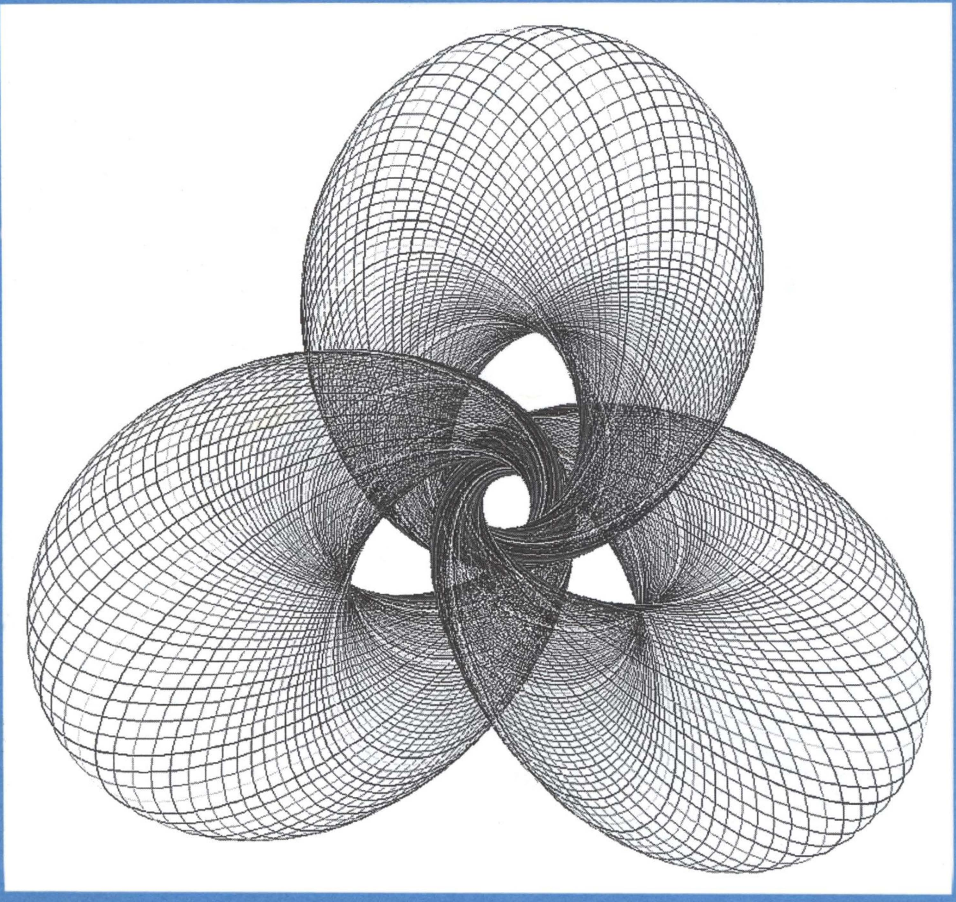
Vol. 76, No. 2, April 2003

MATHEMATICS MAGAZINE



- An Elementary Introduction to the Hopf Fibration
- C. S. Peirce and the Bell Numbers
- The Dinner-Diner Matching Problem

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

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.

Submit new manuscripts to Frank A. Farris, Editor, Mathematics Magazine, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0373. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

Cover image by David Lyons, using software by Paul Hemler, and indispensable help from Jon Pitt. One hundred circles (actual circles in space, not ellipses) lie on the surfaces of each of three linked, lopsided tori. Each of the circles is the stereographic image of a great circle of the 3-sphere. Each of the corresponding circles on the 3-sphere is a preimage set, or fiber, of a single point on the 2-sphere under the Hopf fibration (see Lyons' article).

AUTHORS

David Lyons is an Assistant Professor of Mathematics at Lebanon Valley College in Annville, Pennsylvania. Lyons received his Ph.D. from the University of North Carolina at Chapel Hill in 1996. His current research is in quantum information theory. Outside of mathematics, Lyons leads the college tae kwon do club and coaches youth soccer.

Stephen Pollard is Professor of Philosophy at Truman State University. He received his B.A. from Haverford College and his Ph.D. from the University of Texas at Austin. His opinion about the proper relationship between mathematics and philosophy is best expressed by the closing lines of one of his recent book reviews: "Philosophy will never be as much fun as mathematics; but it might help us to think about the mathematics more clearly. If it does the opposite, good riddance!"

