summation formula for a finite geometric series, i.e., $x + x^2 + x^3 + \cdots + x^n = (x - x^{n+1})/(1 - x)$. He differentiated both sides with respect to x to get $1 + 2x + 3x^2 + \cdots + nx^{n-1} = [1 - (n + 1)x^n + nx^{n+1}]/(1 - x)^2$ and then multiplied through by x to arrive at

$$x + 2x^2 + 3x^3 + \cdots + nx^n = \frac{x - (n + 1)x^{n+1} + nx^{n+2}}{(1 - x)^2}. \tag{2}$$

For $x = 1$, the left-hand side of (2) was simply $1 + 2 + 3 + \cdots + n$, the object of the theorem. But if $x = 1$, the right-hand side reduced to $0/0$ and thus was a ready candidate for l'Hôpital. Euler's first application of the rule yielded

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

which remained indeterminate for $x = 1$. After a second application, he had

$$\frac{-(n + 1)^2 nx^{n-1} + n(n + 2)(n + 1)x^n}{2}.$$

For $x = 1$, this became

$$\frac{-(n + 1)^2 n + n(n + 2)(n + 1)}{2} = \frac{n(n + 1)[-n - 1 + n + 2]}{2} = \frac{n(n + 1)}{2}.$$

By equating the results for the left and right sides of (2), Euler had thus established the summation formula $1 + 2 + 3 + \cdots + n = n(n + 1)/2$. ∎

There are, of course, far easier ways to prove this. Euler's derivation—requiring geometric series, differential calculus, and two doses of l'Hôpital—calls to mind Dr. Johnson's description of "...a dog's walking on his hinder legs. It is not done well; but you are surprised to find it done at all" [9].

Anyone familiar with Euler's methods will not be surprised to learn that he pushed further. Another differentiation of both sides of (2), another wholesale multiplication by x, and yet another application of l'Hôpital for $x = 1$ yielded the summation formula for the squares: $1^2 + 2^2 + 3^2 + \cdots + n^2 = [n(n+1)(2n+1)]/6$. Details are left to the reader, with the caveat that the algebra and calculus get a bit wearisome.

So, Euler could use l'Hôpital to sum these well-known finite series. How he did the same for a famous infinite series is the topic of our last section.

The Basel problem via l'Hôpital

The derivation Euler gave in his differential calculus text rested upon the curious identity

$$\frac{1}{1+x^2}+\frac{1}{4+x^2}+\frac{1}{9+x^2}+\frac{1}{16+x^2}+\cdots=\frac{\pi x-1}{2x^2}+\frac{\pi}{x(e^{2\pi x}-1)}. \tag{3}$$

To understand how he arrived at this, we break the argument into a pair of lemmas, in the process streamlining a few of Euler's steps while retaining the essence of his original reasoning.

LEMMA 1. $\dfrac{1}{1-y^2}+\dfrac{1}{4-y^2}+\dfrac{1}{9-y^2}+\dfrac{1}{16-y^2}+\cdots=\dfrac{1}{2y^2}-\dfrac{\pi\cos\pi y}{2y\sin\pi y}.$

Proof. Euler began with a favorite tactic: expressing the sine of an angle t as the infinite product

$$\sin t = t\left(1-\frac{t}{\pi}\right)\left(1+\frac{t}{\pi}\right)\left(1-\frac{t}{2\pi}\right)\left(1+\frac{t}{2\pi}\right)\left(1-\frac{t}{3\pi}\right)\left(1+\frac{t}{3\pi}\right)\cdots.$$

He justified this by the fact that the equation $\sin t = 0$ has solutions $t = 0, \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$, which generate the corresponding factors on the right side. For $t = \pi y$, this product became

$$\begin{aligned}\sin\pi y &= \pi y(1-y)(1+y)\left(\frac{2-y}{2}\right)\left(\frac{2+y}{2}\right)\left(\frac{3-y}{3}\right)\left(\frac{3+y}{3}\right)\cdots \\ &= \pi y(1-y^2)\left(\frac{4-y^2}{4}\right)\left(\frac{9-y^2}{9}\right)\left(\frac{16-y^2}{16}\right)\cdots. \end{aligned} \tag{4}$$

Euler took logs of both sides of (4) to get

$$\ln(\sin\pi y) = \ln\pi + \ln y + \ln(1-y^2) + \ln(4-y^2) - \ln 4 + \ln(9-y^2) - \ln 9 + \cdots$$

and then differentiated with respect to y to conclude

$$\frac{\pi\cos\pi y}{\sin\pi y} = \frac{1}{y} - \frac{2y}{1-y^2} - \frac{2y}{4-y^2} - \frac{2y}{9-y^2} - \frac{2y}{16-y^2} - \cdots.$$

Lemma 1 follows immediately [6]. ∎

LEMMA 2. $\dfrac{\pi\cos(-i\pi b)}{2ib\sin(-i\pi b)} = \dfrac{\pi}{2b} + \dfrac{\pi}{b(e^{2\pi b}-1)}.$

Proof. This result can be traced back to Euler's *Introductio in analysin infinitorum* of 1748. There he cited it as "... a nice illustration of the reduction of sines and cosines of imaginary arcs to real exponentials" [**5**]. The necessary prerequisites were Euler's formulas for cosine and sine

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \text{ and } \sin z = \frac{e^{iz} - e^{-iz}}{2i},$$

which implied that

$$\frac{\cos z}{\sin z} = \frac{i(e^{iz} + e^{-iz})}{(e^{iz} - e^{-iz})} = \frac{i(e^{2iz} + 1)}{(e^{2iz} - 1)}.$$

Substituting $z = -i\pi b$ in order to introduce the factors that appear in the lemma, we have

$$\frac{\pi \cos(-i\pi b)}{2ib \sin(-i\pi b)} = \frac{\pi}{2ib}\left[\frac{i(e^{2\pi b} + 1)}{e^{2\pi b} - 1}\right] = \frac{\pi}{2b} + \frac{\pi}{b(e^{2\pi b} - 1)},$$

proving the result. ∎

We now combine these results to derive the key series in (3). Euler's idea was to replace y by $-ix$ (and hence y^2 by $-x^2$) in Lemma 1 to get

$$\frac{1}{1+x^2} + \frac{1}{4+x^2} + \frac{1}{9+x^2} + \frac{1}{16+x^2} + \cdots = -\frac{1}{2x^2} + \frac{\pi \cos(-i\pi x)}{2ix \sin(-i\pi x)}.$$

By Lemma 2, the right-hand side is

$$-\frac{1}{2x^2} + \left[\frac{\pi}{2x} + \frac{\pi}{x(e^{2\pi x} - 1)}\right] = \frac{\pi x - 1}{2x^2} + \frac{\pi}{x(e^{2\pi x} - 1)},$$

and so (3) is established. Duncan [**2**, p. 221] gives a modern derivation of the identity within the theory of meromorphic functions.

By now, this dizzying algebraic journey might suggest that Euler has led us into a maze from which escape is hopeless. But the payoff was near, courtesy of l'Hôpital's rule [**6**].

THEOREM. $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{n^2} + \cdots = \frac{\pi^2}{6}$.

Proof. Combining terms on the right side of (3), Euler noted that

$$\begin{aligned}\frac{1}{1+x^2} + \frac{1}{4+x^2} + \frac{1}{9+x^2} + \frac{1}{16+x^2} + \cdots &= \frac{\pi x - 1}{2x^2} + \frac{\pi}{x(e^{2\pi x} - 1)} \\ &= \frac{\pi x e^{2\pi x} - e^{2\pi x} + \pi x + 1}{2x^2 e^{2\pi x} - 2x^2}.\end{aligned}$$

For $x = 0$, the leftmost side was the desired infinite series $1 + 1/4 + 1/9 + \cdots + 1/n^2 + \cdots$, whereas the rightmost fraction became $0/0$. Differentiating numerator and denominator gave

$$\frac{\pi - \pi e^{2\pi x} + 2\pi^2 x e^{2\pi x}}{4x e^{2\pi x} + 4\pi x^2 e^{2\pi x} - 4x},$$

an indeterminate expression when $x = 0$. A second application of l'Hôpital yielded

$$\frac{\pi^3 x e^{2\pi x}}{e^{2\pi x} + 4\pi x e^{2\pi x} + 2\pi^2 x^2 e^{2\pi x} - 1} = \frac{\pi^3 x}{1 + 4\pi x + 2\pi^2 x^2 - e^{-2\pi x}}.$$

Alas, this too was indeterminate, so Euler needed a third application of the rule to get

$$\frac{\pi^3}{4\pi + 4\pi^2 x + 2\pi e^{-2\pi x}},$$

which, for $x = 0$, became $\pi^3/(4\pi + 2\pi) = \pi^2/6$. In this manner, with not one but three helping hands from the Marquis de l'Hôpital, Euler proved that

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{n^2} + \cdots = \frac{\pi^2}{6}. \quad \blacksquare$$

Final thoughts

What a marvelous derivation this was. It boasted an all-star cast of transcendental functions: sines, cosines, logs, and exponentials. It ranged from the real to the complex and back again. And it featured l'Hôpital's rule in a starring role. Of course, none of this would have happened without the fluid imagination of Leonhard Euler, symbol manipulator extraordinaire.

As if that were not sufficient, Euler provided an alternative derivation [**6**]. This time, he employed differentials and began with Lemma 1:

$$\frac{1}{1-y^2} + \frac{1}{4-y^2} + \frac{1}{9-y^2} + \cdots = \frac{1}{2y^2} - \frac{\pi \cos \pi y}{2y \sin \pi y} = \frac{\sin \pi y - \pi y \cos \pi y}{2y^2 \sin \pi y}.$$

As before, for $y = 0$, the left side was $1 + 1/4 + 1/9 + \cdots + 1/n^2 + \cdots$ and the right side was indeterminate. Here, in fact, was another opportunity to invoke l'Hôpital's rule and prove the result (the reader is invited to give it a try).

Euler instead introduced the differential dx in place of y and used series to convert the right-hand side into

$$\begin{aligned}
&\frac{\sin(\pi\, dx) - \pi\, dx \cos(\pi\, dx)}{2(dx)^2 \sin(\pi\, dx)} \\
&= \frac{\left[\pi\, dx - \frac{\pi^3 (dx)^3}{6} + \frac{\pi^5 (dx)^5}{120} - \cdots\right] - \pi\, dx \left[1 - \frac{\pi^2 (dx)^2}{2} + \frac{\pi^4 (dx)^4}{24} - \cdots\right]}{2\,(dx)^2 \left[\pi\,(dx) - \frac{\pi^3 (dx)^3}{6} + \frac{\pi^5 (dx)^5}{120} - \cdots\right]} \\
&= \frac{\frac{\pi^2}{3} - \frac{\pi^4 (dx)^2}{30} + \cdots}{2 - \frac{\pi^2 (dx)^2}{3} + \frac{\pi^4 (dx)^4}{60} - \cdots}.
\end{aligned}$$

But this is just $\pi^2/6$ because the infinitely small dx, in comparison to finite quantities, is indistinguishable from zero. *Voila*!

That such different approaches led to $\pi^2/6$ must have been very satisfying for Euler. To be sure, few mathematicians were more adept at finding multiple paths to the same end. The discussion above illustrates the power—and the charm—of variety. Modern textbook writers should take note.

And modern writers could learn as well from his choice of examples, a choice that delivered more than a routine appreciation of the technique in question. Clever examples can indeed be the icing on our mathematical cake.

It should be evident that here Euler was in his element, operating at the height of his analytic powers. When Euler met l'Hôpital, good things happened.

Acknowledgment. The author thanks Penny Dunham of Muhlenberg College for her many helpful suggestions.

REFERENCES

1. R. Boas, Indeterminate forms revisited, this MAGAZINE **63** (1990) 155–159.
2. J. Duncan, *The Elements of Complex Analysis*, Wiley, 1968.
3. W. Dunham, *Euler: The Master of Us All*, MAA, 1999.
4. L. Euler, *Foundations of Differential Calculus*, trans. John Blanton, Springer-Verlag, 2000.
5. ———, *Introduction to Analysis of the Infinite*, Book I, trans. John Blanton, Springer-Verlag, 1988.
6. ———, *Opera Omnia*, ser. 1, vol. 10, 1913.
7. G. l'Hôpital, *Analyse des infiniment petits* (reprint), ACL-Editions, Paris, 1988.
8. M. McKinzie and C. Tuckey, Higher trigonometry, hyperreal numbers, and Euler's analysis of infinities, this MAGAZINE **74** (2001) 339–368.
9. A. Partington (ed.), *The Oxford Dictionary of Quotations*, Oxford U. Press, 1992.
10. K. Ross, *Elementary Analysis: The Theory of Calculus*, Springer-Verlag, 1980.
11. D. Struik, The origin of l'Hôpital's rule, *Math. Teacher* **56** (1963) 257–260.

Math Bite: A Magic Eight

In many cultures, the number 8 has special significance and is even considered magical. For instance, in China the Olympic Games were scheduled to begin at exactly 8:08 PM on 8/8/08.

Those wishing to advance the cult of the number 8, even in 2009, might enjoy knowing that the sequence

$$\frac{9}{1}, \frac{98}{12}, \frac{987}{123}, \frac{9876}{1234}, \frac{98765}{12345}, \cdots$$

converges exactly to 8.

You probably have realized that the ellipses (...) need a suitable interpretation, since numbers greater than 10 and less than 0 are not normally allowed to serve as digits. Check for yourself that, with a suitable placement of the decimal point, the numerator and denominator become

$$\sum_{k=0}^{\infty} (10 - (k+1)) \left(\frac{1}{10}\right)^k \quad \text{and} \quad \sum_{k=0}^{\infty} (k+1) \left(\frac{1}{10}\right)^k.$$

Summing these series gives 800/81 as the numerator and 100/81 as the denominator. And the magic quotient of these two is 8.

—Paul and Vincent Steinfeld
Darmstadt, Germany

Matroids You Have Known

DAVID L. NEEL
Seattle University
Seattle, Washington 98122
neeld@seattleu.edu

NANCY ANN NEUDAUER
Pacific University
Forest Grove, Oregon 97116
nancy@pacificu.edu

> Anyone who has worked with matroids has come away with the conviction that matroids are one of the richest and most useful ideas of our day.
>
> —Gian Carlo Rota [**10**]

Why matroids?

Have you noticed hidden connections between seemingly unrelated mathematical ideas? Strange that finding roots of polynomials can tell us important things about how to solve certain ordinary differential equations, or that computing a determinant would have anything to do with finding solutions to a linear system of equations. But this is one of the charming features of mathematics—that disparate objects share similar traits. Properties like independence appear in many contexts. Do you find independence everywhere you look? In 1933, three Harvard Junior Fellows unified this recurring theme in mathematics by defining a new mathematical object that they dubbed *matroid* [**4**]. Matroids are everywhere, if only we knew how to look.

What led those junior-fellows to matroids? The same thing that will lead us: Matroids arise from shared behaviors of vector spaces and graphs. We explore this natural motivation for the matroid through two examples and consider how properties of independence surface. We first consider the two matroids arising from these examples, and later introduce three more that are probably less familiar. Delving deeper, we can find matroids in arrangements of hyperplanes, configurations of points, and geometric lattices, if your tastes run in that direction.

While tying together similar structures is important and enlightening, matroids do not reside merely in the halls of pure mathematics; they play an essential role in combinatorial optimization, and we consider their role in two contexts, constructing minimum-weight spanning trees and determining optimal schedules.

What's that, you say? Minimum-weight what? The mathematical details will become clear later, but suppose you move your company into a new office building and your 25 employees need to connect their 25 computers to each other in a network. The cable needed to do this is expensive, so you want to connect them with the least cable possible; this will form a minimum-weight spanning tree, where by *weight* we mean the length of cable needed to connect the computers, by *spanning* we mean that we reach each computer, and by *tree* we mean we have no redundancy in the network. How do we find this minimum length? Test all possible networks for the minimum total cost? That would be $25^{23} \approx 1.4 \times 10^{32}$ networks to consider. (There are n^{n-2} possible trees on n vertices; Bogart [**2**] gives details.) A computer checking one billion configurations per second would take over a quadrillion years to complete the task. (That's 10^{15} years—a very long time.) Matroids provide a more efficient method.

Not only are matroids useful in these optimization settings, it turns out that they are the very characterizations of the problems. Recognizing that a problem involves a matroid tells us whether certain algorithms will return an optimal solution. Knowing that an algorithm effects a solution tells us whether we have a matroid.

In the undergraduate curriculum, notions of independence arise in various contexts, yet are often not tied together. Matroids surface naturally in these situations. We provide a brief, accessible introduction so that matroids can be included in undergraduate courses, and so that students (or faculty!) interested in matroids have a place to start. For further study of matroids, please see Oxley's *Matroid Theory* [**9**], especially its 61-page chapter, *Brief Definitions and Examples*. Only a cursory knowledge of linear algebra and graph theory is assumed, so take out your pencil and work along.

Declaration of (in)dependence

In everyday life, what do we mean by the terms *dependence* and *independence*? In life, we feel dependent if there is something (or someone) upon which (or whom) we must rely. On the other hand, independence is the state of self-sufficiency, and being reliant upon nothing else. Alternatively, we consider something independent if it somehow extends beyond the rest, making new territory accessible, whether that territory is physical, intellectual, or otherwise. In such a case that independent entity is necessary for access to this new territory.

But we use these terms more technically in mathematics, so let us connect the colloquial to the technical by considering two examples where we find independence.

Linear independence of vectors The first and most familiar context where we encounter independence is linear algebra, when we define the linear independence of a set of vectors within a particular vector space. Consider the following finite collection of vectors from the vector space $\mathbb{R}^3$ (or $\mathbb{C}^3$ or $(\mathbb{F}_3)^3$):

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad v_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

$$v_5 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad v_6 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \quad v_7 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

It is not difficult to determine which subsets of this set are *linearly independent* sets of vectors over $\mathbb{R}^3$: subsets in which it is impossible to represent the zero vector as a nontrivial linear combination of the vectors of the subset. To put it another way, no vector within the subset relies upon any of the others. If some vector *were* a linear combination of the others, we would call the set of vectors *linearly dependent*. Clearly, this means v_7 must be excluded from any subset aspiring to linear independence.

Let us identify the *maximal independent sets*. By *maximal* we mean that the set in question is not properly contained within any other independent set of vectors. We know that since the vector space has dimension 3, the size of such a maximal set can be no larger than 3; in fact, we can produce a set of size 3 immediately, since $\{v_1, v_2, v_3\}$ forms the standard basis. It takes little time to find $\mathcal{B}$, the complete set of maximal independent sets. The reader should verify that $\mathcal{B}$ is

$$\{\{v_1, v_2, v_3\}, \{v_1, v_2, v_4\}, \{v_1, v_2, v_5\}, \{v_1, v_3, v_5\}, \{v_1, v_4, v_5\},$$
$$\{v_2, v_3, v_4\}, \{v_2, v_3, v_6\}, \{v_2, v_4, v_5\}, \{v_2, v_4, v_6\},$$
$$\{v_2, v_5, v_6\}, \{v_3, v_4, v_5\}, \{v_3, v_5, v_6\}, \{v_4, v_5, v_6\}\}.$$

Note that each set contains exactly three elements. This will turn out to be a robust characteristic when we expand the scope of our exploration of independence.

We know from linear algebra that every set of vectors has at least one maximal independent set. Two other properties of $\mathcal{B}$ will prove to be important:

- No maximal independent set can be properly contained in another maximal independent set.
- Given any pair of elements, $B_1, B_2 \in \mathcal{B}$, we may take away any v from B_1 and there is some element $w \in B_2$ such that $(B_1 \setminus v) \cup w$ is in $\mathcal{B}$.

The reader is encouraged to check the second property in a few cases, but also strongly encouraged not to bother checking all $\binom{10}{2} = 45$ pairs of maximal sets. (A modest challenge: Using your linear algebraic expertise, explain why this "exchange" must be possible in general.)

Notice that we used only seven vectors from the infinite set of vectors in $\mathbb{R}^3$. In general, given any vector space, we could select some finite set of vectors and then find the maximal linearly independent subsets of that set of vectors. These maximal sets necessarily have size no larger than the dimension of the vector space, but they may not even achieve that size. (Why not?) Whatever the size of these maximal sets, they will always satisfy the two properties listed above.

Graph theory and independence Though not as universally explored as linear algebra, the theory of graphs is hardly a neglected backwater. (West [**11**] and Wilson [**15**] give a general overview of basic graph theory.) We restrict our attention to connected graphs. There are two common ways to define independence in a graph, on the vertices or on the edges. We focus on the edges. What might it mean for a set of edges to be independent?

Revisiting the idea of independence being tied to necessity, and the accessibility of new territory, when would edges be necessary in a connected graph? Edges exist to connect vertices. Put another way, edges are how we move from vertex to vertex in a graph. So some set of edges should be considered independent if, for each edge, the removal of that edge makes some vertex inaccessible to a previously accessible vertex.

Consider the graph in FIGURE 1 with edge set $E = \{e_1, e_2, \ldots, e_7\}$.

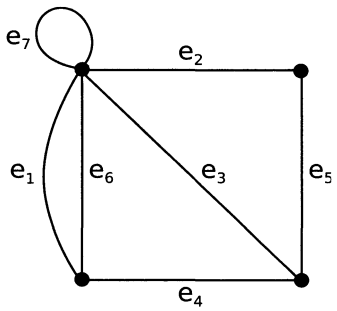


Figure 1 Connected graph G

Now, consider the subset of edges $S = \{e_1, e_3, e_4, e_5\}$. Is this an independent set of edges? No, because the same set of vertices are connected to one another even if, for example, edge e_3 were removed from S. Note that the set S contains a cycle. (A *cycle* is a closed path.) Any time some set of edges contains a cycle, it cannot be an independent set of edges. This also means $\{e_7\}$ is not an independent set, since it is itself a cycle; it doesn't get us anywhere new.

In any connected graph, a set of edges forming a tree or forest (an acyclic subgraph) is independent. This makes sense two different ways: first, a tree or forest never contains a cycle; second, the removal of any edge from a tree or forest disconnects some vertices from one another, decreasing accessibility, and so every edge is necessary. A maximal such set is a set of edges containing no cycles, which also makes all vertices accessible to one another. This is called a *spanning tree*. There must be at least one spanning tree for a connected graph. Here is the set, $\mathcal{T}$, of all spanning trees for G:

$$\mathcal{T} = \{\{e_1, e_2, e_3\}, \{e_1, e_2, e_4\}, \{e_1, e_2, e_5\}, \{e_1, e_3, e_5\}, \{e_1, e_4, e_5\},$$
$$\{e_2, e_3, e_4\}, \{e_2, e_3, e_6\}, \{e_2, e_4, e_5\}, \{e_2, e_4, e_6\},$$
$$\{e_2, e_5, e_6\}, \{e_3, e_4, e_5\}, \{e_3, e_5, e_6\}, \{e_4, e_5, e_6\}\}.$$

Here again we see that all maximal independent sets must have the same size. (How many edges are there in a spanning tree of a connected graph on n vertices?)

Spanning trees also have two other important traits:

- No spanning tree properly contains another spanning tree.
- Given two spanning trees, T_1 and T_2, and an edge e from T_1, we can always find some edge f from T_2 such that $(T_1 \setminus e) \cup f$ will also be a spanning tree.

To demonstrate the second condition, consider the spanning trees T_1 and T_2 shown as bold edges of the graph G in FIGURE 2.

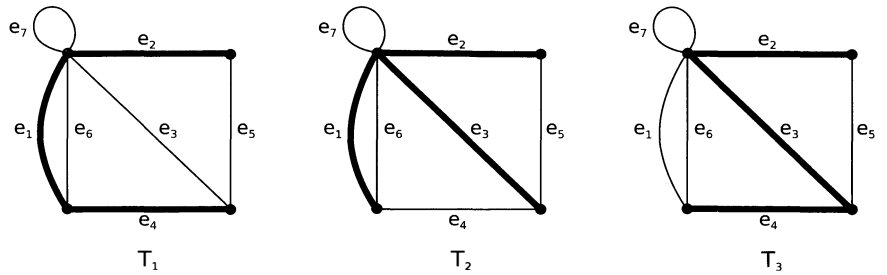


Figure 2 Three spanning trees of G

Suppose we wanted to build a third spanning tree using the edges from T_1 except e_1. Then we must be able to find some edge of T_2 that we can include with the leftover edges from T_1 to form the new spanning tree T_3. We can, indeed, include edge e_3 to produce spanning tree T_3, also shown in FIGURE 2. This exchange property would hold for any edge of T_1.

Motivated by our two examples, now is the proper time for some new terminology and definitions to formally abstract these behaviors.

Thus, matroids

As you notice these similarities between the spanning trees of a graph and the maximal independent sets of a collection of vectors, we should point out that you are not alone. In the 1930s, H. Whitney [**13**], G. Birkhoff [**1**], and S. Maclane [**8**] at Harvard and B. L. van der Waerden [**12**] in Germany were observing these same traits. They noticed these properties of independence that appeared in a graph or a collection of vectors, and wondered if other mathematical objects shared this behavior. To allow for the possibility of other objects sharing this behavior, they defined a matroid on *any* collection of elements that share these traits. We define here a matroid in terms of its maximal independent sets, or *bases*.

The bases A *matroid* M is an ordered pair, $(E, \mathcal{B})$, of a finite set E (the *elements*) and a nonempty collection $\mathcal{B}$ (the *bases*) of subsets of E satisfying the following conditions, usually called the *basis axioms*:

- No basis properly contains another basis.
- If B_1 and B_2 are in $\mathcal{B}$ and $e \in B_1$, then there is an element $f \in B_2$ such that $(B_1 \setminus e) \cup f \in \mathcal{B}$.

The bases of the matroid are its maximal independent sets. By repeatedly applying the second property above, we can show that all bases have the same size.

Returning to our examples, we can define a matroid on a graph. This can be done for any graph, but we will restrict our attention to connected graphs. If G is a graph with edge set E, the *cycle matroid* of G, denoted $M(G)$, is the matroid whose element set, E, is the set of edges of the graph and whose set of bases, $\mathcal{B}$, is the set of spanning trees of G. We can list the bases of the cycle matroid of G by listing all of the spanning trees of the graph.

For the graph in the FIGURE 1, the edges $\{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ are the elements of $M(G)$. We have already listed all of the spanning trees of the graph above, so we already have a list of the bases of this matroid.

We can also define a matroid on a finite set of vectors. The vectors are the elements, or *ground set*, of the matroid, and $\mathcal{B}$ is the set of maximal linearly independent sets of vectors. These maximal independent sets, of course, form bases for the vector space spanned by these vectors. And we recall that all bases of a vector space have the same size.

This helps us see where some of the terminology comes from. The bases of the vector matroid are bases of a vector space. What about the word *matroid*? We can view the vectors of our example as the column vectors of a matrix, which is why Whitney [**13**] called these matroids.

$$\begin{array}{c} \begin{array}{ccccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

These (column) vectors $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ are the elements of this matroid. The bases are the maximal independent sets listed in the previous section.

Now for a quick example not (necessarily) from a matrix or graph. We said that any pair $(E, \mathcal{B})$ that satisfies the two conditions is a matroid. Suppose we take, for example, a set of four elements and let the bases be every subset of two elements. This is a matroid (check the two conditions), called a *uniform matroid*, but is it related to

a graph or a collection of vectors? We will explore this later, but first let us further develop our first two examples.

Beyond the bases You might notice something now that we've looked at our two examples again. The bases of the cycle matroid and the bases of the vector matroid are the same, if we relabel v_i as e_i. Are they the same matroid? Yes. Once we know the elements of the matroid and the bases, the matroid is fully determined, so these matroids are *isomorphic*. An isomorphism is a structure-preserving correspondence. Thus, two matroids are isomorphic if there is a one-to-one correspondence between their elements that preserves the set of bases [**15**].

Knowing the elements and the bases tells us exactly what the matroid is, but can we delve deeper into the structure of this matroid? What else might we like to know about a matroid? Well, what else do we know about a collection of vectors? We know what it means for a set of vectors to be linearly *dependent*, for instance. In a graph, we often look at the *cycles* of the graph. If we had focused on the linearly dependent sets and cycles in our examples, we would have uncovered similar properties they share.

Recall also that, if we take a subset of a linearly independent set of vectors, that subset is linearly independent. (Why? If a vector could not be written as a linear combination of the others, it cannot be written as a linear combination of a smaller set.) Also, if we take a subset of the edges of a tree in a graph, that subset is still independent: If a set of edges contains no cycle, it would be impossible for a subset of those edges to contain a cycle. So any subset of an independent set is independent, and this is true for matroids in general as well.

We can translate some of these familiar traits from linear algebra and graph theory to define some more features of a matroid. Any set of elements of the matroid that is contained in a basis is an *independent set* of the matroid. Further, any independent set can be extended to a basis. On a related note, anytime we have two independent sets of different sizes, say $|I_1| < |I_2|$, then we can always find some element of the larger set to include with the smaller so that it is also independent: There exists some $e \in I_2$ such that $I_1 \cup e$ is independent. This is an important enough fact that if we were to axiomatize matroids according to independence instead of bases—as we mention later—this would be an axiom! It also fits our intuition well, if you think about what it means for vectors.

A subset of E that is not independent is called *dependent*. A minimal dependent set is a *circuit* in the matroid; by minimal we mean that any proper subset of this set is not dependent.

What is an independent set of the cycle matroid? A set of edges is independent in the matroid if it contains no cycle in the graph because a subset of a spanning tree cannot contain a cycle. Thus, a set of edges is dependent in the matroid if it contains a cycle in the graph. A circuit in this matroid is a cycle in the graph.

Get out your pencils! Looking back at the graph in FIGURE 1, we see that $\{e_2, e_4\}$ is an independent set, but not a basis because it is not maximal. The subset $\{e_7\}$ is not independent because it is a cycle; it is a dependent set, and, since it is a minimal dependent set, it is a circuit. (A single-element circuit is called a *loop* in a matroid.) In fact, any set containing $\{e_7\}$ is dependent because it contains a cycle in the graph, or circuit in the matroid. Another dependent set is $\{e_2, e_3, e_4, e_5\}$, but it is not a circuit; $\{e_2, e_3, e_5\}$ is a circuit.

In the vector matroid, a set of elements is independent in the matroid if that collection of vectors is linearly independent; for instance, $\{v_2, v_4\}$ is an independent set. A dependent set in the matroid is a set of linearly dependent vectors, for example $\{v_2, v_3, v_4, v_5\}$. And a circuit is a dependent set, all of whose proper subsets are independent. $\{v_2, v_3, v_5\}$ is a circuit, as is $\{v_7\}$. We noted earlier that any set containing $\{v_7\}$

is a linearly dependent set; we now see that any such set contains a circuit in the vector matroid.

One way to measure the size of a matroid is the cardinality of the ground set, E, but another characteristic of a matroid is the size of the basis, which we call the *rank* of the matroid. If $A \subset E$ is a set of elements of a matroid, the rank of A is the size of a maximal independent set contained in A. In our vector matroid example, let $A = \{v_1, v_2, v_6, v_7\}$. The rank of A is two. The rank of $\{v_7\}$ is zero.

Because it arose naturally from our examples, we defined a matroid in terms of the bases. There are equivalent definitions of a matroid in terms of the independent sets, circuits, and rank; indeed most introductions of matroids will include several such equivalent axiomatizations. Often the first set of exercises is to show the equivalence of these definitions. We spare the reader these theatrics, and refer the interested reader to Oxley [**9**] or Wilson [**14, 15**].

Matroids you may *not* have known

If a matroid can be represented by a collection of vectors in this very natural way, and can also be represented by a graph, why do we need this new notion of matroid? You may ask yourself, given some matroid, M, can we always find a graph such that M is isomorphic to the cycle matroid of that graph? Given some matroid, M, can we always find a matrix over some field such that M is isomorphic to the vector matroid? Happily, the answer to both of these questions is no. (Matroids might be a little boring if they arose only from matrices and graphs.) A graph or matrix does provide a compact way of viewing the matroid, rather than listing all the bases. But this type of representation is just not always possible. When a matroid is isomorphic to the cycle matroid of some graph we say it is *graphic*. A matroid that is isomorphic to the vector matroid of some matrix (over some field) is *representable* (or *matric*). And not every matroid is graphic, nor is every matroid representable.

To demonstrate this, it would be instructive to look at a matroid that is either not graphic or not representable. The smallest nonrepresentable matroid is the Vamos matroid with eight elements [**9**], and it requires a little more space and machinery than we currently have to show that it is not representable. However, it is fairly simple to construct a small example that is not graphic, so let us focus on finding a matroid that is not the cycle matroid of any graph.

Uniform matroids If we take a set E of n elements and let $\mathcal{B}$ be all subsets of E with exactly k elements, we can check that $\mathcal{B}$ forms the set of bases of a matroid on E. This is the *uniform matroid*, $U_{k,n}$, briefly mentioned earlier. In this matroid, any set with k elements is a maximal independent set, any set with fewer than k elements is independent, and any set with more than k elements is dependent. What are the circuits? Precisely the sets of size $k + 1$.

Let's consider an example. Let E be the set $\{a, b, c, d\}$ and let the bases be all sets with two elements. This is the uniform matroid $U_{2,4}$. Is this matroid graphic? To be graphic, $U_{2,4}$ must be isomorphic to the cycle matroid on some graph; so, there would be a graph G, with four edges, such that all of the independent sets of the cycle matroid $M(G)$ are the same as the independent sets of $U_{2,4}$. All of the dependent sets must be the same as well. Since every set with two elements is a basis of $U_{2,4}$, and every set with more than two elements is dependent, we see that each three-element set is a circuit. Is it possible to draw a graph with four edges such that each collection of three edges forms a cycle? Try it. Remember, each collection of two edges is independent, so must *not* contain a cycle.

A careful analysis of cases proves that it is not possible to construct such a graph, so $U_{2,4}$ is not isomorphic to the cycle matroid on any graph, and thus is not graphic. This matroid is, however, representable. A representation over $\mathbb{R}$ is given below. Check for yourself that the vector matroid is isomorphic to $U_{2,4}$ (by listing the bases).

$$\begin{matrix} a & b & c & d \end{matrix}$$
$$\begin{bmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{bmatrix}$$

Notice that this representation is not unique over $\mathbb{R}$ since we could multiply the matrix by any nonzero constant without changing the independent sets. Also notice that this is *not* a representation for $U_{2,4}$ over the field with three elements $\mathbb{F}_3$ (the set $\{0, 1, 2\}$ with addition and multiplication modulo 3). Why? Because, over that field, set $\{c, d\}$ is dependent.

Harvesting a geometric example from a new field We just saw how a collection of vectors can be a representation for a particular matroid over one field but not over another. The ground set of the matroid (the vectors) is the same in each case, but the independent sets are different. Thus, the matroids are not the same. Let's further explore the role the field can play in determining the structure of a vector matroid, with an example over the field of two elements, $\mathbb{F}_2$. As above, the ground set of our matroid is the set of column vectors, and a subset is independent if the vectors form a linearly independent set when considered within the vector space $(\mathbb{F}_2)^3$.

$$\begin{matrix} a & b & c & d & e & f & g \end{matrix}$$
$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Consider the set $\{d, e, f\}$. Accustomed as we are to vectors in $\mathbb{R}^3$, our initial inclination is that this is a linearly independent set of vectors. But recall that $1 + 1 = 0$ over $\mathbb{F}_2$. This means that each vector in $\{d, e, f\}$ is the sum of the other two vectors. This is a linearly dependent set in this vector space, and thus a dependent set in the matroid, and not a basis. In fact, $\{d, e, f\}$ is a minimal dependent set, a circuit, in the matroid, since all of its subsets are independent.

The matroid generated by this matrix has a number of interesting characteristics, which you should take a few moments to explore:

1. Given any two distinct elements, there is a unique third element that completes a 3-element circuit. (That is, any two elements determine a 3-element circuit.)
2. Any two 3-element circuits will intersect in a single element.
3. There is a set of four elements no three of which form a circuit. (This might be a little harder to find, as there are $\binom{7}{4} = 35$ cases to check.)

Geometrically inclined readers might be feeling a tingle of recognition. The traits described above turn out to be precisely the axioms for a finite projective plane, once the language is adjusted accordingly.

A *finite projective plane* is an ordered pair, $(\mathcal{P}, \mathcal{L})$, of a finite set $\mathcal{P}$ *(points)* and a collection $\mathcal{L}$ *(lines)* of subsets of $\mathcal{P}$ satisfying the following [**5**]:

1. Two distinct points of $\mathcal{P}$ are on exactly one line.
2. Any two lines from $\mathcal{L}$ intersect in a unique point.
3. There are four points in $\mathcal{P}$, no three of which are collinear.

Elements of the matroid are the points of the geometry, and 3-element circuits of the matroid are lines of the geometry. Our example has seven points, and this particular projective plane is called the *Fano plane*, denoted F_7. The Fano plane is shown in FIGURE 3, with each point labeled by its associated vector over $\mathbb{F}_2$. Viewed as a matroid, any three points on a line (straight or curved) form a circuit.

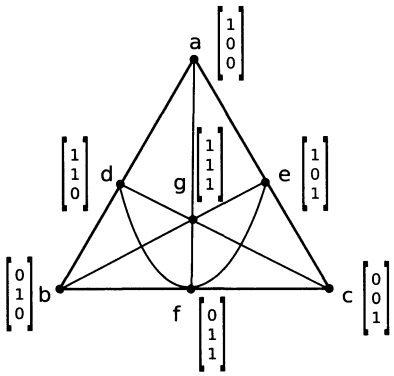


Figure 3 The Fano plane, F_7

We have already seen a variety of structures related to matroids, with still more to come. Ezra Brown wrote in *The many names of* (7, 3, 1) [**3**] in the pages of this MAGAZINE: "In the world of discrete mathematics, we encounter a bewildering variety of topics with no apparent connection between them. But appearances are deceptive." In fact, now that we've recognized the Fano plane as the Fano matroid, we may add this matroid to the list of the "many names of (7, 3, 1)". (For more names of F_7, the interested reader is referred, not surprisingly, to Brown [**3**].)

The Fano plane exemplifies the interesting fact that any projective geometry is also a matroid, though the specific definition of that matroid becomes more complicated once the dimension of the finite geometry grows beyond two. (Although the Fano plane has rank 3 as a matroid it has dimension 2 as a finite geometry, which is, incidentally, why it is called a plane. Oxley [**9**] gives further information.)

We started with a vector matroid and discovered the Fano plane, so we already know that the Fano matroid is representable. The question remains, is it graphic? We attempt to construct a graph, considering the circuits $C_1 = \{a, b, d\}$, $C_2 = \{a, c, e\}$, and $C_3 = \{b, c, f\}$. These would have to correspond to cycles in a graph representation of the Fano matroid. There are two possible configurations for cycles associated with C_1 and C_2, shown in FIGURE 4. In the first we cannot add edge g so that $\{a, f, g\}$ forms a cycle. In the second, we cannot even add f to form a cycle for C_3. (Since the

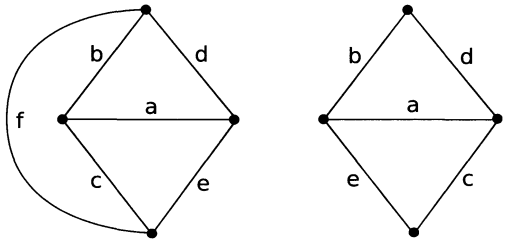


Figure 4 Two possible configurations

matroid has rank 3, the spanning tree must have three edges, so the graph would have 4 vertices and 7 edges.) Thus, the Fano matroid is not a graphic matroid.

One last fact about the Fano plane [**9**]: Viewed as a matroid, the Fano plane is *only* representable over $\mathbb{F}_2$.

Matroids—what are they good for?

Now that we have seen four different types of matroids, we consider their applications. Beyond unifying distinct areas of discrete mathematics, matroids are essential in combinatorial optimization. The greedy algorithm, a powerful optimization technique, can be recognized as a matroid optimization technique. In fact, the greedy algorithm guarantees an optimal solution only if the fundamental structure is a matroid. Once we've familiarized ourselves with the algorithm, we explore how to adapt it to a different style of problem. Finally, we will explore the ramifications, with respect to matroids, of the greedy algorithm's success in finding a solution to this different style of problem.

Walking, ever uphill, and arriving atop Everest Suppose each edge of a graph has been assigned a weight. How would you find a spanning tree of minimum total weight? You could start with an edge of minimal weight, then continue to add the next smallest weight edge available, unless that edge would introduce a cycle. Does this simple and intuitive idea work? Yes, but only because the operative structure is a matroid.

An algorithm that, at each stage, chooses the best option (cheapest, shortest, highest profit) is called *greedy*. The greedy algorithm allows us to construct a minimum-weight spanning tree. (This particular incarnation of the greedy algorithm is called Kruskal's algorithm.) Here are the steps:

In graph G with weight function w on the edges, initialize our set B: $B = \emptyset$.

1. Choose edge e_i of minimal weight. In case of ties, choose any of the tied edges.
2. If $B \cup \{e_i\}$ contains no cycle, then set $B := B \cup \{e_i\}$, else remove e_i from consideration and repeat previous step.

The greedy algorithm concludes, returning a minimum-weight spanning tree B.

We will later see that, perhaps surprisingly, this approach will always construct a minimum-weight spanning tree. The surprise is that a sequence of locally best choices results in a globally optimal solution. In other situations, opting for a locally best choice may, in fact, lead you astray. For example, the person who decides she will always walk in the steepest uphill direction need not end up atop Mount Everest, and, indeed, most of the time such a walk would end instead atop some hill (that is, a local maximum) near her starting point. Or, back to thinking about graphs, suppose a traveling salesperson has to visit several cities and return back home. We can think of the cities as the vertices of a graph, the edges as connecting each pair of cities, and the weight of an edge as the distance he must drive between those cities. What we seek here is a minimum-weight spanning cycle. It turns out that the *greedy algorithm* will not usually lead you to an optimal solution to the *Traveling Salesperson Problem*. Right now, the only way to guarantee an optimal solution is to check all possible routes. For only 10 cities this is $9! = 362{,}880$ possible routes. But for the minimum-weight spanning tree problem, the greedy algorithm guarantees success.

What does this have to do with matroids? The greedy algorithm constructs a minimum-weight spanning tree, and we know what role a spanning tree plays in a graph's associated cycle matroid. Thus, the greedy algorithm finds a minimum-weight basis of the cycle matroid. (Once weights have been assigned to the edges of G, they have also been assigned to the elements of $M(G)$.) Further, for *any* matroid, graphic or otherwise, the greedy algorithm finds a minimum-weight basis.

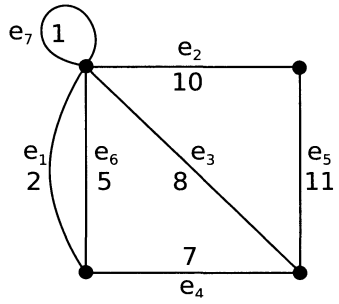


Figure 5 Graph G with weights assigned to each of its edges

Let's work through an example, based on the cycle matroid of the weighted graph shown in FIGURE 5. The greedy algorithm will identify a minimum-weight basis from the set of bases, $\mathcal{B}$. It will build up this basis, element by element; thus, in the algorithm below, the set B will not actually be a basis for the matroid until the algorithm has concluded. (It will be an independent set throughout, but only maximal when the algorithm concludes.) We will use matroid terminology to emphasize the matroidal nature of the algorithm:

Initialize our set B as $B = \emptyset$.