Barbara Margolius received her Ph.D. from Case Western Reserve University in 1996 under the direction of Wojbor Woyczynski. She has masters degrees in Operations Research (CWRU) and Public Policy (University of Michigan). Since 1996 she has been teaching at Cleveland State University. Her research interests include queueing theory and classical probability. When not working on mathematics, she enjoys playing tennis poorly, and watching her sons play tennis well.

Vol. 76, No. 2, April 2003

MATHEMATICS MAGAZINE

EDITOR
Frank A. Farris
Santa Clara University

ASSOCIATE EDITORS
Glenn D. Appleby
Santa Clara University
Arthur T. Benjamin
Harvey Mudd College
Paul J. Campbell
Beloit College
Annalisa Crannell
Franklin & Marshall College
David M. James
Howard University
Elgin H. Johnston
Iowa State University
Victor J. Katz
University of District of Columbia
Jennifer J. Quinn
Occidental College
David R. Scott
University of Puget Sound
Sanford L. Segal
University of Rochester
Harry Waldman
MAA, Washington, DC

EDITORIAL ASSISTANT
Martha L. Giannini

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 Dave Riska (*driska@maa.org*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

Copyright © by the Mathematical Association of America (Incorporated), 2003, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. Permission to make copies of individual articles, in paper or electronic form, including posting on personal and class web pages, for educational and scientific use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear the following copyright notice:

Copyright the Mathematical Association of America 2003. All rights reserved.

Abstracting with credit is permitted. To copy otherwise, or to republish, requires specific permission of the MAA's Director of Publication and possibly a fee.

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

An Elementary Introduction to the Hopf Fibration

DAVID W. LYONS
Department of Mathematical Sciences
Lebanon Valley College
101 N. College Avenue
Annville, PA 17003
lyons@lvc.edu

The Hopf fibration, named after Heinz Hopf who studied it in a 1931 paper [**9**], is an important object in mathematics and physics. It was a landmark discovery in topology and is a fundamental object in the theory of Lie groups. The Hopf fibration has a wide variety of physical applications including magnetic monopoles [**14**], rigid body mechanics [**11**], and quantum information theory [**13**].

Unfortunately, the Hopf fibration is little known in the undergraduate curriculum, in part because presentations usually assume background in abstract algebra or manifolds. However, this is not a necessary restriction. We present in this article an introduction to the Hopf fibration that requires only linear algebra and analytic geometry. In particular, no vector calculus, abstract algebra, or topology is needed. Our approach uses the algebra of quaternions and illustrates some of the algebraic and geometric properties of the Hopf fibration. We explain the intimate connection of the Hopf fibration with rotations of 3-space that is the basis for its natural applications to physics.

We deliberately leave some of the development as exercises, called "Investigations," for the reader. The Investigations contain key ideas and are meant to be fun to think about. The reader may also take them as statements of facts that we wish to assume without interrupting the narrative.

Hopf's mapping

The *standard unit n-sphere* S^n is the set of points $(x_0, x_1, \ldots, x_n)$ in $\mathbb{R}^{n+1}$ that satisfy the equation

$$x_0^2 + x_1^2 + \cdots + x_n^2 = 1.$$

Geometrically, S^n is the set of points in $\mathbb{R}^{n+1}$ whose distance from the origin is 1. Thus the 1-sphere S^1 is the familiar unit circle in the plane, and the 2-sphere S^2 is the surface of the solid unit ball in 3-space. The thoughtful reader may wonder what higher dimensional spheres look like. We address this issue at the end of this article, where we explain how *stereographic projection* is used to "see" S^3.

The Hopf fibration is the mapping $h\colon S^3 \to S^2$ defined by

$$h(a, b, c, d) = (a^2 + b^2 - c^2 - d^2, 2(ad + bc), 2(bd - ac)). \tag{1}$$

To be historically precise, Hopf's original formula differs from that given here by a reordering of coordinates. We use this altered version to be consistent with the quaternion approach explained later in this article. It is easy to check that the squares of the three coordinates on the right-hand side sum to $(a^2 + b^2 + c^2 + d^2)^2 = 1$, so that the image of h is indeed contained in S^2.

What problem was Hopf trying to solve when he invented this map? And how can one see any connection with physical rotations, as we have claimed?

Hopf's paper [**9**] represented an early achievement in the modern subject of *homotopy theory*. In loose terms, homotopy theory seeks to determine those properties of a space that are not altered by continuous deformations. One way to discover the properties of an unfamiliar space X is to compare X with a familiar one Y via the set of all continuous maps $Y \to X$. Two maps whose images can be continuously deformed from one to the other are called *homotopically equivalent*. Knowing something about Y and also about the set of homotopically equivalent maps from Y to X helps us understand X. This seemingly indirect method provides a powerful way to analyze spaces.

Ironically, one of the most intractable problems in homotopy theory is to determine the homotopy equivalence classes of maps $Y \to X$ when X and Y are both spheres and the dimension of X is smaller than the dimension of Y. Many individual cases for particular pairs of dimensions of X and Y are understood, but there remain interesting unsolved problems. Hopf's map $h: S^3 \to S^2$ was a spectacular breakthrough in this area. We cannot give the full story of this discovery here, but we can explain the Hopf fibration in a geometric way that indicates its connection to rotations.

Rotations and quaternions

First, notice that a rotation about the origin in $\mathbb{R}^3$ can be specified by giving a vector for the axis of rotation and an angle of rotation about that axis. We make the convention that the rotation will be counterclockwise for positive angles, when viewed from the tip of the vector (as in FIGURE 1), and clockwise for negative angles.

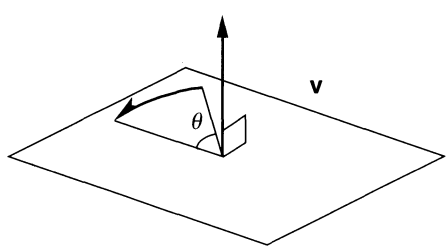


Figure 1 A rotation in $\mathbb{R}^3$ is specified by an angle θ and a vector **v** giving the axis

The specification of a rotation by an axis vector and an angle is far from unique. The rotation determined by the vector **v** and the angle θ is the same as the rotation determined by the pair $(k\mathbf{v}, \theta + 2n\pi)$, where k is any positive scalar and n is any integer. The pair $(-\mathbf{v}, -\theta)$ also determines the same rotation. Nonetheless, we see that four real numbers are sufficient to specify a rotation: three coordinates for a vector and one real number to give the angle. This is far fewer than the *nine* entries of a 3×3 orthogonal matrix we learn to use in linear algebra. In fact, we can cut the number of parameters needed to specify a rotation from four to three, for example, by giving an axis vector whose length determines the angle of rotation. However, we shall not pursue that here; it is the 4-tuple approach that turns out to be practical. Is there an efficient way to work with 4-tuples of real numbers to do practical calculations with rotations? Here are some questions that we recommend you ponder long enough to realize that they are cumbersome to answer by matrix methods. Revisit this topic after doing Investigation E below.

INVESTIGATION A. Show that the composition of two rotations is another rotation. (The composition of two rotations is the motion obtained by performing first one rotation, then the other. Show by example that order counts.) Given geometric data (axes and angles) for two rotations, how do you determine the axis and angle for their composition?

The problem of finding a convenient algebraic method for computing with rotations led William Rowan Hamilton to invent the *quaternions* in the mid-19th century. The discovery of quaternions, and Hamilton's life in general, is a fascinating bit of history. For further reading, see biographies by Hankins [**7**] and O'Donnell [**15**]. Kuipers [**10**, § 6.2 ff] gives an exposition of the rotation problem in Investigation A and its solution, beyond what appears in this section.

Hamilton was inspired by the solution to the analogous problem in two dimensions: rotations of the plane about the origin can be encoded by unit length *complex numbers*. The angle of a rotation is the same as the angle made by its corresponding complex number, thought of as a vector in $\mathbb{R}^2$, with the positive real axis. The composition of rotations corresponds to the multiplication of the corresponding complex numbers. Hamilton tried for years to make an algebra of rotations in $\mathbb{R}^3$ using ordered triples of real numbers. One day he realized he could achieve his goal using 4-tuples.