1. The minimum weight element is e_7, but it is rejected since its inclusion would create a circuit. (It is a loop.) $B = \emptyset$.
2. Consider next smallest weight element e_1. It creates no circuits with the edges in B, so set $B = \{e_1\}$.
3. Consider e_6: It creates a circuit with e_1, so do not add it to B. $B = \{e_1\}$.
4. Consider e_4: It creates no circuits with e_1, so set $B = \{e_1, e_4\}$.
5. Consider e_3: It creates a circuit with the current elements of B, so do not add it to B. $B = \{e_1, e_4\}$.
6. Consider e_2: It creates no circuits with the elements of B, so set $B = \{e_1, e_2, e_4\}$.
7. Consider the remaining element, e_5. It creates a circuit with the elements of B. $B = \{e_1, e_2, e_4\}$.

The greedy algorithm concludes, returning a minimum-weight basis $B = \{e_1, e_2, e_4\}$.

None of these steps was actually specific to the graph—they all involve avoiding circuits in the matroid. This is a matroid algorithm for constructing a minimum-weight basis, whether the matroid is graphic or not.

Let us sketch a proof of why this algorithm will always produce a minimum-weight basis. Suppose the greedy algorithm generates some basis $B = \{e_1, e_2, \ldots, e_n\}$, yet

there exists some other basis $B' = \{f_1, f_2, \ldots, f_n\}$ with smaller total weight. Further, without loss of generality, let the elements of each basis be arranged in order of ascending weight. Then $w(e_1) = w(f_1)$, necessarily. Let k be the smallest integer such that $w(f_k) < w(e_k)$. Consider the two independent sets $I_1 = \{e_1, \ldots, e_{k-1}\}$ and $I_2 = \{f_1, \ldots, f_k\}$, and recall the observation we made earlier about two independent sets of different size. Since $|I_1| < |I_2|$ we know there must be some $f_l, l \leq k$ such that $I_1 \cup f_l$ is independent. But this means f_l is both not dependent on $e_1, \ldots, e_{k-1}$ and has weight smaller than e_k. So this is a contradiction, because the greedy algorithm would have selected f_l over e_k in constructing B. This contradiction proves that the greedy algorithm will find a minimum-weight basis. (Oxley [**9**] gives more details.)

What is fascinating and quite stunning is that one may go further and *define* matroids using the greedy algorithm. That is, it turns out that any time the greedy algorithm, in any of its guises, guarantees an optimal solution for all weight functions, we may be sure that the operative mathematical structure *must* be a matroid. Stated another way, only when the structure is a matroid is the greedy algorithm guaranteed to return an optimal solution. (See Oxley [**9**] or Lawler [**7**].) We may, however, have to dig deep to find out what that particular matroid might be.

$$\left(\begin{array}{c}\text{The greedy algorithm}\\ \text{guarantees an optimal solution.}\end{array}\right) \Longleftrightarrow \left(\begin{array}{c}\text{The underlying structure}\\ \text{is actually a matroid.}\end{array}\right)$$

Figure 6 A stunning truth

Finally, one other observation on the nature of matroids is in order. Once a particular matroid is defined, another matroid on the same ground set naturally arises, the *dual matroid*. The set of bases of this new matroid are precisely the set of all complements of bases of the original matroid. That is, given a matroid M on ground set E, with set of bases $\mathcal{B}$, we may always construct the dual matroid with the same ground set and the set of bases $\{B' \subseteq E \mid B' = E \setminus B, B \in \mathcal{B}\}$. What is surprising is that this new collection of sets does in fact satisfy the basis axioms, and this fact has kept many matroid theorists employed for many years. In our current context, the reason this is particularly interesting is that any time the greedy algorithm is used to find a minimum-weight basis for a matroid, it has simultaneously found a maximum-weight basis for the dual matroid. Pause for a moment to grasp, and then savor, that fact. (In fact, the greedy algorithm is sometimes presented first as a method of finding a maximum-weight set of bases, in which case the adjective "greedy" makes a little more sense.)

Is a schedule(d) digression really a digression? Lest we forget how important mathematics can be in the so-called "real world," let us imagine a student with a constrained schedule. This student, call her Imogen, can only take classes at 1 PM, 2 PM, 3 PM, and 4 PM. She's found seven classes that she must take sooner or later, but at the moment she has prioritized them as follows in descending order of importance: Geometry (g), English (e), Chemistry (c), Art (a), Biology (b), Drama (d), French (f). The classes offered at i PM, denoted H_i, are

$$H_1 = \{c, e, f, g\}, \quad H_2 = \{a, b, d\}, \quad H_3 = \{c, e, g\}, \quad H_4 = \{d, f\}.$$

Now the question is perhaps an obvious one: Which classes should Imogen take to best satisfy the prioritization she has set up for herself? Granted, it can be tempting in a small example to stumble our way through some process of trial and error, but let's demonstrate ourselves a trifle more evolved. Casting ourselves in the role of Imogen's advisor, we will attempt something akin to the greedy approach we saw above. Though

it would be a rather busy schedule, we will allow Imogen to take four courses, if there is indeed a way to fill her hours.

Since Geometry is Imogen's top priority, any schedule leaving it out must be considered less than optimal, so we make sure she signs up for g. (This class is offered at two times, but for the moment we must suppress our desire to specify which hour we choose.) What should our next step be? If we can add English, e, without blocking out Geometry, then we should do so. Is it possible to be signed up for both g and e? Yes, each is offered at 1 PM and 3 PM. (We still need not commit her to a time for either class.) Can she take her third priority, Chemistry, c, without dislodging either of those two classes? No, because Chemistry is only offered at 1 PM and 3 PM. There are only two possible times for her top three priorities. What about her fourth priority, Art, a? Yes, she could take Art at 2 PM, the only time it is offered. Imogen's next priority is Biology, b, but it is only offered at 2 PM, where it conflicts with Art. Finally we may fill one more slot in her schedule by signing her up for Drama, d, at 4 PM.

Now her schedule is full, and she is signed up for her first, second, fourth, and sixth most important classes, and filled all her time slots. Better yet, she even still has some flexibility, and can choose whether she'd like to take Geometry at 1 PM and English at 3 PM or vice versa. As her advisor, we leave our office feeling satisfied with our performance, as we should, for if we were to search all her possible schedules, we would find that this is the best we could do.

Why do we need powerful concepts like matroids and the greedy algorithm to tackle this problem? In this example, the problem and constraints are simple enough that trial-and-error may have allowed us to find the solution. But in more complicated situations the number of possibilities grows massive. (This is affectionately referred to as the "combinatorial explosion.") If Imogen had eight possible times to take a class and a prioritized list of 17 classes, trial-and-error would likely be a fool's errand. Similarly, in our earlier example with 25 computers in a network, constructing a minimum-weight spanning tree without the algorithm would be miserable: we would need a quadrillion years to compare all possible spanning trees. Imagine an actual company with hundreds of computers! Knowing that our structure is a matroid tells us that the algorithm will work, and the algorithm is an efficient way to tackle a problem where an exhaustive search might take the fastest computer longer than a human lifetime to compute.

The hidden matroid For Imogen's schedule, at each stage we chose the best option available that would not conflict with previous choices we had made. This is another incarnation of the greedy algorithm, and in this type of scheduling problem it will always produce an optimal solution. (We omit the proof here for brevity's sake. See Bogart [**2**] or Lawler [**7**] for details.) Since the greedy algorithm is inextricably connected to matroids, it must also be true that a matroid lurks in this scheduling example. Let's ferret out that matroid!

To identify the matroid, we need to identify the two sets in $(E, \mathcal{B})$. The first is fairly simple: E is the set of seven courses. Now which subsets of E are bases?

The solution to the scheduling problem is actually an example of a *system of distinct representatives* (or SDR) [**7**]. We have four possible class periods available, and a certain number of courses offered during each period. A desirable feature of a course schedule for Imogen would be that she actually *takes* a class during each hour when she is available. We have a set of seven courses, $\mathcal{C} = \{a, b, c, d, e, f, g\}$, and a family of four subsets of $\mathcal{C}$ representing the time slots available, which we've denoted H_1, H_2, H_3, and H_4. We seek a set of four courses from $\mathcal{C}$ so that each course (element of the set) is taken at some distinct time (H_i). The classes we helped Imogen choose, $\{a, d, e, g\}$, form just such a set; a distinct course can represent each time. Formally, a

set S is a *system of distinct representatives* for a family of sets $A_1, \ldots, A_n$ if there exists a one-to-one correspondence $f : S \to \{A_1, \ldots, A_n\}$ such that for all $s \in S$, $s \in f(s)$. These problems are often modeled using bipartite graphs and matchings. Bogart [**2**] and Oxley [**9**] give details.

The greedy algorithm returns a minimum-weight basis; in our example that basis was an SDR. It turns out that any system of distinct representatives for the family of sets H_1, H_2, H_3, H_4 is a basis for the matroid; that is, $\mathcal{B} = \{S \subseteq \mathcal{C} \mid S \text{ is an SDR for family } H_1, H_2, H_3, H_4\}$. The SDR we found was minimum-weight. (Imogen defined a weight function when she prioritized the classes.)

We now know the matroid, but which sets are independent in this matroid? Gone are the familiar characterizations like "Are the vectors linearly independent?" or "Do the edges form any cycles?" Thinking back to the definition of independence, a set is independent if it is a subset of a basis. In our example, a set $S \subseteq \mathcal{C}$ will be independent when it can be extended into a system of distinct representatives for H_1, H_2, H_3, H_4. This is a more unwieldy definition for independence. But there is a simpler way to characterize it. As long as a subset of size k (from $\mathcal{C}$) can represent k members of the family of sets (the H_is), it will be possible for that set to be extended to a full SDR (assuming that a full SDR is indeed possible, as it was in our example). Naturally, this preserves the property that any subset of an independent set is independent; if some set S can represent $|S|$ members of the family of sets, then clearly any $S' \subseteq S$ can also represent $|S'|$ members of the family of sets.

So it turns out that $(E, \mathcal{B})$ forms a matroid. By definition, SDRs must have the same number of elements, and thus no SDR can properly contain another, satisfying the first condition for the bases of a matroid.

A full proof of the basis exchange property for bases would be rather involved, so let's examine one example to see how it works in this matroid. Consider two bases, $B_1 = \{b, c, d, f\}$ and $B_2 = \{a, d, e, g\}$. Again, each of these is an SDR for the family of sets H_1, H_2, H_3, H_4. In a full proof, we would show that for any element x of B_1, there exists some element y of B_2 such that $(B_1 \setminus x) \cup y$ is a basis, which in this case is an SDR. For this example, consider element d of B_1. We must find an element of B_2 to replace d, and if we do so with e we find that, indeed, the resulting set $B_3 = \{b, c, e, f\}$ is an SDR, as shown in TABLE 1.

TABLE 1: Three SDRs and the sets they represent

Time	Set	B_1	B_2	$B_3 = (B_1 \setminus d) \cup e$
1 PM	H_1	French	Engl. or Geom.	Chemistry
2 PM	H_2	Biology	Art	Biology
3 PM	H_3	Chemistry	Engl. or Geom.	English
4 PM	H_4	Drama	Drama	French

Notice that in B_2, there are options for which class will represent H_1 and H_3. In defining the SDR, we need not pick a certain one-to-one correspondence, we just need to know that at least one such correspondence exists. Note also that the sets represented by f and c changed from B_1 to B_3. Finding a replacement for d from B_2 forced the other classes to shuffle around. This is a more subtle matroid than we've yet seen. (You may also have noticed that we could have just replaced d from B_2 when building B_3. But, wouldn't that have been boring?)

What if Imogen had chosen a list of classes and times such that it was only possible for her to take at most two or three classes? Even in situations where no full system

of distinct representatives is possible, there exists a matroid such that the bases are the partial SDRs of maximum size. Any matroid that can be realized in such a way is called a *transversal matroid*, since SDRs are usually called transversals by matroid theorists. Such a matroid need not be graphic (but certainly could be). The relationship between the types of matroids we have discussed is summarized in the Venn diagram in FIGURE 7.

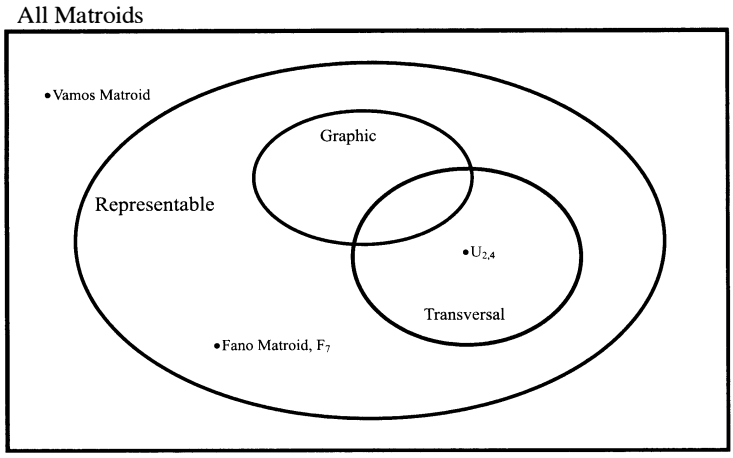


Figure 7 Matroids you have seen

Matroids you have now seen

Where you previously saw independence you might now see matroids. We have encountered five matroids: the cycle matroid, the vector matroid, the uniform matroid, the Fano matroid, and the transversal matroid. Some of these matroids you have known, some are new. With the matroid, we travel to the worlds of linear algebra, graph theory, finite geometry, and combinatorial optimization. The matroid is also tied to endless other discrete structures that we have not yet seen. We have learned that the greedy algorithm is a characterization of a matroid: when we have a matroid, a greedy algorithm will find an optimal solution, but, even more surprisingly, when a greedy approach finds an optimal solution (for all weight functions), we must have a matroid lurking. Once, we have even found that lurking matroid.

Do you now see matroids everywhere you look?

REFERENCES

1. Garrett Birkhoff, Abstract linear dependence and lattices, *Amer. J. Math.* **57** (1935) 800–804.
2. Kenneth P. Bogart, *Introductory Combinatorics*, 3rd ed., Harcourt Academic Press, San Diego, 2000.
3. Ezra Brown, *The Many Names of* (7, 3, 1), this MAGAZINE **75** (2002) 83–94.
4. Tom Brylawski, A partially-anecdotal history of matroids, talk given at "Matroids in Montana" workshop, November, 2006.
5. Ralph P. Grimaldi, *Discrete and Combinatorial Mathematics: An Applied Introduction*, 5th ed., Pearson/Addison-Wesley, Boston, 2004.
6. Frank Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1972.
7. Eugene Lawler, *Combinatorial Optimization: Networks and Matroids*, Dover Publications, Mineola, NY, 2001.

8. Saunders MacLane, Some interpretations of abstract linear dependence in terms of projective geometry, *Amer. J. Math.* **58** (1936) 236–240.
9. James G. Oxley, *Matroid Theory*, Oxford University Press, Oxford, 1992.
10. Gian-Carlo Rota and Fabrizio Palombi, *Indiscrete Thoughts*, Birkhäuser, Boston, 1997.
11. Douglas B. West, *Introduction to Graph Theory*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2000.
12. B. L. van der Waerden, *Moderne Algebra*, 2nd ed., Springer, Berlin, 1937.
13. Hassler Whitney, On the abstract properties of linear dependence, *Amer. J. Math.* **57** (1935) 509–533.
14. Robin J. Wilson, An introduction to matroid theory, *Amer. Math. Monthly* **80** (1973) 500–525.
15. Robin J. Wilson, *Graph Theory*, 4th ed., Addison Wesley Longman, Harlow, Essex, UK, 1996.

Letter to the Editor: Archimedes, Taylor, and Richardson

The enjoyable article "What if Archimedes Had Met Taylor?" (this MAGAZINE, October 2008, pp. 285–290) can be understood in terms of eliminating error terms. This leads to a different concluding approximation that is more in the spirit of the note by combining previous estimates for improvement. We denote the paper's weighted-average estimates for π based on an n-gon by A_n, using area, and P_n, using perimeter. The last section shows two formulas,

$$\text{Error(perim)} = P_n - \pi = \frac{\pi^5}{20n^4} + \frac{\pi^7}{56n^6} + \cdots \quad \text{and}$$

$$\text{Error(area)} = A_n - \pi = \frac{2\pi^5}{15n^4} + \frac{2\pi^7}{63n^6} + \cdots,$$

where we have corrected the first term in the latter. A combination of $8/5$ of the first and $-3/5$ of the second will leave $O(1/n^6)$ error. So, the last table could show

$$\frac{8}{5}P_{96} - \frac{3}{5}A_{96} = 3.14159265363.$$

This approach could be used alternatively to justify

$$A_n = \frac{1}{3}AI_n + \frac{2}{3}AC_n,$$

where AI_n and AC_n are inscribed and circumscribed areas respectively, by subtracting out the $1/n^2$ error terms and leaving the corrected error formula above. For general integration, a similar derivation motivates Simpson's rule as the combination of $1/3$ trapezoidal rule plus $2/3$ midpoint rule. This is more than a coincidence since the inscribed area connects arc endpoints as in trapezoidal rule and circumscribed area uses the arc midpoint.

The technique of combining estimates to eliminate error terms is known as Richardson's Extrapolation in most numerical analysis textbooks. It is usually applied to halving step-size in the same approximation formula. For example, Archimedes could have computed

$$\frac{16}{15}P_{96} - \frac{1}{15}P_{48} = 3.14159265337,$$

if Taylor could have whispered these magical combinations.

—Richard D. Neidinger
Davidson College
Davidson, NC 28035

NOTES

Series that Probably Converge to One

THOMAS J. PFAFF
Ithaca College
Ithaca, NY 14850
tpfaff@ithaca.edu

MAX M. TRAN
Kingsborough Community College
Brooklyn, NY 11235
mtran@kingsborough.edu

What is your favorite way to prove that

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots=1? \tag{1}$$

Ours involves probability. Say you flip a fair coin until it lands heads and count the number of flips that takes. The probability that it takes one try is $1/2$, the probability that it takes two tries is $1/4$, and so on. The probability that it *never* lands heads is zero, so (1) just says that the sum of all probabilities is certainty, or 1.

In the language of probability, we say that the set of all possible outcomes to our flipping experiment, the *sample space*, has been partitioned into disjoint *events*—they *exhaust* the sample space. An intuitively appealing fact is that when disjoint events exhaust the sample space, their probabilities sum up to 1, and that is the heart of (1).

Students are sometimes unsatisfied when all they can say about an infinite series is that it converges, with no information about the sum. Of course, they have their geometric and telescoping series, whose limits are easily computed. And once Taylor series are on the table, students can use them to sum more complicated series, as long as the corresponding functions are known.

As our first example suggests, probability is a rich source for constructing infinite series with known sums. In a variety of examples, we define a random phenomenon and disjoint events E_i for $i = 1, 2, 3, \ldots$ that exhaust the sample space, $\mathcal{S}$, so that

$$\sum_{i=1}^{\infty} P(E_i) = P(\mathcal{S}) = 1.$$

Whenever we can calculate $P(E_i)$, we have another series that "probably" converges to 1. Some of these series are familiar, others are less common. Of course, our students might think we are cheating here, since we are not finding the sum of a given convergent series; still we enjoyed finding lots of series that converge to one.

Coin flipping series

Let's flip a possibly unfair coin until it lands heads. Say that heads occurs with probability p, where $0 < p < 1$, and call E_i the event that the first head appears on the

ith flip of the coin. In order for this to happen, the coin must have landed tails $i-1$ times, each time with probability $q = 1-p$, and then landed heads. This means that $P(E_i) = p(1-p)^{i-1} = (1-q)q^{i-1}$. If we are sure that these events exhaust the sample space, we conclude

$$1 = P(\mathcal{S}) = \sum_{i=1}^{\infty} P(E_i) = \sum_{i=1}^{\infty}(1-p)^{i-1}p = \sum_{i=1}^{\infty}(1-q)q^{i-1},$$

which is the known sum of the geometric series for $0 < q < 1$. Unfortunately, since probabilities must be positive our methods do not work on the other half of the domain of the geometric series where $-1 < q < 0$.

It seems intuitively obvious that our coin must lands heads at some point. For a formal justification that the events E_i exhaust the sample space, let's focus on long runs of tails. Let F_i be the event that all i tosses are tails, so that $\mathcal{S} \setminus \cup_{i=1}^{\infty} E_i = \cap_{i=1}^{\infty} F_i$. Now $P(F_i) = q^i$, so $\lim_{i\to\infty} P(F_i) = 0$, which means $\mathcal{S} = \cup_{i=1}^{\infty} E_i$.

We pause for a moment to note that in all our examples $\mathcal{S} \setminus \cup_{i=1}^{\infty} E_i$ may be nonempty, but it can always be shown to have measure 0. In terms of probability, it is safe to say that $\mathcal{S} = \cup_{i=1}^{\infty} E_i$.

Generalizing this approach, let us fix an integer j and continue to flip our coin (the one that lands heads with probability $0 < p < 1$) until j heads appear. Let E_i^j be the event that the jth head appears on the ith flip and let us assume for the moment that these events exhaust the sample space.

Since $P(E_i^j) = \binom{i-1}{j-1}p^j(1-p)^{i-j}$, a count of total probability gives

$$1 = P(\mathcal{S}) = \sum_{i=j}^{\infty} P(E_i^j) = \sum_{i=j}^{\infty}\binom{i-1}{j-1}p^j(1-p)^{i-j}$$

$$= \sum_{k=0}^{\infty}\binom{k+j-1}{j-1}p^j(1-p)^k,$$

after re-indexing. Readers might recognize this last expression as the total probability of a negative binomial random variable. This series may also be familiar as

$$\frac{1}{(1-q)^{j+1}} = \sum_{k=0}^{\infty}\binom{k+j}{k}q^k,$$

which can be obtained by taking j derivatives with respect to q of the standard geometric series.

To verify that the events exhaust the sample space, let F_i^k be the event of obtaining exactly k heads among i tosses of the coin. We compute

$$\lim_{i\to\infty} P(F_i^k) = \lim_{i\to\infty}\binom{i}{k}p^k q^{i-k} = \lim_{i\to\infty}\frac{i(i-1)\cdots(i-k+1)}{k!}\left(\frac{p}{q}\right)^k q^i$$

$$\leq \lim_{i\to\infty}\frac{i^k}{k!}\left(\frac{p}{q}\right)^k q^i = 0,$$

since an exponential function dominates a power function. Hence, we must eventually obtain j heads and so $\mathcal{S} = \cup_{i=j}^{\infty} E_i^j$.

We now consider flipping a coin until a run of 2 *consecutive* heads appears. Let E_i be the event that the first instance of 2 heads in a row occurs on flips $i-1$ and i, and let F_i be the event that the first i tosses contain no such run. To calculate $P(E_i)$ we

use the general idea from Berresford's paper [**1**]. The first few events are exceptional: $P(E_1) = 0$, $P(E_2) = p^2$, and $P(E_3) = qp^2$. For $i > 3$, $P(E_i) = P(F_{i-3})qp^2$, since a run of two heads appearing for the first time on flip i requires that the last three tosses are THH and no run of two heads occurs earlier.

We now must calculate $P(F_i)$. First, $P(F_1) = 1$ and $P(F_2) = 1 - p^2$. In the event F_i, the ith toss may be either T or H, and in the latter case the previous toss was T, so $P(F_i) = qP(F_{i-1}) + pqP(F_{i-2})$. Solving this recurrence relation for $p = 1/2$, we obtain

$$P(F_i) = \frac{5+3\sqrt{5}}{10}\left(\frac{1+\sqrt{5}}{4}\right)^i + \frac{5-3\sqrt{5}}{10}\left(\frac{1-\sqrt{5}}{4}\right)^i.$$

Note that each term in parentheses is less than 1, so $\lim_{i\to\infty} P(F_i) = 0$. This is what we need to conclude that

$$1 = \frac{1}{4} + \frac{1}{8} + \frac{1}{8}\sum_{i=1}^{\infty} \frac{5+3\sqrt{5}}{10}\left(\frac{1+\sqrt{5}}{4}\right)^i + \frac{5-3\sqrt{5}}{10}\left(\frac{1-\sqrt{5}}{4}\right)^i.$$

Readers may enjoy working this out for general p and q. With a bit of work, one finds that $(1+\sqrt{5})/4$ is replaced by $(q+\sqrt{q^2+4pq})/2$.

Factorials and products from marbles in bags Start with a bag containing one blue and one red marble. We continue to remove marbles from the bag under the rule that if we select a red marble from the bag we put back the red marble along with an additional red marble. The game ends once a blue marble is selected. Some games may take a long time to end, so let's suppose we have a magical bag that can accommodate any number of marbles. This is a variation of Polya's Urn Scheme.

Let E_i be the event that the blue marble is selected on the ith draw from the bag. To begin with, $P(E_1) = 1/2$, and

$$\begin{aligned} P(E_i) &= (1 - P(E_1))(1 - P(E_2)) \cdots (1 - P(E_{i-1}))(1/(i+1)) \\ &= \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4}\cdots\frac{i-1}{i}\cdot\frac{1}{i+1} = \frac{1}{i(i+1)}. \end{aligned}$$

Again, we need to show that the events E_i exhaust the sample space. Let F_i be the event that i red marbles are selected and note that $P(F_i) = 1/(i+1)$ so that $\lim_{i\to\infty} P(F_i) = 0$. The series here is just the standard first example of a telescoping series, but things quickly become more interesting.

Let us change the rules so that if a red marble is chosen we put back the original red along with *two* more red marbles. Again $P(E_1) = 1/2$, but subsequent steps require *double factorials*:

$$\begin{aligned} P(E_i) &= (1 - P(E_1))(1 - P(E_2)) \cdots (1 - P(E_{i-1}))(1/(2i)) \\ &= \frac{(2i-3)!!}{(2i)!!}, \end{aligned}$$

where $n!! = n(n-2)(n-4)\cdots 2$ (or 1) is the double factorial. (By convention, $(-1)!! = 1$.) With F_i defined as above we can show that

$$\lim_{i\to\infty} P(F_i) = \lim_{i\to\infty} \frac{(2i-1)!!}{(2i)!!} = \lim_{i\to\infty} \frac{(2i)!}{(2^i i!)^2} = 0,$$

using Stirling's formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$ to prove that $(2i)!/(2^i i!)^2 \sim \sqrt{4\pi i}/2\pi i$. Hence, we conclude that

$$1 = P(\mathcal{S}) = \sum_{i=1}^{\infty} P(E_i) = \sum_{i=1}^{\infty} \frac{(2i-3)!!}{(2i)!!}. \tag{2}$$

To generalize further, start with a bag that has $b > 0$ blue marbles and $r > 0$ red marbles. Remove marbles from the bag until a blue marble is chosen, under the rule that when a red marble is removed we add l blue marbles and $d + 1$ red marbles. Events E_i and F_i have the same definitions. In this scenario we see that $P(E_1) = b/(b + r)$, and

$$P(E_i) = (1 - P(E_1))(1 - P(E_2)) \cdots (1 - P(E_{i-1}))\frac{b + (i-1)l}{b + (i-1)l + r + (i-1)d}.$$

To show that the whole sample space is covered, notice that

$$P(F_i) = \prod_{j=0}^{i-1} \frac{r + jd}{b + r + j(l + d)}, \text{ and so } \lim_{i\to\infty} P(F_i) = \prod_{j=0}^{\infty} \frac{r + jd}{b + r + j(l + d)}.$$

If the the log of the product approaches $-\infty$, the product approaches 0, so consider

$$\sum_{j=1}^{\infty} \ln\left(\frac{r + jd}{b + r + j(l + d)}\right).$$

The integral test gives the desired result:

$$\int_1^{\infty} \ln\left(\frac{r + xd}{b + r + x(l + d)}\right) dx = -\infty.$$

Intuitively, for j sufficiently large, $(r + jd)/[b + r + j(l + d)] \approx d/(l + d)$, which is less than one when $l \neq 0$; thus the limit of the product is zero.

Calculating the first few probabilities

$$P(E_1) = \frac{b}{b + r}, \qquad P(E_2) = \frac{r(b + l)}{(b + r)(b + r + l + d)}, \qquad \text{and}$$

$$P(E_3) = \frac{r(r + d)(b + 2l)}{(b + r)(b + r + l + d)(b + r + 2l + 2d)},$$

leads to the general result

$$P(E_i) = \frac{b + (i-1)l}{r + (i-1)d} \prod_{j=0}^{i-1} \frac{r + jd}{b + r + j(l + d)}. \tag{3}$$

The two cases above, the original telescoping series and (2), arose from (b, r, l, d) equal to $(1, 1, 0, 1)$ and $(1, 1, 0, 2)$. In the case $(1, 1, 0, 0)$ and, in fact, whenever $b = r$ and $l = d$, the complicated formula (3) reduces to $1/2^i$ and the series is just (1), since these cases are the same as flipping a fair coin. Some other cases that work out nicely are $(1, 1, 1, 0)$, $(1, 1, 2, 0)$, $(k, 1, 0, 1)$ $(k > 1)$, and (b, r, b, r):

$$1 = \sum_{i=1}^{\infty} \frac{i}{(i + 1)!}, \qquad 1 = \sum_{i=1}^{\infty} \frac{2i - 1}{2^i (i!)},$$

$$1 = \sum_{i=1}^{\infty} \frac{k(i-1)!\,k!}{(k + i)!}, \qquad \text{and} \qquad 1 = \sum_{i=1}^{\infty} \frac{b}{r}\left(\frac{r}{b + r}\right)^i.$$

As an exercise, readers can show that the first three are telescoping series and the last is a special case of the geometric series.

For one more layer of complexity, let's start with a bag containing $b > 0$ blue marbles and $r > 0$ red marbles. In this case, if a red marble is removed at the ith turn we add $i \times l$ blue marbles and $i \times d + 1$ red marbles. Again we have $P(E_1) = b/(b + r)$. Conditioning on the previous draw and using $\sum_{k=1}^{n} k = n(n + 1)/2$ we discover that

$$P(E_i) = \frac{2b + i(i-1)l}{2r + i(i-1)d} \prod_{j=1}^{i} \frac{2r + j(j-1)d}{2b + 2r + j(j-1)(l+d)}.$$

As usual we let F_i be the event that i consecutive red marbles are selected. The event F_i is calculated in essentially the same manner as E_i except that the last draw is now a red instead of a blue. Hence,

$$P(F_i) = \prod_{j=1}^{i} \frac{2r + j(j-1)d}{2b + 2r + j(j-1)(l+d)}.$$

In this case, $\lim_{i\to\infty} P(F_i)$ does not always equal 0. For instance, when $l = 0$ and $d = r = b$ we discover, according to Hansen [3] and *Mathematica*, that

$$\lim_{i\to\infty} P(F_i) = \prod_{j=1}^{\infty} \frac{2 + j(j-1)}{4 + j(j-1)} = \cosh\left(\frac{\sqrt{7}\pi}{2}\right) \operatorname{sech}\left(\frac{\sqrt{15}\pi}{2}\right) \approx 0.1455.$$

Intuitively this makes sense since we are continually adding a lot of red marbles and no blue marbles thus making it possible never to draw a blue marble. But now $\mathcal{S} = F \cup \left(\cup_{i=1}^{\infty} E_i\right)$, where F is the event that a blue marble is never drawn. Hence,

$$1 = P(\mathcal{S}) = P(F) + \sum_{i=1}^{\infty} P(E_i) = \cosh\left(\frac{\sqrt{7}\pi}{2}\right) \operatorname{sech}\left(\frac{\sqrt{15}\pi}{2}\right) + \sum_{i=1}^{\infty}\left[\frac{2}{2 + i(i-1)} \prod_{j=1}^{i} \frac{2 + j(j-1)}{4 + j(j-1)}\right].$$

On the other hand, if we let $d = 0$ and $r = b = l$, adding more blue marbles than red, then

$$\lim_{i\to\infty} P(F_i) = \lim_{i\to\infty} \prod_{j=1}^{i} \frac{2}{4 + j(j-1)} \le \lim_{i\to\infty} \prod_{j=1}^{i} \frac{2}{4} \le \lim_{i\to\infty} \left(\frac{2}{4}\right)^i = 0,$$

and so

$$1 = P(\mathcal{S}) = \sum_{i=1}^{\infty} P(E_i) = \sum_{i=1}^{\infty} \frac{2^{i-1}[2 + i(i-1)]}{\prod_{j=1}^{i}[4 + j(j-1)]}. \tag{4}$$

For a slight variation we can start with $b > 0$ blue marbles in the bag and $r > 0$ red marbles and continue to select marbles until a blue is chosen, under the rule that if a red is chosen at the ith turn we add $l \cdot i!$ blue marbles and $d \cdot i! + 1$ red marbles. In this case, calculating in the same manner, we find that $P(E_1) = b/(b + r)$, $P(E_2) = r(b + l)/(b + r)(b + r + l + d)$, and for $i > 2$

$$P(E_i) = \frac{r\left[b + l\sum_{k=1}^{i-1} k!\right]}{(b + r)\left[r + d\sum_{k=1}^{i-1} k!\right]} \prod_{j=1}^{i-1} \frac{r + d\sum_{k=1}^{j} k!}{b + r + (l + d)\sum_{k=1}^{j} k!},$$

and

$$P(F_i) = \frac{r}{b+r} \prod_{j=1}^{i-1} \frac{r + d\sum_{k=1}^{j} k!}{b + r + (l+d)\sum_{k=1}^{j} k!}.$$

For $d = 0$ and $r = b = l$ we have

$$P(F_i) = \frac{1}{2} \prod_{j=1}^{i-1} \frac{1}{1 + \sum_{k=0}^{j} k!},$$

and clearly $\lim_{i\to\infty} P(F_i) = 0$ and so

$$1 = \frac{1}{2} + \frac{1}{3} + \sum_{i=3}^{\infty} \frac{\sum_{k=0}^{i-1} k!}{2\prod_{j=1}^{i-1}\left[1 + \sum_{k=0}^{j} k!\right]}. \tag{5}$$

On the other hand, for $l = 0$ and $r = d = b$, we find that

$$P(F_i) = \frac{1}{2} \prod_{j=1}^{i-1} \frac{\sum_{k=0}^{j} k!}{1 + \sum_{k=0}^{j} k!}. \tag{6}$$

In this case, we don't necessarily expect the limit to be 0 since we are adding a lot of red marbles but no blue marbles. In fact, *Mathematica* suggests that the limit is around 0.23. We leave it as a challenge to the reader to calculate the limit or at least show that it is positive (or wait until the next section to see why).

In yet another variation, we start with $b > 0$ blue marbles and $r > 0$ red marbles in the bag, but this time if a red is chosen at the ith turn add l^i blue marbles and $d^i + 1$ red marbles. Here $P(E_1) = b/(b+r)$, $P(E_2) = r(b+l)/(r+b)(r+d+b+l)$,

$$P(E_i) = \frac{r\left[b + \sum_{k=1}^{i} l^k\right]}{(r+b)\left[r + \sum_{k=1}^{i} d^k\right]} \prod_{j=1}^{i-1} \frac{r + \sum_{k=1}^{j} d^k}{r + b + \sum_{k=1}^{j} d^k + \sum_{k=1}^{j} l^k},$$

and

$$P(F_i) = \frac{r}{r+b} \prod_{j=1}^{i-1} \frac{r + \sum_{k=1}^{j} d^k}{r + b + \sum_{k=1}^{j} d^k + \sum_{k=1}^{j} l^k}.$$

When $d = 0$ and $b = r = l > 1$, we have

$$\lim_{i\to\infty} P(F_i) = \frac{1}{2\prod_{j=1}^{i-1}[2 + (1 - r^{j+1})/(1-r)]} = 0,$$

$P(E_1) = 1/2$, $P(E_2) = 1/3$, and for $i > 2$

$$P(E_i) = \frac{(r^{i-1} + r - 2)(r-1)^{i-2}}{2\prod_{j=1}^{i-1}[r^j + 2r - 3]},$$

which yields, in the specific case when $r = 2$, the series

$$1 = \frac{1}{2} + \frac{1}{3} + \sum_{i=3}^{\infty} \frac{2^{i-2}}{\prod_{j=1}^{i-1}[2^j + 1]}. \tag{7}$$

Of course, when $r = b$ and $l = d$, we again come up with the sum given in (1) since the bag always has equal amounts of red and blue marbles. We leave it to the reader

to check this if they wish. For the case $l = 0$ and $b = r = d$ we don't expect $P(F_i)$ to converge to 0 so we leave that case alone.

Certainly there are other avenues to pursue since there are many possibilities of what to put back in the bag. We could also stop when we have chosen k blue marbles in a row instead of just one. We could also put more than two colors of marbles in the bag: we could have red, blue, and green marbles and stop if a red is chosen on an even draw or a blue is chosen on an odd draw, while adding marbles after each draw.

Connections and conclusions

It is well known [**2**, Theorem 3.20] that $\sum_{n=1}^{\infty} a_n$ and $\prod_{n=1}^{\infty}(1 + a_n)$ either both converge or both diverge. Taking the log of the product the series $\sum_{n=1}^{\infty} \log(1 + a_n)$ behaves just like the other two. As an application, we can show that for $\{a_n\}$ positive, $\prod_{n=1}^{\infty} a_n/(1 + a_n)$ is positive if and only if $\sum_{n=1}^{\infty} 1/a_n$ converges, since

$$\log\left(\prod_{n=1}^{\infty} \frac{a_n}{1 + a_n}\right) = \sum_{n=1}^{\infty} -\log\left(1 + \frac{1}{a_n}\right),$$

which converges to a nonzero value if and only if $\sum_{n=1}^{\infty} 1/a_n$ converges. This shows that the limit in (6) must be positive, since the choice $a_n = \sum_{k=0}^{n} k!$ guarantees the convergence of $\sum_{n=1}^{\infty} 1/a_n$.

To prove another corollary we need the fact that for $\{a_n\}$ positive,

$$\sum_{i=1}^{n} \frac{a_i}{(1 + a_1)(1 + a_2)\cdots(1 + a_i)} = 1 - \frac{1}{(1 + a_1)(1 + a_2)\cdots(1 + a_n)},$$

which follows by replacing a_i with $1 + a_i - 1$ to make the sum telescope. Hence, for a_n positive, we conclude

$$\sum_{i=1}^{\infty} \frac{a_i}{(1 + a_1)(1 + a_2)\cdots(1 + a_i)} = 1 \quad \text{if} \quad \sum_{i=1}^{\infty} a_i \quad \text{diverges,} \tag{8}$$

and

$$\sum_{i=1}^{\infty} \frac{a_i}{(1 + a_1)(1 + a_2)\cdots(1 + a_i)} = 1 - \frac{1}{\prod_{i=1}^{\infty}(1 + a_i)} \quad \text{if} \quad \sum_{i=1}^{\infty} a_i \quad \text{converges,} \tag{9}$$

both of which follow from the relationship between $\sum_{i=1}^{\infty} a_i$ and $\prod_{n=1}^{\infty}(1 + a_n)$.

Many of our results can be derived from these facts. For example, take $a_i = 1 + \binom{i}{2}$ to derive (4), $a_i = \sum_{k=1}^{i}(k - 1)!$ to obtain (5) and $a_i = 2^i$ to get (7). Since our methods provide a probabilistic interpretation of these special cases of (8) and (9), it is natural to wonder whether they can be used to prove (8) and (9) in general. Indeed, they can as the following shows:

Given a sequence of positive numbers a_n, define a game where on the ith turn we throw a dart at a unit square board. The game ends if the dart lands in the region of the square below height $a_i/(1 + a_i)$ and continues if it lands above that height. Let E_i be the event that the game ends at the ith turn and F_i be the event that the game continues to the next turn. Calculation shows that $P(F_1) = 1/(1 + a_1)$ and so

$$P(F_i) = P(F_{i-1})\frac{1}{1 + a_i} = \prod_{n=1}^{i} \frac{1}{1 + a_n}$$

and

$$P(E_i) = P(F_{i-1})P(\text{hitting target area on the } i\text{th throw}) = a_i \prod_{n=1}^{i} \frac{1}{1+a_n}.$$

Finally, with $F = \cap_{i=1}^{\infty} F_i$ and $\mathcal{S} = \cup_{i=1}^{\infty} E_i \cup F$ we see that (8) and (9) are equivalent to the fact that $P(\mathcal{S}) = 1$, and that $\sum_{i=1}^{\infty} a_i$ converges if and only if $\prod_{i=1}^{\infty}(1+a_i)$ converges.

Of course, there are series that we have not yet derived by these methods, such as $\sum_{n=1}^{\infty} 6/(\pi n)^2 = 1$ and $\sum_{n=0}^{\infty} 1/(n!\,e) = 1$. More generally, if we have a series of positive terms is there always a corresponding probabilistic scenario?

Acknowledgment. We would like to thank Warren Johnson who carefully read this paper and provided numerous thoughts that greatly improved it.

REFERENCES

1. G. Berresford, Runs in coin tossing: Randomness revealed, *College Math. J.* **33** (2002) 391–394.
2. Daniel D. Bonar and Michael J. Khoury, *Real Infinite Series*, MAA, 2006.
3. E. R. Hansen, *A Table of Series and Products*, Prentice Hall, Englewood Cliffs, NJ, 1975.

Flip the Script: From Probability to Integration

DAVID A. ROLLS
The University of Melbourne
Parkville, VIC, 3010, AUSTRALIA
D.Rolls@ms.unimelb.edu.au

Sometimes in mathematics we can reinterpret a problem or a result to advantage. The problem becomes easier to solve, or the result becomes even more useful. For an easy example from probability, imagine calculating the integral

$$\int_0^{\infty} e^{-3x}\,dx. \tag{1}$$

The integrand resembles $f(x) = 3e^{-3x}$, $x \geq 0$, the probability density function (p.d.f.) of a rate 3 exponential probability distribution. Although they differ by a multiplying constant, that's easy to address, and we can write

$$\int_0^{\infty} e^{-3x}\,dx = \frac{1}{3}\int_0^{\infty} 3e^{-3x}\,dx = \frac{1}{3}(1) = \frac{1}{3}.$$

The key point is that the integral of a p.d.f. over its domain must be one. Compare that with the standard calculus approach using a substitution $u = -3x$ and a limit for the improper integral

$$\int_0^{\infty} e^{-3x}\,dx = \lim_{t\to\infty}\int_0^{t} e^{-3x}\,dx = \lim_{t\to\infty}\frac{1}{3}\int_{-3t}^{0} e^{u}\,du$$

$$= \frac{1}{3}\lim_{t\to\infty} e^{u}\Big|_{-3t}^{0} = \frac{1}{3} - 0 = \frac{1}{3}.$$

Which solution would you rather use? Now, imagine calculating

$$\int_0^\infty \int_1^5 2xy\, e^{-2x}\, dy\, dx \tag{2}$$

in your head. Think about it for a moment, then read on!

Obviously, students take calculus-based probability after a number of calculus courses where they exercise their skills of integration, including integration by parts and improper integrals. But after the course in probability, what might be said about calculus that couldn't be said before? This note illustrates a few such points, primarily by interpreting integrands as familiar probability density functions. Monte Carlo integration is also discussed as a technique to approximate integrals. To do this, integrands must again be interpreted in terms of probability density functions.

Definitions from probability

Suppose X is a continuous random variable, meaning that X assigns a real number to every possible outcome of some particular experiment [**10**, p. 187]. Then there is a nonnegative function f called the *probability density function* (p.d.f.), or simply the *probability density*, that we use to compute probabilities. A key property of probability densities is

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

The *expected value* (also called the *mean*) of X and $g(X)$ [**10**, pp. 191–192] are defined to be

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{and} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \tag{3}$$

respectively, where g is a real-valued function.

Let us review two common examples of continuous distributions. The rate λ exponential distribution has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

and mean $1/\lambda$. The uniform distribution on the interval $[a, b]$ has density

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

and mean given by the midpoint $(a + b)/2$. (Note that inclusion of the endpoints is optional for continuous distributions since they don't change the value of an integral.) Sometimes probability densities are specified using only the nonzero portion, with the function understood to be zero elsewhere.

The univariate definitions extend to n dimensions, imagining jointly continuous random variables $X_1, \ldots, X_n$ [**10**, Section 6.1]. Then for the *joint p.d.f.* $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n = 1$$

and for the expected value of $g(X_1, \ldots, X_n)$ where g is real-valued, we have

$$E[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Expected value behaves in a linear fashion: For any constants a and b

$$E[aX + b] = aE[X] + b.$$

The condition that $X_1, \ldots, X_n$ are *independent* [**10**, Section 6.2] is equivalent to saying for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$ the joint density factors as

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

where $f_{X_1}(x_1), \ldots, f_{X_1}(x_n)$ are all univariate densities. In this case

$$E[g_1(X_1) \cdots g_n(X_n)] = E[g_1(X_1)] \cdots E[g_n(X_n)]$$

where $g_1, g_2, \ldots, g_n$ are real-valued functions.

Making quick work of integrals

So, how might we reinterpret to advantage? Imagine calculating the integral

$$\int_0^{\infty} xe^{-3x}\, dx$$

which resembles the integral in (1) but adds integration by parts into the mix. Again, we can think of a rate 3 exponential distribution, which has mean $1/3$. By (3) we have

$$\int_0^{\infty} xe^{-3x}\, dx = \frac{1}{3}\int_0^{\infty} x(3e^{-3x})\, dx = \frac{1}{3}E[X] = \frac{1}{9}.$$

It takes longer to write down the integral than to compute it!

For a simple bivariate example, consider the slightly tedious but straightforward integral

$$I_1 = \int_0^1 \int_0^1 (x + y)\, dx\, dy.$$

To the student of probability, if X and Y are random variables with uniform distributions on [0, 1] we have

$$I_1 = E[X + Y] = E[X] + E[Y] = \frac{1}{2} + \frac{1}{2} = 1$$

since expected value behaves in a linear fashion and the mean of a uniform distribution is the midpoint of its interval of definition. The joint p.d.f. here is the constant 1 over the interval of integration.

A slightly more advanced example is the integral

$$I_2 = \int_1^3 \int_1^3 (x + y)\, dx\, dy. \tag{4}$$

Corresponding to each integral is an interval $[1, 3]$. A uniform distribution on this interval has p.d.f.

$$f(x) = \begin{cases} \frac{1}{2}, & 1 \le x \le 3 \\ 0, & \text{otherwise.} \end{cases}$$

We can form a joint p.d.f. $f(x, y)$ here as the product $f(x, y) = f(x)f(y)$, so

$$I_2 = 2^2 \int_1^3 \int_1^3 \frac{(x+y)}{2^2}\, dx\, dy = 4E[X+Y]$$
$$= 4(E[X] + E[Y]) = 4(2+2) = 16.$$

This idea extends so easily. Try calculating

$$\int_1^3 \cdots \int_1^3 (x_1 + \cdots + x_{10})\, dx_1 \cdots dx_{10}. \tag{5}$$

Again, the idea is to write the integral using an integrand that is a legitimate density. How does your calculation time compare with the time just to type the iterated integrals into *Maple*?

The intervals of integration need not be the same, and the integrand doesn't have to be a sum. In fact, the distributions need not be the same either. Consider the integral

$$I_3 = \int_0^\infty \int_1^5 2xy\, e^{-2x}\, dy\, dx$$

mentioned earlier in (2). Thinking of X as a rate 2 exponential distribution, and Y as a uniform distribution on $[1, 5]$, the integrand resembles (up to a constant) an expected value using a joint density which is the product of the two densities

$$f_X(x) = \begin{cases} 2e^{-2x}, & x \ge 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} \frac{1}{4}, & 1 \le y \le 5 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, we can treat X and Y as independent. Since we know $E[X] = 1/2$ and $E[Y] = 3$ we have

$$I_3 = 4 \int_0^\infty \int_1^5 xy\, f_Y(y)\, f_X(x)\, dy\, dx = 4E[XY] = 4E[X]E[Y] = 6.$$

Monte Carlo integration

The idea of using probability to calculate integrals really shows benefits in one of the applications of *Monte Carlo methods* [**9**] called *Monte Carlo integration* [**2**, pp. 670–672], [**3**, pp. 213–221]. Monte Carlo methods are a popularization of statistical sampling techniques that started with an idea of the mathematician Stanislaw Ulam in the early 1940s. He conceived the idea as an attempt to approximate the probability of winning at solitaire, where combinatorics leads to an exponential explosion in the number of possible configurations of playing cards. About this he said, "In a sufficiently complicated problem, actual sampling is better than an examination of all the chains of possibilities" [**11**, p. 197]. The term itself seems to have been coined by Nicholas Metropolis [**8**, p. 127], a colleague and co-author of Ulam, partly in connection to an uncle of Ulam's who would borrow money to visit the casino in Monte Carlo, one of the districts of the Principality of Monaco.

The idea of Monte Carlo integration is to use probabilistic simulation to compute an estimate of the true value of the integral. As above, the key point is to find a density that allows interpreting the integrand in terms of probability. For the integral in (4), imagine simulating $2N$ independent random values $\{X_1, \ldots, X_N\}$ and $\{Y_1, \ldots, Y_N\}$, all from a uniform distribution on $[1, 3]$. Then an estimate of I_2 is given by

$$\hat{I}_N = 4\frac{1}{N}\sum_{i=1}^{N}(X_i + Y_i) \tag{6}$$

and by the Strong Law of Large numbers, with probability 1, as $N \to \infty$

$$\hat{I}_N = 4\frac{1}{N}\sum_{i=1}^{N}(X_i + Y_i) \to 4E[X + Y] = I_2.$$

Unlike with deterministic methods such as Riemann sums, 10 simulations using Monte Carlo integration can yield 10 different estimates, depending on the particular random values that have been generated. For example, five simulations using (6) and either $N = 5$ or $N = 5000$ yielded the estimates in TABLE 1. Larger N reduces the variance of the estimate—that is, a collection of estimates is more closely clustered around the true value. Such simulations can by done with a variety of computer tools and languages such as *R*, *Maple*, and *C*, which provide *pseudorandom* values using some mathematical algorithm.

A measure of the quality of the estimate $\hat{I}_N$ is its variance. The smaller the variance, the tighter the clustering of the estimates, thus increasing the confidence that the error is small. It is also possible several distributions, and so several densities, could be used to generate estimates of the same integral. Smaller variance for similar N is a good reason to prefer one over the other, assuming computation time is comparable. In fact, there are a number of more advanced schemes designed to reduce variance in the estimate.

TABLE 1: Five estimates of (4) using $N = 5$ (top row) and $N = 5000$ (bottom row). Larger N lowers the variance—the estimates are more closely clustered around the true value of sixteen.

$\hat{I}_5$	16.81970	14.46337	17.409612	14.60413	16.40183
$\hat{I}_{5000}$	16.01292	15.98240	16.08013	15.89388	16.02809

Notice that no antiderivative is required for Monte Carlo integration. In fact, if you had one, why use an approximation at all? So, generally you can think of Monte Carlo integration as an alternative to a deterministic numerical integration method (such as the trapezoid rule). In one dimension, those methods are preferable. But an advantage of the Monte Carlo approach is how it scales with dimension. Notice that a Monte Carlo estimate of (4) requires N points of the form (X_i, Y_i), $i = 1, \ldots, N$ so $2N$ random values. Similarly, an estimate for the integral in (5) generalized to d dimensions requires N points $(X_1^{(i)}, \ldots, X_d^{(i)})$, $i = 1, \ldots, N$ so Nd random values. On the other hand, a numerical approximation using a Riemann sum on a grid with M values in each dimension involves M^2 points for (4) and M^d points for the generalization of (5). Since M^d grows exponentially in d, it is potentially a much larger number than Nd, requiring more time to generate the estimate. This is sometimes called the *curse of dimensionality*. A rule of thumb is that Monte Carlo integration is preferable when the dimension is about eight or more. In fact, Monte Carlo integration can provide

estimates of integrals that appear virtually impossible otherwise. Suppose $X_1, \ldots, X_d$ are random variables, possibly not independent, with joint p.d.f. $f_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$. The *marginal expectation* of X_1, is the d-dimensional integral

$$E[X_1] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 f_{X_1,\ldots,X_d}(x_1, \ldots, x_d)\, dx_1 \cdots dx_d.$$

The support of f may be a d-dimensional hypercube, $\mathbb{R}^d$, or some more complicated volume. The dimension d might be large (the number of days in the year or observations in a large experiment). One can always sample from the volume and thereby create an estimate of the expected value.

Final thoughts

The examples shown here use distributions whose features can be reasonably remembered. If you can remember the density and mean of other distributions, you can calculate more integrals similarly. Maybe you can extend to variances and higher moments too. But remember, not all distributions have finite expected mean or variance—so be careful which distributions you use! Also, all the iterated integrals here had constant or infinite limits of integration—no variables. This allows using ideas of independence. Unfortunately, not all integrals are so simple.

What we've seen here is the use of probability to reinterpret some problems from calculus. The references below give more information about probability, in general, and Monte Carlo methods, in particular. But we could equally consider any topic through the lens of a later course. What else can you reinterpret to advantage?

Acknowledgment. I would like to thank Edward Boone and the anonymous referees for their helpful comments that improved this paper.

REFERENCES

1. Jay L. Devore, *Probability and Statistics for Engineering and the Sciences*, 6th ed., Brooks/Cole, Belmont, CA, 2004.
2. Haym Benaroya and Seon Han, *Probability Models in Engineering and Science*, Taylor and Francis, New York, 2005.
3. Michael J. Evans and Jeffrey S. Rosenthal, *Probability and Statistics: The Science of Uncertainty*, W. H. Freeman, New York, 2004.
4. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.
5. Paul Hoffman, *The Man Who Loved Only Numbers*, Hyperion Books, New York, 1998.
6. Malvin H. Kalos and Paula A. Whitlock, *Monte Carlo Methods: Volume 1*, John Wiley, New York, 1986.
7. Frances Y. Kuo and Ian H. Sloan, Lifting the curse of dimensionality, *Notices Amer. Math. Soc.* **52** (2005) 1320–1329.
8. N. Metropolis, The beginning of the Monte Carlo method, *Los Alamos Sci.* **15** Special Issue (1987) 125–130.
9. Nicholas Metropolis and S. Ulam, The Monte Carlo method, *J. Amer. Statist. Assoc.* **44** (1949) 335–341.
10. Sheldon Ross, *A First Course in Probability*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
11. S. M. Ulam, *Adventures of a Mathematician*, Charles Scribner, New York, 1976.

Dirichletino

INTA BERTUCCIONI
Chemin de la Raye, 11
CH-1024 Écublens, SWITZERLAND
Inta.Bertuccioni@gmail.com

In 1837 P. G. Lejeune Dirichlet published his celebrated theorem [**1**], stating that any arithmetic progression $a, a+b, a+2b, a+3b, \ldots$ wherein a and b have no common factor contains infinitely many prime numbers. All known proofs are difficult, and the most readable ones (see, for instance, Serre [**3**]) use nontrivial results of complex analysis. It can be shown [**4, 5**] that, in a very precise technical sense, elementary proofs are only possible when $a^2 \equiv 1$ modulo b, in particular for $a = 1$ and $a = -1$. Recently, Hillel Gauchman published a simple proof [**2**] for the case $a = 1$. We show here what we believe to be an even simpler proof, using an idea of Gauchman's [**2**] (the only idea we need, in fact) and a little lemma. To fix the notation, we restate what we want to prove.

THEOREM. *If N is any positive integer, there are infinitely many primes of the form $1 + mN$, $m = 1, 2, \ldots$.*

Proof. Let $q_1, \ldots, q_r$ be r primes of the form $1 + mN$. We will find another prime of this form. Let $p_1, \ldots, p_s$ be the distinct prime divisors of N. Consider, as in Gauchman [**2**], for $k = 1, \ldots, s$ the polynomials

$$f_k(X) = \frac{X^N - 1}{X^{N/p_k} - 1} = \left(X^{N/p_k}\right)^{p_k - 1} + \left(X^{N/p_k}\right)^{p_k - 2} + \cdots + 1.$$

Fixing an index k, we can decompose $f_k(X)$ into a product of irreducible monic polynomials in $\mathbb{Z}[X]$. Since $f_k(e^{2\pi i/N}) = 0$, one of these irreducible factors, say $f(X)$, must vanish at $e^{2\pi i/N}$. By a well-known lemma of Gauss, an integral polynomial that is irreducible in $\mathbb{Z}[X]$ is also irreducible in $\mathbb{Q}[X]$, and therefore $f(X)$ must be the minimal polynomial of $e^{2\pi i/N}$ over $\mathbb{Q}$. Thus $f(X)$ is the same for every k.

If t is a sufficiently large integer, then for $c = tp_1 \cdots p_s q_1 \cdots q_r$, we will have $f(c) \geq 2$. Let q be a prime divisor of $f(c)$, hence of $c^N - 1$. It must be different from each p_k and each q_k, because none of the primes p_k and none of the primes q_k divides $c^N - 1$. Furthermore, by the lemma below, q does not divide any of the $c^{N/p_k} - 1$. This means that $c^N \equiv 1 \pmod{q}$, whereas $c^{N/p_k} \not\equiv 1 \pmod{q}$. In other words, the multiplicative order of c modulo q is exactly N. On the other hand, by Fermat's little theorem, $c^{q-1} \equiv 1 \pmod{q}$; hence N divides $q - 1$, which is the same as $q = 1 + mN$ for some integer m. We have proved that, given any number of primes $q_1, \ldots, q_r$ of the form $1 + mN$, we can find another one. Thus there are infinitely many such primes. ∎

It remains to state and prove the lemma, that, as noticed by van der Waerden [**7**], goes back at least to Sylvester [**6**].

LEMMA. *Let c be any integer different from 1 and -1, N a positive integer and p a prime divisor of N. Let $q \neq p$ be a prime divisor of*

$$A = \frac{c^N - 1}{c^{N/p} - 1}.$$

Then q does not divide $c^{N/p} - 1$.

Proof. Setting $b = c^{N/p} - 1$ we have

$$A = \frac{(b+1)^p - 1}{b} = b^{p-1} + \binom{p}{1} b^{p-2} + \cdots + \binom{p}{p-1}$$

and therefore, if q were a divisor of b, it would also divide $\binom{p}{p-1} = p$, which contradicts the assumption $q \neq p$. ■

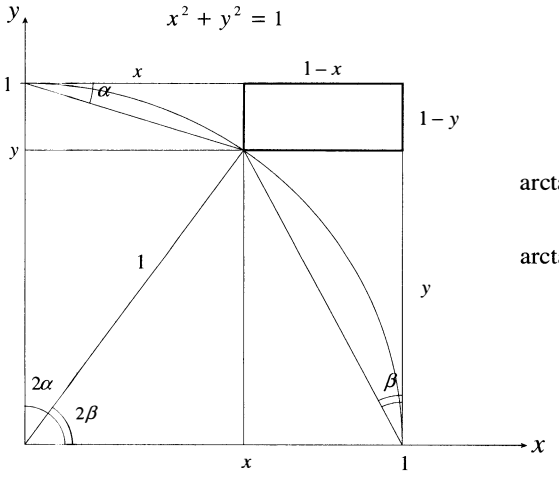
REFERENCES

1. P. G. Lejeune Dirichlet, *Beweis des Satzes, dass jede unbegrenzte arithmetische Progression, deren erstes Glied und Differenz ganze Zahlen ohne gemeinschaftlichen Factor sind, unendlich viele Primzahlen enthält*, *Abh. Preuss. Akad. Wiss.* (1837) 45–81. (Werke I, 313–342.)
2. Hillel Gauchman, A special case of Dirichlet's theorem on primes in an arithmetic progression, this MAGAZINE **74** (2001) 397–399.
3. Jean-Pierre Serre, *Cours d'arithmétique*, 2me édition, Presses Universitaires de France, Paris, 1977.
4. M. Ram Murty, Primes in certain arithmetic progressions, *J. Madras Univ.* **51** (1988) 161–169.
5. M. Ram Murty and Nithum Thain, Prime numbers in certain arithmetic progressions, *Funct. Approx. Comment. Math.* **35** (2006) 249–259.
6. James Joseph Sylvester, On the divisors of the sum of a geometrical series whose first term is unity and common ratio any positive or negative integer, *Nature* **37** (1888) 417–418. (Collected papers IV, 625–629.)
7. B. L. van der Waerden, *Elementarer Beweis eines zahlentheoretischen Existenztheorems*, *J. reine angew. Math.* **171** (1934) 1–3.

Proof Without Words: An Arctangent Identity

If $x, y > 0$ and $x^2 + y^2 = 1$, then

$$\arctan\left(\frac{1-x}{y}\right) + \arctan\left(\frac{1-y}{x}\right) = \frac{\pi}{4}.$$



$$\arctan\left(\frac{1-y}{x}\right) = \alpha$$

$$\arctan\left(\frac{1-x}{y}\right) = \beta$$

$$2\alpha + 2\beta = \frac{\pi}{2} \rightarrow \alpha + \beta = \frac{\pi}{4}$$

—HASAN UNAL
Yildiz Technical University
Istanbul 34210, TURKEY

Evil Twins Alternate with Odious Twins

CHRIS BERNHARDT
Fairfield University
Fairfield, CT 06824
cbernhardt@mail.fairfield.edu

An integer is called *evil* if the number of ones in its binary expansion is *ev*en and *odious* if the number of ones in its binary expansion is *od*d. If we look at the integers between 0 and 15 we find that 0, 3, 5, 6, 9, 10, 12, 15 are evil and that 1, 2, 4, 7, 8, 11, 13, 14 are odious.

Next, we say that if two consecutive integers are evil then this is a pair of *evil twins* and that if two consecutive integers are odious then this is a pair of *odious twins*. Returning to the integers from 0 to 15 we see that {5, 6} and {9, 10} are two sets of evil twins and that {1, 2}, {7, 8} and {13, 14} are three pairs of odious twins. We can now state the new result of this paper:

EVIL TWIN THEOREM. *Evil twins alternate with odious twins.*

The terminology of evil and odious is fairly new coming from combinatorial game theory [**2**], but the theory connected to these numbers has many applications and a long history. One of the first results in the area is due to Prouhet [**8**].

If we look at the numbers from 0 to 15 we see that the number of evil numbers equals the number of odious numbers and that the sum of the evil numbers equals the sum of the odious numbers. More surprisingly, the sum of the squares of the evil numbers equals the sum of the squares of the odious numbers and the sum of the cubes of the evil numbers equals the sum of the cubes of the odious numbers. If we let $0^0 = 1$ we can write the preceding statements as:

$$1^0 + 2^0 + 4^0 + 7^0 + 8^0 + 11^0 + 13^0 + 14^0 = 0^0 + 3^0 + 5^0 + 6^0 + 9^0 + 10^0 + 12^0 + 15^0$$
$$1^1 + 2^1 + 4^1 + 7^1 + 8^1 + 11^1 + 13^1 + 14^1 = 0^1 + 3^1 + 5^1 + 6^1 + 9^1 + 10^1 + 12^1 + 15^1$$
$$1^2 + 2^2 + 4^2 + 7^2 + 8^2 + 11^2 + 13^2 + 14^2 = 0^2 + 3^2 + 5^2 + 6^2 + 9^2 + 10^2 + 12^2 + 15^2$$
$$1^3 + 2^3 + 4^3 + 7^3 + 8^3 + 11^3 + 13^3 + 14^3 = 0^3 + 3^3 + 5^3 + 6^3 + 9^3 + 10^3 + 12^3 + 15^3$$

More succinctly we can write the preceding statements as:

$$\sum_{\substack{k=0 \\ k \text{ is evil}}}^{15} k^j = \sum_{\substack{k=0 \\ k \text{ is odious}}}^{15} k^j \quad \text{for} \quad 0 \le j \le 3.$$

Prouhet proved the remarkable result:

PROUHET'S THEOREM.

$$\sum_{\substack{k=0 \\ k \text{ is evil}}}^{2^n-1} k^j = \sum_{\substack{k=0 \\ k \text{ is odious}}}^{2^n-1} k^j \quad \text{for} \quad 0 \le j \le n-1.$$

In fact Prouhet proved a more general theorem. He looked at integers written in any given base b, and then partitioned $0, 1, \ldots, b^n - 1$ according to the sum of their digits modulo b. He showed that the sums of the integers raised to the power j in each class were equal for any j satisfying $0 \le j \le n-1$. There have been generalizations of this

theorem. Lehmer [**5**] proved a generalization in 1947 and then Sinha [**10**] generalized it further in 1972.

We postpone the proof of Prouhet's Theorem, in favor of introducing the Thue-Morse sequence—a good place to start.

The Thue-Morse sequence

For an integer n, we will define $\alpha(n) = 1$ if n is evil and $\alpha(n) = -1$ if n is odious. Then the *Thue-Morse sequence* is $\alpha(0), \alpha(1), \alpha(2), \ldots$. So the Thue-Morse sequence begins

$$1, -1, -1, 1, -1, 1, 1, -1, -1, 1, 1, -1, 1, -1, -1, 1, \ldots.$$

Often we will slightly simplify the notation by omitting the ones and just writing the signs; so the sequence will be written as

$$+--+-++--++-+--+ \cdots.$$

Notice that the sequence has a nice property that after the first term $+$ comes the negation of that term $-$, after the first two terms $+-$ comes the negation of those terms $-+$, after the first four terms $+--+$ comes the negation $-++-$. In general the terms from 2^n to $2^{n+1} - 1$ are just the negation of the terms from 0 to $2^n - 1$. This follows from the fact that if $0 \leq k \leq 2^n - 1$ then the binary expansion of $2^n + k$ has one more 1 in its binary expansion than does k (namely it has an extra 1 in the 2^n place) and so $\alpha(k)$ and $\alpha(2^n + k)$ have opposite signs. This property makes it very easy to write down the sequence, and will be important later when we come to generating functions.

The Thue-Morse sequence has several other useful properties and has been used in a variety of areas of mathematics. We will briefly discuss some of these. Alloche and Shallit [**1**] give a more complete list of applications and history.

One important area where the sequence occurs is in the application of symbolic dynamics to dynamical systems. Consider the map

$$f(x) = 4x(1 - x).$$

We look at points $x \in [0, 1]$ and see what we can say about the sequence

$$x, f(x), f^2(x), f^3(x), \ldots$$

where f^i denotes f composed (not multiplied) with itself i times. Now define

$$\beta(x) = \begin{cases} -1 & \text{if } 0 \leq x < .5 \\ c & \text{if } x = .5 \\ 1 & \text{if } .5 < x \leq 1. \end{cases}$$

That is, β tells us whether x is to the left or right of .5 or equal to .5. To greatly simplify things instead of considering our sequence

$$x, f(x), f^2(x), f^3(x), \ldots$$

we consider the string of -1s and 1s and possibly a c given by

$$\beta(x), \beta(f(x)), \beta(f^2(x)), \beta(f^3(x)), \ldots.$$

An amazing fact is that for any sequence of -1s and 1s we can find an x with exactly that sequence. The Thue-Morse sequence then shows that there must be a point x that is not periodic and not eventually periodic. (Devaney [**3**, Ch. 1.6] gives a good introduction to symbolic dynamics.) This lack of periodicity was the property that Morse exploited in order to prove a result in differential geometry about geodesics being recurrent but nonperiodic on certain surfaces of negative curvature [**6**].

A useful fact is that the Thue-Morse sequence contains examples of blocks of symbols X such that the block XX also occurs in the sequence. For example, if we take X to be $+$ we find in the sequence the block $++$; if we take X to be $+-$ we can find $+-+-$ as consecutive terms in the sequence. Of course we have to be careful how we choose X. If we take X to be $++$ we cannot find $++++$, because that never occurs. The important point is that there do exist some blocks X such that XX occurs in the sequence. More interestingly, Axel Thue [**11, 12**] (both reprinted in Nagel [**7**]) showed that there is no block of symbols X such that XXX occurs. This is sometimes referred to as saying that the Thue-Morse sequence is cube-free. Max Euwe, the Dutch chess grandmaster and mathematician, used this property to show that there could be infinite games of chess wherein the same sequence of moves never occurred three times in succession [**4**]. In particular, for us, this cube-free property means that there are no evil or odious triplets.

One important way of proving results about this sequence is by looking at its *generating function*, which is a formal power series with the terms of the Thue-Morse sequence as coefficients. This is what we do next.

The Thue-Morse generating function

Notice that $(1-x)(1-x^2) = 1 - x - x^2 + x^3$ and that the coefficients of the polynomial on the right are just the first four terms of the Thue-Morse sequence. If we take $1 - x - x^2 + x^3$ and multiply by $1 - x^4$ we obtain a polynomial whose first four coefficients remain as before, and the coefficients of x^4 to x^7 are just the first four coefficients with the signs reversed. So we end up with a polynomial of degree 7 whose coefficients are the first eight terms in the Thue-Morse sequence. Inductively we can show, and the reader may wish to check, that

$$\prod_{k=0}^{n-1}(1 - x^{2^k}) = \sum_{k=0}^{2^n-1} \alpha(k)x^k.$$

Letting n tend toward infinity, we obtain our generating function

$$\prod_{k=0}^{\infty}(1 - x^{2^k}) = \sum_{k=0}^{\infty} \alpha(k)x^k.$$

Note that we are not interested in the convergence of either the product or the sum. In every application we will truncate at a finite stage, but it is convenient not to indicate any specific stopping point.

We now begin our study of twins. Suppose we multiply both sides of the equation above by $1 + x$. We obtain

$$(1+x)\prod_{k=0}^{\infty}(1 - x^{2^k}) = (1+x)\sum_{k=0}^{\infty} \alpha(k)x^k = \sum_{k=0}^{\infty} \alpha(k)(x^k + x^{k+1})$$

$$= 1 + \sum_{k=1}^{\infty}(\alpha(k) + \alpha(k-1))x^k.$$

Notice that $\alpha(k) + \alpha(k-1)$ will take values of $+2$, 0 or -2 depending on whether $k-1$ and k are evil twins, not twins, or odious twins, respectively. This formal infinite product is like a generating function for evil and odious pairs.

For small values of n it is easy to compute $(1+x)\prod_{k=0}^{n-1}(1-x^{2^k})$ using a computer algebra system. It is suggested that the reader try various examples. For example,

$$(1+x)\prod_{k=0}^{3}(1-x^{2^k}) = 1 - 2x^2 + 2x^6 - 2x^8 + 2x^{10} - 2x^{14} + x^{16}$$

tells us as we observed before that $\{5, 6\}$ and $\{9, 10\}$ are sets of evil twins, and that $\{1, 2\}$, $\{7, 8\}$, and $\{13, 14\}$ are three pairs of odious twins.

In the next two short sections we will give proofs of both the theorems stated in the introduction. Both involve the generating function of the Thue-Morse sequence.

Proof that evil and odious twins alternate

We consider the evil-and-odious-pairs "generating function"

$$(1+x)\prod_{k=0}^{\infty}(1-x^{2^k}) = 1 + \sum_{k=1}^{\infty}(\alpha(k) + \alpha(k-1))x^k.$$

As was stated before, with the exception of the first term, we know that the coefficients are $+2$, 0, or -2 depending on whether $k-1$ and k are evil twins, not twins or odious twins. Now

$$\begin{aligned}(1+x)\prod_{k=0}^{\infty}(1-x^{2^k}) &= (1-x^2)\prod_{k=1}^{\infty}(1-x^{2^k}) = (1-x^2)\prod_{k=0}^{\infty}(1-(x^2)^{2^k}) \\ &= (1-x^2)\sum_{k=0}^{\infty}\alpha(k)(x^2)^k = \sum_{k=0}^{\infty}\alpha(k)((x^2)^k - (x^2)^{k+1}) \\ &= 1 + \sum_{k=1}^{\infty}(\alpha(k) - \alpha(k-1))x^{2k}. \qquad (1)\end{aligned}$$

We see that the coefficient of x^k is zero if k is odd. So that for $\{k-1, k\}$ to be a pair of twins k must be even. (We didn't really need such a complicated argument to deduce this as it is clear that if $k-1$ is even then its last binary digit must be 0 so $\alpha(k-1)$ and $\alpha(k)$ must have opposite signs.) The more interesting conclusion is that $\{2k-1, 2k\}$ is an evil pair if and only if $\alpha(k) = 1$ and $\alpha(k-1) = -1$; and $\{2k-1, 2k\}$ is an odious pair if and only $\alpha(k) = -1$ and $\alpha(k-1) = +1$. This means that as we look along the Thue-Morse sequence whenever we see $-+$ in the $k-1$, k positions we know that $2k-1$ and $2k$ will be evil twins; and whenever we see $+-$ in the $k-1$, k positions we know that $2k-1$ and $2k$ will be odious twins; and there are no other sets of twins. Suppose we look along a string that begins with $+-$ and we know that it contains at least one other $+$ sign. The next $+$ after the first that we see must be preceded by a $-$. This means that any finite string that begins and ends with $+-$ must contain a $-+$. Similarly any string that begins and ends with $-+$ must contain a $+-$. So evil twins must alternate with odious twins.

Prouhet's Theorem

In this final section we give a proof of Prouhet's theorem, which has been proved in many ways. We follow Roberts [**9**] and Wright [**13**] using difference operators. (Wright [**13**] contains interesting history about this theorem and the connection with the Tarry-Escott problem.)

Given any polynomial in one variable, $P(x)$ say, let E denote the operator defined by $EP(x) = P(x+1)$. So E can be thought of as translating the graph of $P(x)$ one unit horizontally. For any positive integer m we let E^m denote E composed with itself m times. Thus $E^m P(x) = P(x+m)$. We also let I denote the identity operator, $IP(x) = P(x)$.

The key observation is that if $P(x)$ is a polynomial of degree d then

$$(I - E^m)P(x) = IP(x) - E^m P(x) = P(x) - P(x+m)$$

is a polynomial of degree $d-1$, and that if $P(x)$ is a constant, $P(x) = c$, then

$$(I - E^m)P(x) = IP(x) - E^m P(x) = c - c = 0.$$

We can now re-write the Thue-Morse generating function using the operator E. We obtain

$$\prod_{k=0}^{n-1}(I - E^{2^k}) = \sum_{k=0}^{2^n-1}\alpha(k)E^k.$$

Suppose that $P(x)$ is a polynomial; then

$$\prod_{k=0}^{n-1}(I - E^{2^k})P(x) = \sum_{k=0}^{2^n-1}\alpha(k)E^k P(x).$$

Suppose that $P(x)$ has degree $d \le n-1$ and examine the left-hand side: It must be 0 because each time we operate on P with a term of the product, we reduce the degree by one, by the key observation above, finally arriving at 0. We obtain

$$0 = \sum_{k=0}^{2^n-1}\alpha(k)E^k P(x) = \sum_{k=0}^{2^n-1}\alpha(k)P(x+k).$$

Letting $P(x) = x^j$, for $0 \le j \le n-1$ gives

$$0 = \sum_{k=0}^{2^n-1}\alpha(k)(x+k)^j.$$

Finally, putting $x = 0$ and re-arranging completes the proof:

$$\sum_{\substack{k=0\\ k \text{ is evil}}}^{2^n-1} k^j = \sum_{\substack{k=0\\ k \text{ is odious}}}^{2^n-1} k^j \quad \text{for} \quad 0 \le j \le n-1.$$

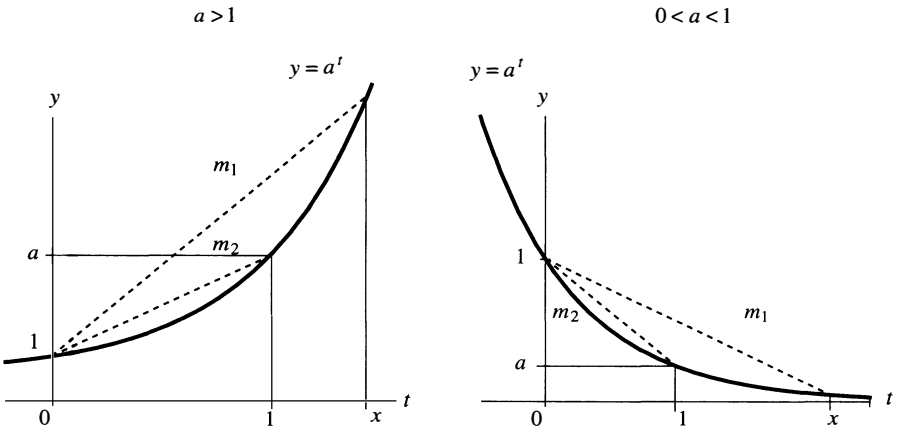
Acknowledgment. We thank the anonymous referees for their many helpful suggestions.

REFERENCES

1. J.-P. Allouche and J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, *Sequences and Their Applications: Proceedings of Seta '98*, Springer-Verlag, New York, 1999, 1–16.
2. E. R. Berlekamp, J. H. Conway, and R. K. Guy, *Winning Ways for Your Mathematical Plays*, **3**, A. K. Peters, New York, 2003, p. 463.
3. R. Devaney, *An Introduction to Chaotic Dynamical Systems*, Benjamin Cummings, Menlo Park, CA, 1986.
4. M. Euwe, Mengentheoretische Betrachtungen über das Schachspiel, Proc. Konin. Akad. Wetenschappen, Amsterdam **32** (1929) 633–642.
5. D. H. Lehmer, The Tarry-Escott problem, *Scripta Math.* **13** (1947) 37–41.
6. M. Morse, Recurrent geodesics on a surface of negative curvature, *Trans. Amer. Math. Soc.* **22** (1921) 84–100.
7. T. Nagel, ed., *Selected Mathematical Papers of Axel Thue*, Universtetsforlaget, Oslo, 1977, pp. 139–158, pp. 413–478.
8. E. Prouhet, Mémoire sur quelques relations entre puissances des nombres, *C. R. Acad. Sci. Paris Sér. I* **33** (1851) 225.
9. J. B. Roberts, A curious sequence of signs, *Amer. Math. Monthly* **64** (1957) 317–322.
10. T. N. Sinha, A note on a theorem of Lehmer, *J. London Math. Soc. (2)* **4** (1971/72) 541–544.
11. A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I Math-Nat. Kl.* **7** (1906) 1–22.
12. A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr. I Math-Nat. Kl. Chris.* **1** (1912) 1–67.
13. E. M. Wright, Prouhet's 1851 solution of the Tarry-Escott problem of 1910, *Amer. Math. Monthly* **66** (1959) 199–201.

Proof Without Words: Bernoulli's Inequality

If $a > 0$, $a \neq 1$, and $x > 1$, then $a^x - 1 > x(a-1)$.



$$m_1 > m_2 \Rightarrow \frac{a^x - 1}{x} > a - 1.$$

—ÁNGEL PLAZA
University of Las Palmas de Gran Canaria
35017-Las Palmas G.C., SPAIN
aplaza@dmat.ulpgc.es

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN GELCA, Texas Tech University; ROBERT GREGORAC, Iowa State University; GERALD HEUER, Concordia College; VANIA MASCIONI, Ball State University; BYRON WALDEN, Santa Clara University; PAUL ZEITZ, The University of San Francisco

PROPOSALS

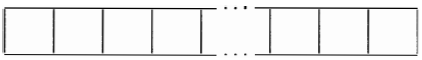
To be considered for publication, solutions should be received by July 1, 2009.

1811. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.*

Given a connected graph G with vertices $v_1, v_2, \ldots, v_n$, let $d_{i,j}$ denote the distance from v_i to v_j. (That is, $d_{i,j}$ is the minimal number of edges that must be traversed in traveling from v_i to v_j.) The Wiener index $W(G)$ of G is defined by

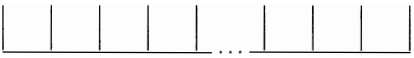
$$W(G) = \sum_{1 \le i < j \le n} d_{i,j}.$$

a. Find the Wiener index for the grid-like graph



on $2n$ vertices.

b. Find the Wiener index for the comb-like graph



on $2n$ vertices.

1812. *Proposed by Bob Tomper, University of North Dakota, Grand Forks, ND.*

Let m and n be relatively prime positive integers. Prove that

$$\sum_{k=1}^{n} k^2 \left\lfloor \frac{km}{n} \right\rfloor = n \sum_{k=1}^{n} k \left\lfloor \frac{km}{n} \right\rfloor - \frac{n(n^2-1)(m-1)}{12}.$$

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames IA 50011, or mailed electronically (ideally as a LaTeX file) to `ehjohnst@iastate.edu`. All communications, written or electronic, should include **on each page** the reader's name, full address, and an e-mail address and/or FAX number.

1813. *Proposed by Elton Bojaxhiu, Albania, and Enkel Hysnelaj, Australia.*

Let a, b, and c be positive real numbers. Prove that

$$\frac{1}{a(1+b)}+\frac{1}{b(1+c)}+\frac{1}{c(1+a)} \geq \frac{3}{\sqrt[3]{abc}(1+\sqrt[3]{abc})}.$$

1814. *Proposed by Michael Goldenberg and Mark Kaplan, The Ingenuity Project, Baltimore Polytechnic Institute, Baltimore, MD.*

Let $A_1A_2A_3$ be a triangle with circumcenter O, and let B_1 be the midpoint of A_2A_3, B_2 be the midpoint of A_3A_1, and B_3 be the midpoint of A_1A_2. For $-\infty < t \leq \infty$ and $k = 1, 2, 3$, let $B_{k,t}$ be the point defined by $\overrightarrow{OB_{k,t}} = t\overrightarrow{OB_k}$ (where by $B_{k,\infty}$ we mean the point at infinity in the direction of $\overrightarrow{OB_k}$). Prove that for any $t \in (-\infty, \infty]$, the lines $A_kB_{k,t}$, $k = 1, 2, 3$, are concurrent, and that the locus of all such points of concurrency is the Euler line of triangle $A_1A_2A_3$.

1815. *Proposed by Stephen J. Herschkorn, Rutgers University, New Brunswick, NJ.*

It is well known that if R is a subring of the ring $\mathbb{Z}$ of integers, then there is a unique positive integer m such that $R = m\mathbb{Z}$. Determine a similar unique characterization for any subring of the ring $\mathbb{Q}$ of rational numbers. What is the cardinality of the class of all subrings of $\mathbb{Q}$? (We do not assume that a ring has a multiplicative identity.)

Quickies

Answers to the Quickies are on page 69.

Q987. *Proposed by Scott Duke Kominers, student, Harvard University, Cambridge, MA.*

Let A and B be $n \times n$ commuting, idempotent matrices such that $A - B$ is invertible. Prove that $A + B$ is the $n \times n$ identity matrix.

Q988. *Proposed by Ovidiu Furdui, Campia-Turzii, Cluj, Romania.*

Let k and p be positive integers. Prove that

$$1^k + 2^k + \cdots + n^k = (1 + 2 + \cdots + n)^p$$

is true for all positive integers n if and only if $k = p = 1$ or $k = 3$ and $p = 2$.

Solutions

Odd sums **February 2008**

1786. *Proposed by Marian Tetiva, Bîrlad, Romania.*

Let $n \geq 2$ be a positive integer and let $O_n = \{1, 3, \ldots, 2n - 1\}$ be the set of odd positive integers less than or equal to $2n - 1$.

a. Prove that if m is a positive integer with $3 \leq m \leq n^2$ and $m \neq n^2 - 2$, then m can be written as a sum of distinct elements from O_n.

b. Prove that $n^2 - 2$ cannot be written as a sum of distinct elements of O_n.

Solution by David Nacin, William Peterson University, Wayne, NJ.
First notice that a set of elements in O_n sums to k if and only if the complement of that set sums to $n^2 - k$. This immediately proves part b, since 2 is not expressible as a sum of distinct elements from O_n.

We prove part a. by induction on n. The base case of $n = 2$ is clear. Now assume that with the exception of 2 and $n^2 - 2$, every integer in between 1 and n^2 can be expressed as a sum of distinct elements of O_n , We show that with the exception of 2 and $(n + 1)^2 - 2$, every m between 1 and $(n + 1)^2$ can be written as a sum of distinct elements of O_{n+1}.

If $1 \leq m \leq n^2$, $m \neq n^2 - 2$ and $m \neq 2$ then m can be written as a sum of distinct elements from O_n, and hence as a sum of distinct elements of O_{n+1}. If $m = n^2 - 2$ and $n > 2$ then

$$m = n^2 - 2 = (n^2 - 2n - 3) + (2n + 1). \tag{1}$$

Because $n^2 - 2n - 3 < n^2 - 2$ and is $\neq 2$, it can be written as a sum of distinct elements of O_n. Since $2n + 1 \in O_{n+1} \backslash O_n$, it follows from (1) that $n^2 - 2$ can be written as a sum of distinct elements of O_{n+1}.

If $m = (n + 1)^2$, then m is the sum of the elements of O_{n+1}. This leaves the numbers of the form $m = n^2 + a$ for $1 \leq a < 2n + 1$. By our earlier remarks, it suffices to show that

$$b = (n + 1)^2 - (n^2 + a) = 2n + 1 - a$$

can be written as a sum of distinct elements of O_{n+1}. However, it is easy to check that $1 \leq b \leq 2n$ and $b \neq 2$, and that any such number satisfies either

$$b \in O_n \qquad \text{or} \qquad b - 1 \in O_n \quad \text{with} \quad b - 1 \geq 3.$$

In either case it is immediate that b is a sum of distinct elements in $O_n \subset O_{n+1}$. This completes the proof.

Also solved by Michael Andreoli, Michel Bataille (France), Brian D. Beasley, J. C. Binz (Switzerland), Jean Bogaert (Belgium), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Cal Poly Pomona Problem Solving Group, Robert Calcaterra, John Christopher, CMC 328 Carleton College, Elliot Cohen, Chip Curtis, Joe DeMaio and Andy Lightcap, Gregory Dresden, Eric Errthum, John Ferdinands, Dmitry Fleischman, David Gould and Chi-Kwong Li and Victoria Perrigan, G.R.A.20 Problem Solving Group (Italy), Natalio H. Guersenzvaig (Argentina), Arup Guha, Russell Jay Hendel, Houghton College Problem Solving Group, Tom Jager, Harris Kwong, Kee-Wai Lau (China), Kathleen E. Lewis, Ronald G. Mosier, Northwestern University Math Problem Solving Group, Rob Pratt, Gary Raduns, Harry Sedinger, Skidmore College Problem Group, Albert Stadler (Switzerland), Phillip D. Straffin, Bob Tomper, S. M. Vaidehi (India), Michael Vowe (Switzerland), Gary L. Walls, Yaming Yu, and the proposer.

A product/sum inequality **February 2008**

1787. *Proposed by Ovidiu Bagdasar, Babes Bolyai University, Cluj Napoca, Romania.*
Let k and n be positive integers with $k \leq n$, and let $0 \leq a_1 \leq a_2 \cdots \leq a_n$. Prove that

$$a_1 a_2 \cdots a_k + a_2 a_3 \cdots a_{k+1} + \cdots + a_{n-k+1} a_{n-k+2} \cdots a_n \leq \left(\frac{a_1 + a_2 + \cdots + a_n}{k} \right)^k.$$

Solution by Chip Curtis, Missouri Southern State University, Joplin, MO.
Write $n = qk + r$, where q and r are nonnegative integers with $r < k$. Then

$$\begin{aligned} &a_1 a_2 \cdots a_k + a_2 a_3 \cdots a_{k+1} + \cdots + a_{n-k+1} a_{n-k+2} \cdots a_n \\ &\quad \leq (a_1 + a_{k+1} + \cdots + a_{qk+1})(a_2 + a_{k+2} + \cdots + a_{qk+2}) \\ &\qquad \cdots (a_r + a_{k+r} + \cdots + a_{qk+r})(a_{r+1} + a_{k+r+1} + \cdots + a_{(q-1)k+r+1}) \\ &\qquad \cdots (a_k + a_{2k} + \cdots + a_{qk}), \end{aligned}$$

where the first r factors each have $q+1$ terms with successive indices differing by k, and the remaining $k-r$ factors each have q terms with successive indices differing by k. By the AM-GM inequality, this product of k factors is less than or equal to $(S/k)^k$, where S is the sum of these factors. Since $S = a_1 + a_2 + \cdots a_n$, the claim is proved.

Also solved by Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Robert Calcaterra, David Gould and Chi-Kwong Li and Victoria Perrigan, Hidefumi Katsuura, Paolo Perfetti (Italy), Nicholas C. Singer, Albert Stadler (Switzerland), Marian Tetiva (Romania), Yaming Yu, and the proposer.

Uniform convergence and continuity **February 2008**

1788. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let D be a nonempty compact set of real numbers, let $\{f_n\}$ be a sequence of real valued functions on D, and let f be a real valued function defined on D. Suppose that $\lim_{n\to\infty} f_n(x_n) = f(x)$ for any sequence $\{x_n\}$ in D with $x_n \to x \in D$.

a. Must it be the case that $f_n \to f$ uniformly on D?

b. Must it be the case that f is continuous on D?

Solution by Tom Jager, Calvin College, Grand Rapids, MI.

We prove that f is continuous and that the convergence is uniform. Observe that if $\{f_{n_k}\}$ is any subsequence of $\{f_n\}$ and $x_k \to x$ in D, then $f_{n_k}(x_k) \to f(x)$. Also note that the constant sequence defined by $x_k = x$ converges to x, so $f_n(x) \to f(x)$ for each x in D.

We first prove that f is continuous on D. Suppose that $x_n \to x$ in D. Because $f_n(x_1) \to f(x_1)$, there is a positive integer n_1 with $|f_{n_1}(x_1) - f(x_1)| < 1/2$. Since $f_n(x_2) \to f(x_2)$, there is an integer $n_2 > n_1$ with $|f_{n_2}(x_2) - f(x_2)| < 1/2^2$. Continuing in this way we obtain an increasing sequence $\{n_k\}$ of positive integers with $|f_{n_k}(x_k) - f(x_k)| < 1/2^k$. As noted above, $f_{n_k}(x_k) \to f(x)$. It follows that $f(x_k) \to f(x)$, so f is continuous at each $x \in D$.

Now assume that the convergence is not uniform. Then there is an $\epsilon > 0$ such that for each N, there is an $n > N$ and an $x \in D$ with $|f_n(x) - f(x)| \geq \epsilon$. Thus there is an increasing sequence $\{n_k\}$ of positive integers and a sequence $\{x_k\}$ of elements of D such that $|f_{n_k}(x_k) - f(x_k)| \geq \epsilon$ for all k. Because D is compact, there is a subsequence $\{x_{k_m}\}$ of $\{x_k\}$ with $x_{k_m} \to x$ for some $x \in D$. Now by the above observation, $f(x_{k_m}) \to f(x)$ and, because f is continuous, $f(x_{k_m}) \to f(x)$. But this implies that $|f_{n_{k_m}}(x_{k_m}) - f(x_{k_m})| \to 0$ and contradicts the fact that $|f_{n_k}(x_k) - f(x_k)| \geq \epsilon$ for all k. We conclude that $f_n \to f$ uniformly on D.

Also solved by Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Paul Budney, Robert Calcaterra, Eugene A. Herman, Shoban Mandel (India), Nicholas C. Singer, Bob Tomper, Stuart V. Witt, and the proposer.

A determinant and an inverse **February 2008**

1789. *Proposed by Harris Kwong, SUNY Fredonia, Fredonia, NY.*

For nonzero real numbers $a_1, a_2, \ldots, a_n$, define $s = \sum_{k=1}^{n} 1/a_k$ and

$$A = \begin{pmatrix} t+a_1 & t & \cdots & t \\ t & t+a_2 & \cdots & t \\ \vdots & \vdots & \ddots & \vdots \\ t & t & \cdots & t+a_n \end{pmatrix},$$

where t is a real number with $st \neq -1$. Find A^{-1} and $\det(A)$.

Solution by Robert Calcaterra, University of Wisconsin Platteville, Platteville, WI.
Let I be the $n \times n$ identity matrix, B be the $n \times n$ matrix with $b_{ij} = 1/a_j$, and D be the $n \times n$ diagonal matrix with $d_{i,i} = a_i$. It is easy to check that $B^2 = sB$ and that $AD^{-1} = I + tB$. It then follows that

$$\begin{aligned} AD^{-1}\left(I - t(1+st)^{-1}B\right) &= (I + tB)\left(I - t(1+st)^{-1}B\right) \\ &= I - t(1+st)^{-1}B + tB - t^2(1+st)^{-1}B^2 = I. \end{aligned}$$

Thus, $A^{-1} = D^{-1}\left(I - t(1+st)^{-1}B\right)$.

To determine the determinant, subtract the last row of A from each of the other rows of A. The result is the matrix

$$T = \begin{pmatrix} a_1 & 0 & \cdots & -a_n \\ 0 & a_2 & \cdots & -a_n \\ \vdots & \vdots & \ddots & \vdots \\ t & t & \cdots & t + a_n \end{pmatrix},$$

which has the same determinant as A. For each i, $1 \le i \le n-1$, add

$$(-1)\frac{t}{a_i} \text{ (row } i \text{ of } T) \text{ to row } n \text{ of } T.$$

The result is an upper triangular matrix S with the same determinant as A and with

$$s_{ii} = \begin{cases} a_i & 1 \le i \le n-1 \\ (st+1)a_n & i = n. \end{cases}$$

It follows that $\det(A) = (st+1)\prod_{i=1}^{n} a_i$.

Also solved by Ricardo Alfaro, Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Hongwei Chen, Chip Curtis, Luz M. DeAlba, John Ferdinands, Fisher Problem Solving Group, Dmitry Fleischman, David Gould and Chi-Kwong Li and Victoria Perrigan, Russell Jay Hendel, Eugene A. Herman, Parviz Khalili, Kim McInturff, Éric Pité (France), Rob Pratt, Ken Ross, Nicholas C. Singer, Albert Stadler (Switzerland), Jeffrey Stuart, Marian Tetiva (Romania), Bob Tomper, Dave Trautman, Yaming Yu, Chris Zin and Samuel Otten, and the proposer.

A ring of power? **February 2008**

1790. *Proposed by Erwin Just (Emeritus), Bronx Community College of the City University of New York, Bronx, NY.*

Let R be a ring and assume that for each $x \in R$,

$$x + x^2 + x^3 + x^4 = x^{11} + x^{12} + x^{13} + x^{28}.$$

Prove that there is an integer $N > 1$ such that for each $x \in R$, we have $x = x^N$.

Solution by the proposer.
We show that $x = x^{127}$ for each $x \in R$. We first prove the following lemma:

LEMMA. *If $u \in R$ with $u^2 = 0$, then $u = 0$.*

Proof. By the hypotheses we have $u + u^2 + u^3 + u^4 = u^{11} + u^{12} + u^{13} + u^{28}$. Because all terms in the sum but the first are 0, we conclude $u = 0$. ∎

Now let $x \in R$ and set

$$V = -x - x^2 - x^3 + x^{10} + x^{11} + x^{12} + x^{27}.$$

Then $Vx = x$ and it follows by induction that $V^n x = x$ for all positive integers n. Next, let

$$W = -V - x - x^2 + x^9 + x^{10} + x^{11} + x^{26}.$$

Then $Wx^2 = x$ and it follows by induction that for positive integers $p < q$,

$$W^p x^q = x^{q-p}. \tag{1}$$

Because $W \in R$ we have, by hypothesis,

$$W + W^2 + W^3 + W^4 = W^{11} + W^{12} + W^{13} + W^{28}.$$

Multiplying both sides of this expression by x^{29} and then applying (1) we obtain

$$x^{28} + x^{27} + x^{26} + x^{25} = x^{18} + x^{17} + x^{16} + x. \tag{2}$$

Adding (2) to the equation in the problem statement leads to

$$(x^2 + x^3 + x^4) + (x^{27} + x^{26} + x^{25}) = (x^{11} + x^{12} + x^{13}) + (x^{18} + x^{17} + x^{16}).$$

Multiply both sides of this expression by W, apply (1), and rearrange to obtain

$$(x + x^2 + x^3) - (x^{10} + x^{11} + x^{12}) = (x^{15} + x^{16} + x^{17}) - (x^{24} + x^{25} + x^{26}). \tag{3}$$

Let $P = x + x^2 + x^3$. Then (3) can be rewritten as $P - x^9 P = x^{14}(P - x^9 P)$, and we can deduce by induction that for any positive integer t

$$P - x^9 P = x^{14t}(P - x^9 P) \qquad \text{or} \qquad P - x^{14t} P = x^9(P - x^{14t} P).$$

By a similar induction we find that for any positive integers s and t,

$$P - x^{14t} P = x^{9s}(P - x^{14t}). \tag{4}$$

Now set $z = P - x^{14t} P$, and then set $t = 9$ and $s = 14$. Then (4) implies that $z = x^{126} z$ and then that $Pz = x^{126} Pz$. Thus,

$$z^2 = (P - x^{126} P)z = Pz - x^{126} Pz = 0.$$

It follows from the Lemma that $z = 0$ and hence that $P = x^{126} P$, that is,

$$x + x^2 + x^3 = x^{127} + x^{128} + x^{129}. \tag{5}$$

Let $Y = x - x^{127}$ and note that (5) can be rewritten as $Y + xY + x^2 Y = 0$ which implies that $xY + x^2 Y + x^3 Y = 0$. Combining these last two equation we find $Y = x^3 Y$, and it follows by induction that $Y = x^{3r} Y$ for any positive integer r. let $r = 42$ and multiply by x to see $xY = x^{127} Y$. Then

$$Y^2 = (x - x^{127})Y = xY - x^{127} Y = 0.$$

By the Lemma, $Y = 0$, so $x = x^{127}$.

Also solved by J. C. Binz (Switzerland), Robert Calcaterra, FAU Problem Solving Group, Tom Jager, Northwestern University Math Problem Solving Group, Nicholas C. Singer.

Answers

Solutions to the Quickies from page 64.

A987. By the hypotheses we have

$$(A + B)(A - B) = A^2 - B^2 + (BA - AB) = A^2 - B^2 = A - B.$$

Because $A - B$ is invertible, $A + B$ is the $n \times n$ identity matrix.

A988. One implication is easy to prove. To prove the other we note that if $k = 1$ then $p = 1$. Thus we consider only the case $k \geq 2$. In this case we have

$$1^k + 2^k + 3^k + \cdots + n^k = \left(\frac{n(n+1)}{2}\right)^p$$

for all positive integers n. Dividing both sides of the preceding equality by n^{k+1} we obtain

$$\frac{1}{n}\left(\left(\frac{1}{n}\right)^k + \left(\frac{2}{n}\right)^k + \cdots + \left(\frac{n}{n}\right)^k\right) = \left(\frac{1}{2}\right)^p \frac{(n(n+1))^p}{n^{k+1}}.$$

Next let $n \to \infty$ to get

$$\int_0^1 x^k\,dx = \frac{1}{k+1} = \left(\frac{1}{2}\right)^p \cdot \lim_{n\to\infty} \frac{(n(n+1))^p}{n^{k+1}}.$$

It follows that $2p = k + 1$ and that $(k + 1)^2 = 2^{k+1}$. It is straight forward to show that the positive solutions of this equation are $k = 1$ and $k = 3$. This completes the proof.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of mathematics literature. Readers are invited to suggest items for review to the editors.

Watkins, John J., *Across the Board: The Mathematics of Chessboard Problems*, Princeton University Press, 2004; x + 257 pp, $42; 2007, $20.95 (P). ISBN 0-691-11503-6; 0-691-13062-0.

I would be happy to recommend this book to you, even if it did not contain the definitive theorem by this MAGAZINE's editor Allen Schwenk telling exactly when a rectangular board has a knight's tour, or a paragraph urging you to read a paper that I wrote half a lifetime ago. Beware, though: The book is not about traditional chess problems but solely about mathematical problems involving the board itself, with generalizations to nonrectangular shapes and higher dimensions (torus, cylinder, Klein bottle). A major focus is knight's tours, including their use in constructing magic squares. Other topics include problems concerned with domination (how few of a chess piece can command every square) and independence (how many pieces can be placed without attacking one another—e.g., the Eight Queens Problem). Also featured are paving boards with dominoes and polyominoes, and (generalizing the checkering of the board) Latin squares and Graeco-Latin squares. Each chapter has problems with solutions. The book is an easy and entertaining read that shows numerous paths into various branches of discrete mathematics and graph theory.

Borwein, J. M., E. M. Rocha, and J. F. Rodrigues, *Communicating Mathematics in the Digital Era*, A K Peters, 2008; xii + 325 pp, $49. ISBN 978-1-56881-410-0.

Do you and your colleagues send papers to arXiv (`http://arxiv.org/`)? Will your institution pay page charges to publish your works? Have you been offered (as I was last year) the "opportunity" to pay a publisher for others to have "open access" to a paper of yours? And who should pay for that? The mechanisms and technologies of dissemination of mathematical and scientific knowledge are in flux. This volume gives some snapshots of ideas, projects, and insights. Notable is the essay by John Ewing of the American Mathematical Society, "The digital downside," in which he cites problems that arise from the "human frailties of carelessness, greed, myopia, dogmatism, and infatuation."

Borwein, Jonathan, and Keith Devlin, *The Computer as Crucible: An Introduction to Experimental Mathematics*, A K Peters, 2009; xi + 158 pp, $29.95 (P). ISBN 978-1-56881-343-1.

This is a book full of astonishments, wondrous and unlikely results that Ramanujan would have been proud to find. One prosaic example: A group of mathematicians "guessed" that

$$4\int_0^1\int_0^1 \frac{dx\,dy}{\sqrt{1+x^2+y^2}} = 8\log\bigl(1+\sqrt{3}\bigr) - 4\log 2 - \frac{2\pi}{3}.$$

The guess was based on experience that related integrals involve combinations of $\log\bigl(1+\sqrt{3}\bigr)$, $\log 2$, and π. The group let a computer algebra system find the combination indicated by comparing numerical evaluation of both sides of the equation to 12 decimal places (and then they confirmed equality to 20 places). The book begins with an apologia for experimental mathematics, followed by experiments in calculating digits of π, "identifying" a number, closed form for

hypergeometric functions and zeta function values, finding the limit of a sum, and much more. Crucial tools are the "3Ms" (Mathematica, Maple, Matlab), Sloane's On-Line Encyclopedia of Integer Sequences (`http://www.research.att.com/~njas/sequences/index.html`), the Inverse Symbolic Calculator (`http://ddrive.cs.dal.ca/~isc/standard.html`—this URL is a correction to the one given on p. 31), and serendipity. Each chapter includes "Explorations" for the reader, with "Answers and Reflections" at the back of the book. One chapter, entitled "The Computer Knows More Math than You Do," shows how explorations with a computer algebra system can lead you to learn about the Lambert *W* function, the HeunG function, and more intricacies and wonders.

Gray, Jeremy, *Plato's Ghost: The Modernist Transformation of Mathematics*, Princeton University Press, 2008; viii + 515 pp, $45. ISBN 978-0-691-13610-3.

You no doubt have some sense of modernism, as in modern art, modern music, and modern literature, as they developed around the turn of the 20th century. Author Gray proposes in detail an interpretation of mathematical developments of that era as having followed the same spirit. His definition of modernism: "an autonomous body of ideas, having little or no outward reference, placing considerable emphasis on formal aspects. . . ." Gray follows the progress through stages of "modernization" by following the threads of geometry, analysis, algebra, and foundations, from "before modernism" to the 1920s and the death of Hilbert's program, by which time modernism had "won." And today? "The most visible failing in the present situation concerns the question of what mathematics actually is. . . . There is not a vision of mathematics as an enterprise to which a large audience can relate."

Coe, Penelope A., *A Handbook: Mathematical Thinking and the Structure of Proofs*, CMT Publishers, 2008; xii + 402 pp, $22 plus shipping from the author, 112 Hazelmere Rd., New Britain, CT 06053-2116, `dr-coe@hotmail.com`. No ISBN.

This book is intended not as a textbook but as a handbook for students in upper-level mathematics courses. It has valuable advice about how to: read mathematics (read in words without citing the letter symbols used), look for overall structure, state a theorem in good rhetoric (with examples of good and bad), and lay out a proof on the page. The book is distinguished from the many others by clever choice of typography and exceptionally well-thought-out layout that emphasize and implement its principles, including insertion of reasons between steps of derivations.

Robson, Eleanor, *Mathematics in Ancient Iraq: A Social History*, Princeton University Press, 2008; xxvii + 441 pp, $49.50. ISBN 978-0691-09182-2.

Lost to contemporary U.S. consciousness about Iraq is any sense of the origins of civilization—and origins of mathematics—there. But the names Iraq, Sumer, Babylonia, and Mesopotamia evoke a culture where 5,000 years ago there arose arithmetic and accounting, not to mention the base-60 system that remains part of our measurement of time and angles. Author Robson deals admirably with an enormous scope (more than 3,000 years, with roughly equal space devoted to each 500-year epoch); numerous sources (950 published clay tablets, all of which are available at a single Website); and the cultural context (social history, an ethnomathematical approach).

Madison, Bernard L. (ed.), *Assessment of Student Learning in College Mathematics: Towards Improved Programs and Courses*, Association for Institutional Research, 2006; ii + 181 pp, $30 ($25 to AIR member or MAA member). No ISBN. Steen, Lynn Arthur (ed.), *Supporting Assessment in Undergraduate Mathematics*, MAA, 2006; vii + 239 pp, free to MAA members or free with copy of Madison book. ISBN 0-88385-820-7. Available online at `http://www.maa.org/saum/cases/welcome.html`.

These two volumes detail case studies of assessment of courses for various audiences in mathematics programs at a variety of institutions. The lesser-known first volume contains "more extensive and detailed" case studies of assessment programs that are "more mature," with some overlap with institutions in the second volume. Additional case studies are at `http://www.maa.org/saum/new_case.html`.

NEWS AND LETTERS

68th Annual William Lowell Putnam Mathematical Competition

Editor's Note: Additional solutions will be printed in the *Monthly* later in the year. A question for our readers: Do you enjoy seeing these solutions in our February issue, or would the space be better used for more Notes and Articles?

PROBLEMS

A1. Let $f: \mathbb{R}^2 \to \mathbb{R}$ be a function such that $f(x, y) + f(y, z) + f(z, x) = 0$ for all real numbers x, y, and z. Prove that there exists a function $g: \mathbb{R} \to \mathbb{R}$ such that $f(x, y) = g(x) - g(y)$ for all real numbers x and y.

A2. Alan and Barbara play a game in which they take turns filling entries of an initially empty 2008×2008 array. Alan plays first. At each turn, a player chooses a real number and places it in a vacant entry. The game ends when all the entries are filled. Alan wins if the determinant of the resulting matrix is nonzero; Barbara wins if it is zero. Which player has a winning strategy?

A3. Start with a finite sequence $a_1, a_2, \ldots, a_n$ of positive integers. If possible, choose two indices $j < k$ such that a_j does not divide a_k, and replace a_j and a_k by $\gcd(a_j, a_k)$ and $\operatorname{lcm}(a_j, a_k)$, respectively. Prove that if this process is repeated, it must eventually stop and the final sequence does not depend on the choices made. (Note: gcd means greatest common divisor and lcm means least common multiple.)

A4. Define $f: \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x & \text{if } x \leq e, \\ xf(\ln x) & \text{if } x > e. \end{cases}$$

Does $\sum_{n=1}^{\infty} 1/f(n)$ converge?

A5. Let $n \geq 3$ be an integer. Let $f(x)$ and $g(x)$ be polynomials with real coefficients such that the points $(f(1), g(1)), (f(2), g(2)), \ldots, (f(n), g(n))$ in $\mathbb{R}^2$ are the vertices of a regular n-gon in counterclockwise order. Prove that at least one of $f(x)$ and $g(x)$ has degree greater than or equal to $n - 1$.

A6. Prove that there exists a constant $c > 0$ such that in every nontrivial finite group G there exists a sequence of length at most $c \log |G|$ with the property that each element of G equals the product of some subsequence. (The elements of G in the sequence are not required to be distinct. A *subsequence* of a sequence is obtained by selecting some of the terms, not necessarily consecutive, without reordering them; for example, 4, 4, 2 is a subsequence of 2, 4, 6, 4, 2, but 2, 2, 4 is not.)

B1. What is the maximum number of rational points that can lie on a circle in $\mathbb{R}^2$ whose center is not a rational point? (A *rational point* is a point both of whose coordinates are rational numbers.)

B2. Let $F_0(x) = \ln x$. For $n \geq 0$ and $x > 0$, let $F_{n+1}(x) = \int_0^x F_n(t)\,dt$. Evaluate

$$\lim_{n\to\infty} \frac{n!\,F_n(1)}{\ln n}.$$

B3. What is the largest possible radius of a circle contained in a 4-dimensional hypercube of side length 1?

B4. Let p be a prime number. Let $h(x)$ be a polynomial with integer coefficients such that $h(0), h(1), \ldots, h(p^2 - 1)$ are distinct modulo p^2. Show that $h(0), h(1), \ldots, h(p^3 - 1)$ are distinct modulo p^3.

B5. Find all continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ such that for every rational number q, the number $f(q)$ is rational and has the same denominator as q. (The *denominator* of a rational number q is the unique positive integer b such that $q = a/b$ for some integer a with $\gcd(a, b) = 1$.)

B6. Let n and k be positive integers. Say that a permutation σ of $\{1, 2, \ldots, n\}$ is *k-limited* if $|\sigma(i) - i| \leq k$ for all i. Prove that the number of k-limited permutations of $\{1, 2, \ldots, n\}$ is odd if and only if $n \equiv 0$ or $1 \pmod{2k+1}$.

SOLUTIONS

Solution to A1. Let $(*)$ denote the given functional relation for f and fix any $a \in \mathbb{R}$. Then from $(*)$, $f(x, y) + f(y, a) + f(a, x) = 0$, so $f(x, y) = -f(a, x) - f(y, a)$. Let $g(x) = f(x, a)$. Then $f(x, y) = -f(a, x) - g(y)$, so it's enough to show $g(x) = -f(a, x)$. From $(*)$ for $y = z = a$, we get $f(x, a) + f(a, a) + f(a, x) = 0$, so $g(x) = -f(a, a) - f(a, x)$ and it's enough to show $f(a, a) = 0$. To do so, take $x = y = z = 0$ in $(*)$ to get $3f(a, a) = 0$, and the result follows.

Solution to A2. Barbara should win. Here's a winning strategy for her: Once Alan places an entry on his first move, pick a different column and place the same number in that column, in the same row as Alan's entry. Call these two columns the *relevant* columns and the other 2006 columns the *irrelevant* columns. From now on, every time Alan plays in an irrelevant column, do so as well. (Because there is an even number of entries in the irrelevant columns, this can be done.) Every time Alan moves in one of the relevant columns, duplicate that move in the same row of the other relevant column. At the end, the relevant columns will be equal, so the determinant will be 0, and Barbara will win.

Solution to A3. At any stage in the process, consider the number N of ordered pairs (a, b) for which a precedes b in the sequence and a does not divide b. When (a_j, a_k) is replaced by $(\gcd(a_j, a_k), \operatorname{lcm}(a_j, a_k))$, a straightforward case analysis shows that the number N always decreases as long as a_j does not divide a_k. [If x precedes both a_j and a_k and divides both, it will also divide both $\gcd(a_j, a_k)$ and $\operatorname{lcm}(a_j, a_k)$. If x divides just one of a_j, a_k, it will certainly divide $\operatorname{lcm}(a_j, a_k)$. Similar considerations hold for x between a_j and a_k and for x beyond a_k. So the number of ordered pairs counted by N goes down by at least 1, since $\gcd(a_j, a_k)$ does divide $\operatorname{lcm}(a_j, a_k)$.] Therefore, the process must stop, and at that stage each term in the sequence will divide its successor.

For any given prime p and any exponent $k \geq 0$, the number of terms x in the sequence for which $p^k | x$ is unaffected by the process. Thus in the final sequence we know exactly which terms (consecutive backwards from the end) are divisible by p^k. Because this is true for all p and k, the terms of the final sequence are completely determined.

Solution to A4. No. The partial sums of the series are increasing, so the question is whether they are bounded. Let $e_1 = e, e_2 = e^e, e_3 = e^{e^e}, \ldots, e_k = e^{e_{k-1}}$ and let $N_1 = \lfloor e_1 \rfloor = 2, N_2 = \lfloor e_2 \rfloor, \ldots, N_k = \lfloor e_k \rfloor$. Then

$$f(x) = \begin{cases} x & x \le e_1, \\ x \ln x & e_1 < x \le e_2, \\ x \ln x \ln(\ln x) & e_2 < x \le e_3, \\ & \vdots \\ x \ln x \ln(\ln x) \cdots \ln^{(k)}(x) & e_k < x \le e_{k+1}, \end{cases}$$

where $\ln^{(k)}$ denotes the composition of k "factors" ln. Therefore

$$\begin{aligned} \sum_{n=1}^{N_k} \frac{1}{f(n)} &\ge \int_1^{e_k} \frac{1}{f(x)}\, dx \\ &= \int_1^{e_1} \frac{1}{x}\, dx + \int_{e_1}^{e_2} \frac{1}{x \ln x}\, dx + \cdots + \int_{e_{k-1}}^{e_k} \frac{1}{x \ln x \cdots \ln^{(k-1)}(x)}\, dx, \\ &= \ln x \Big|_1^{e_1} + \ln(\ln x) \Big|_{e_1}^{e_2} + \cdots + \ln^{(k)} x \Big|_{e_{k-1}}^{e_k} \\ &= (1-0) + (1-0) + \cdots + (1-0) = k. \end{aligned}$$

So the partial sums are unbounded, and the series diverges.

Solution to A5. Consider the "difference" vectors $\mathbf{v}_1 = (f(2) - f(1), g(2) - g(1))$, $\mathbf{v}_2 = (f(3) - f(2), g(3) - g(2)), \ldots, \mathbf{v}_{n-1} = (f(n) - f(n-1), g(n) - g(n-1))$. Each of these vectors is obtained from the previous one by rotating through $2\pi/n$. Therefore, this is also true of the "second difference" vectors $\mathbf{w}_1 = \mathbf{v}_2 - \mathbf{v}_1, \mathbf{w}_2 = \mathbf{v}_3 - \mathbf{v}_2$, $\ldots, \mathbf{w}_{n-2} = \mathbf{v}_{n-1} - \mathbf{v}_{n-2}$, and we can continue in this way. Because two vectors that are rotated from each other through $2\pi/n$ can never be equal, none of the vectors formed in this process can be $\mathbf{0}$. The process continues (with one fewer vector at each step) until we get the $(n-1)$st difference vector. If both polynomials $f(x), g(x)$ had degree $\le n-2$, the $(n-1)$st difference vector would be $\mathbf{0}$, a contradiction.

Solution to A6. The idea is that the *greedy* choice of each successive element (the choice maximizing the set of products reached so far) does at least as well as the *random* choice. Notation: If $P \subseteq G$ and $g \in G$, let $Pg := \{pg \mid p \in P\}$ and $P^{-1}g := \{p^{-1}g \mid p \in P\}$. Let $n = |G|$.

LEMMA 1. *If $P \subseteq G$ and g is chosen randomly from G, then the average size of $P \cup Pg$ is $|P| + |G \setminus P| \cdot |P|/n = 2|P| - |P|^2/n$.*

Proof of Lemma. Any particular $h \in G$ belongs to Pg for $|P^{-1}h| = |P|$ choices of g, so the probability that a particular $h \in G$ belongs to Pg is $|P|/n$. Therefore, the expected number of elements of $G \setminus P$ that lie in Pg is $|G \setminus P| \cdot |P|/n$. ∎

Let $P_0 := \{1\} \subseteq G$. For $i \ge 1$, define $s_i \in G$ and $P_i \subseteq G$ inductively as follows: choose $s_i \in G$ so as to maximize the size of $P_i := P_{i-1} \cup P_{i-1}s_i$. Note that P_i is exactly the set of products obtained from subsequences of $s_1, \ldots, s_i$. By the Lemma, $|P_i| \ge 2|P_{i-1}| - |P_{i-1}|^2/n$. Equivalently, the numbers $\alpha_i := 1 - |P_i|/n$ satisfy $\alpha_i \le \alpha_{i-1}^2$. We have $\alpha_0 = 1 - 1/n < e^{-1/n}$, so $\alpha_m < e^{-2^m/n}$. Let m be the smallest integer such that $2^m \ge n \log n$. Then $\alpha_m < e^{-\log n} = 1/n$, so $|P_m| > n - 1$. Thus $P_m = G$, so the sequence $s_1, \ldots, s_m$ has the required property. Also, $m = O(\log n)$, so the sequence has length at most $c \log |G|$ for

a suitable constant c. [See L. Babai and P. Erdős, Representation of group elements as Short Products, *Ann. Discrete Math.* **12** (1982) 27–30.]

Solution to B1. The answer is 2. This is attained by the circle having center $(\sqrt{2}, 0)$ and radius $\sqrt{3}$, which passes through $(0, 1)$ and $(0, -1)$.

Suppose that a circle with nonrational center passes through three distinct rational points P_1, P_2, P_3. Let L_1 be the perpendicular bisector of segment P_1P_2, and let L_2 be the perpendicular bisector of segment P_2P_3. If L_1 and L_2 were parallel, then P_1P_2 and P_2P_3 would be in the same line, contradicting the fact that a line can intersect a circle in at most two points. Thus L_1 and L_2 meet at the center of the circle. Elementary analytic geometry shows that L_1 and L_2 are defined by linear equations with rational coefficients, and solving this system of two equations shows that the center is a rational point, contradicting the hypothesis.

Solution to B2. The limit is -1. Computing $F_n(x)$ for the first few n suggests that for each $n \geq 0$ there exists $\alpha_n \in \mathbb{R}$ such that $F_n(x) = \frac{x^n}{n!} \ln x - \alpha_n x^n$ for all $x > 0$. We prove this by induction. For $n = 0$, it holds with $\alpha_0 := 0$. If it holds for a given n, then integration by parts yields

$$\begin{aligned} F_{n+1}(x) &= \int F_n(x)\,dx = \int \left(\frac{x^n}{n!} \ln x - \alpha_n x^n \right) dx \\ &= \frac{x^{n+1}}{(n+1)!} \ln x - \int \frac{x^n}{(n+1)!}\,dx - \int \alpha_n x^n\,dx \\ &= \frac{x^{n+1}}{(n+1)!} \ln x - \left(\frac{1}{(n+1)!\,(n+1)} + \frac{\alpha_n}{n+1} \right) x^{n+1} + C \end{aligned}$$

for some constant C. Taking the limit as $x \to 0^+$ shows that $C = 0$, so the inductive step holds with

$$\alpha_{n+1} = \frac{\alpha_n}{n+1} + \frac{1}{(n+1)!\,(n+1)}.$$

Hence $(n+1)!\,\alpha_{n+1} = n!\alpha_n + \frac{1}{n+1}$, so by induction we obtain

$$n!\,\alpha_n = H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \quad \text{for } n \geq 1.$$

Comparison with $\int_1^n \frac{1}{t}\,dt$ shows that $H_n - \ln n$ is bounded as $n \to \infty$. Therefore,

$$\lim_{n\to\infty} \frac{n!\,F_n(1)}{\ln n} = \lim_{n\to\infty} \frac{n!\,(-\alpha_n)}{\ln n} = \lim_{n\to\infty} -\frac{H_n}{\ln n} = -1.$$

Solution to B3. The answer is $\sqrt{2}/2$. The circle can be parametrized as

$$\{\mathbf{c} + \mathbf{v}\cos t + \mathbf{w}\sin t \mid t \in [0, 2\pi)\}$$

for some $\mathbf{c} = (c_1, c_2, c_3, c_4)$, $\mathbf{v} = (v_1, v_2, v_3, v_4)$, $\mathbf{w} = (w_1, w_2, w_3, w_4) \in \mathbb{R}^4$ with $\mathbf{v}$ and $\mathbf{w}$ orthogonal and of equal length. Let (R, θ) be the polar coordinates for (w_1, v_1), so that $(v_1, w_1) = (R\sin\theta, R\cos\theta)$. Then the first coordinate in the parmetrization is $c_1 + R\sin(t + \theta)$. If this is to lie in $[0, 1]$, we must have $|R| \leq 1/2$, so $v_1^2 + w_1^2 \leq 1/4$. Summing the analogous inequalities yields $|v|^2 + |w|^2 \leq 1$, so the radius $|v| = |w|$ is at most $\sqrt{1/2} = \sqrt{2}/2$. Equality is attained for $\mathbf{c} = \mathbf{0}$, $\mathbf{v} = (1/2, 1/2, 0, 0)$, $\mathbf{w} = (0, 0, 1/2, 1/2)$.

Solution to B4. Suppose not; then there exist a and b with $0 \le a < b < p^3$ such that $h(a) \equiv h(b) \bmod p^3$. By assumption, h induces an injection $\mathbb{Z}/p^2\mathbb{Z} \to \mathbb{Z}/p^2\mathbb{Z}$, so $a \equiv b \bmod p^2$. We have

$$h(a+x) = h(a) + h'(a)x + x^2 d(x)$$

for some $d(x) \in \mathbb{Z}[x]$. Take $x = p$: then because $h(a+p) \not\equiv h(a) \bmod p^2$, p does not divide $h'(a)$. Now take $x = b - a$: then $x = p^2 c$ for some $c \in \{1, \ldots, p-1\}$, so

$$0 \equiv h(b) - h(a) \equiv h'(a)p^2 c + p^4 c^2 d(p^2 c) \equiv h'(a)p^2 c \mod p^3,$$

contradicting the fact that $h'(a)$ and c are not divisible by p.

Solution to B5. If $t \in \mathbb{Z}$, then the functions $f(x) = x + t$ and $f(x) = -x + t$ satisfy the conditions. We now assume that f satisfies the conditions and prove that $f(x)$ is one of these functions.

Let $q \in \mathbb{Q}$. If n is a positive multiple of the denominator of q, then $nf(q)$ and $nf(q + \frac{1}{n})$ are integers, so $\frac{f(q+\frac{1}{n}) - f(q)}{1/n} \in \mathbb{Z}$. As $n \to \infty$ through multiples of the denominator of q, the quotients $\frac{f(q+\frac{1}{n}) - f(q)}{1/n}$ converge to $f'(q)$, so $f'(q) \in \mathbb{Z}$. If $r \in \mathbb{R}$, then we can choose $q_n \in \mathbb{Q}$ converging to r. Since f' is continuous, $f'(r) = \lim_{n\to\infty} f'(q_n)$, and because this is a limit of integers, $f'(r) \in \mathbb{Z}$. Since f' is a continuous function taking integer values, the Intermediate Value Theorem implies f' is a constant. Hence $f(x) = sx + t$ for some $s \in \mathbb{Z}$. Also, $t = f(0/1)$ has denominator 1, so $t \in \mathbb{Z}$. If $s = 0$, then $f(1/2)$ has the wrong denominator. If $|s| > 1$, then $f(1/s)$ has the wrong denominator. Thus $s = \pm 1$.

Solution to B6. Notation: $[a, b] = \{a, a+1, \ldots, b\}$. Let $S_{n,k}$ denote the set of k-limited permutations of $[1, n]$ and set $P_{n,k} = \{\sigma \in S_{n,k} \mid \sigma([2, k+1]) \not\subseteq [k+2, 2k+1]\}$.

(i) If $2 \le n \le 2k$, $P_{n,k} = S_{n,k}$.

Proof. $|[2, k+1] \cap [1, n]| > |[k+2, 2k+1] \cap [1, n]|$.

(ii) If $2k + 1 \le n$ and $\sigma \in S_{n,k} \setminus P_{n,k}$, then σ stabilizes $[1, 2k+1]$ whereupon it induces the unique involution fixing 1 and switching i and $i + k$ for $2 \le i \le k + 1$.

Proof. That $\sigma([1, 2k+1]) = [1, 2k+1]$ follows from the fact that $\sigma(1) \cup \sigma([k+2, 2k+1])$ must fill the positions in $[1, k+1]$. The specified form of $\sigma|_{[1,2k+1]}$ follows by induction on i.

(iii) $|P_{n,k}|$ is even.

Proof. In fact, we exhibit a pairing of the elements of $P_{n,k}$. For $\sigma \in P_{n,k}$, let $j \ge 2$ be minimal such that $\sigma(j) \le k + 1$. Define $\sigma' \in S_{n,k}$ by

$$\sigma'(i) = \begin{cases} \sigma(j) & \text{if } i = 1, \\ \sigma(1) & \text{if } i = j, \\ \sigma(i) & \text{otherwise.} \end{cases}$$

Then $\sigma' \in P_{n,k}$ and $\sigma'' = \sigma$.

We now establish the required result by induction on n. For $n \le 1$, $|S_{n,k}|$ is 1 and therefore odd, while for $2 \le n \le 2k$, $|S_{n,k}|$ is even by virtue of (i) and (iii). For the inductive step, we note, using (ii), that when $n \ge 2k + 1$, there is a bijection $f : S_{n,k} \setminus P_{n,k} \to S_{n-2k-1,k}$ whereby $f(\sigma)(i) = \sigma(i + 2k + 1) - 2k - 1$, and so, by (iii), $|S_{n,k}| \equiv |S_{n-2k-1,k}| \pmod 2$.

Letter to the Editor: Isosceles Dissections

I enjoyed Des MacHale's "Proof Without Words: Isosceles Dissections" in the December 2008 issue of the MAGAZINE. However, his third claim (that a triangle can be dissected into two isosceles triangles if and only if one of its angles is three times another or if the triangle is right angled) is incorrect. There is a third case to consider.

Consider an acute $\triangle ABC$, two of whose angles are x and $2x$ ($0 < x < 45°$), as illustrated in FIGURE 1a. Such a triangle can be dissected into two isosceles triangles as follows. As in FIGURE 1b, draw the cevian CD equal in length to BC so that $\triangle BCD$ is isosceles. It now follows that $\angle ACD = x$, so that $\triangle ACD$ is also isosceles.

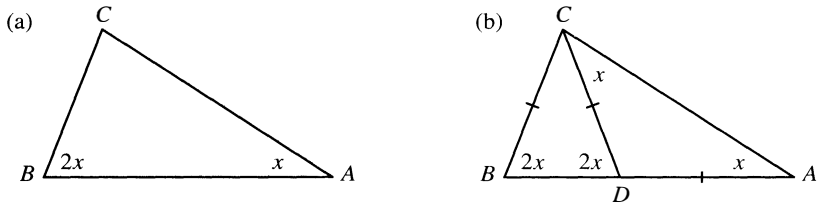


Figure 1 A triangle with one angle twice another

Two familiar members of this family are the triangles with $x = 30°$ (which makes $\triangle ABC$ a right triangle) and $x = 36°$ (when $\triangle ABC$ is itself isosceles). MacHale also showed how every acute triangle can be dissected into three isosceles triangles. The construction in FIGURE 1 can be extended to show that every isosceles obtuse triangle can also be dissected into three isosceles triangles, as in FIGURE 2.

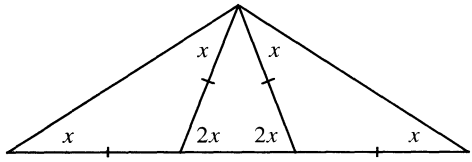


Figure 2 An obtuse isosceles triangle dissection

—ROGER B. NELSEN
Lewis and Clark College
Portland, OR 97219

New Editor of MATHEMATICS MAGAZINE

Please submit new manuscripts by email to

mathmag@maa.org

A brief message with an attached PDF file is preferred. Word-processor and DVI files can also be considered. Alternatively, manuscripts may be mailed to

Walter Stromquist
Mathematics Magazine
132 Bodine Road
Berwyn, PA 19312-1027

If possible, please include an email address for further correspondence.

Please read the editorial guidelines posted at our website, `www.maa.org/pubs/mathmag.html`. Remember that a good expository article begins with an introduction that grabs readers' attention and encourages them to keep reading. A good bibliography is also very helpful. We recommend that you search the electronic database of *Mathematics Magazine* and the *College Mathematics Journal* for articles on subjects related to yours. Articles with no references are rarely published.

If your article is accepted, we will ask you to provide (if possible) a LaTeX file using one of the templates provided at our website, along with electronic versions of figures. We will also ask for a copyright agreement, an abstract for inclusion in the database, and in the case of articles, brief biographies of the authors for publication in the MAGAZINE.

If you wish to provide one or more electronic complements to your manuscript, such as color illustrations, Java applets, or statistical datasets, please include links to these materials with your manuscript. If your article is accepted, the complements will be hosted at our site.

Finally, consider including a referee's appendix with your manuscript, not for publication, but as a guide to referees. Here you may expand on such statements as "A simple calculation shows. . . ." It is often appropriate to suppress such things in exposition, but a referee might find further details helpful.

Subscribers can access current and recent issues of the MAGAZINE electronically from our website, `http://www.maa.org/pubs/mathmag.html`. An easy shortcut to the site is `mathematicsmagazine.org`. The MAGAZINE site also offers free access to summaries of recent articles and electronic supplements prepared by authors.