Here is Hamilton's invention: As a set (and as a vector space) the set of quaternions is identical to $\mathbb{R}^4$. The three distinguished coordinate vectors $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$ are given the names i, j, and k, respectively. The vector (a, b, c, d) is written $a + bi + cj + dk$ when thought of as a quaternion. The number a is referred to as the *real part* and b, c, and d are called the *i, j, and k parts*, respectively. Like real and complex numbers, quaternions can be *multiplied*. The multiplication rules are encapsulated by the following relations.

$$i^2 = j^2 = k^2 = -1$$

$$ij = k \quad jk = i \quad ki = j$$

The elements i, j, and k do *not* commute. Reversing the left-right order changes the sign of the product.

$$ji = -k \quad kj = -i \quad ik = -j$$

Here is a sample multiplication.

$$\begin{aligned}(3 + 2j)(1 - 4i + k) &= 3 - 12i + 3k + 2j - 8ji + 2jk && \text{(distributing)}\\ &= 2 - 12i + 3k + 2i + 8k + 2i && \text{(applying relations)}\\ &= 3 - 10i + 2j + 11k && \text{(combining terms)}\end{aligned}$$

The *conjugate* of a quaternion $r = a + bi + cj + dk$, denoted $\bar{r}$, is defined to be $\bar{r} = a - bi - cj - dk$, which resembles the complex conjugate. The *length* or *norm* of a quaternion r, denoted $\|r\|$, is its length as a vector in $\mathbb{R}^4$, $\sqrt{a^2 + b^2 + c^2 + d^2}$. (The term *norm*, when applied to quaternions, is sometimes used in other treatments to denote the *square* of the Euclidean norm defined here.).

INVESTIGATION B. What algebraic properties do the quaternions share with the real or complex numbers? How are they different? In particular, verify the following things: Show that quaternion multiplication is associative but noncommutative. (Associativity means that $p(qr) = (pq)r$ for all quaternions p, q and r.) The norm of $r = a + bi + cj + dk$ can also be written as $\|r\| = \sqrt{r\bar{r}}$. The norm has the property $\|rs\| = \|r\| \, \|s\|$ for all quaternions r and s. (Because of this, multiplying two unit length quaternions yields another unit length quaternion.) The set of unit length quaternions, viewed as points in $\mathbb{R}^4$, is the 3-sphere S^3. Each nonzero quaternion r has a *multiplicative inverse*, denoted r^{-1}, given by

$$r^{-1} = \frac{\bar{r}}{\|r\|^2}.$$

When r is a unit quaternion, r^{-1} is the same as $\bar{r}$. (Kuipers [**10**, Ch. 5] is a good source for other details about quaternion algebra.)

Here is how a quaternion r determines a linear mapping $R_r \colon \mathbb{R}^3 \to \mathbb{R}^3$. To a point $p = (x, y, z)$ in 3-space, we associate a quaternion $xi + yj + zk$. By slight abuse of notation, we will also call this p. Since the real part of p is zero we call it a *pure* quaternion. The quaternion product rpr^{-1} can be shown to be pure, and hence can be thought of as a point $x'i + y'j + z'k = (x', y', z')$ in 3-space. We define the mapping R_r by

$$R_r(x, y, z) = rpr^{-1} = (x', y', z'). \tag{2}$$

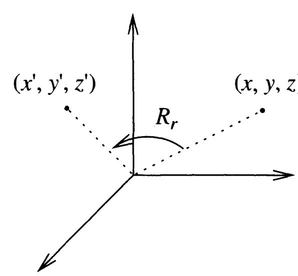


Figure 2 A nonzero quaternion r gives rise to a rotation R_r in $\mathbb{R}^3$

INVESTIGATION C. Is the mapping R_r described in the previous paragraph indeed a linear map? Verify that this is the case. Moreover, show that the map determined by any nonzero real scalar multiple of r is equal to R_r, that is, show that $R_{kr} = R_r$ for any quaternion r and any nonzero real number k. Show that when $r \neq 0$, R_r is invertible with inverse $(R_r)^{-1} = R_{(r^{-1})}$.

From the "moreover" statement in this Investigation, whenever $r \neq 0$, we are free to choose r to have norm 1 when working with the map R_r, and we shall do so since this makes the analysis simpler; we may restrict our consideration to points on the 3-sphere S^3 in order to work with rotations given by quaternions.

For $r \neq 0$, it turns out that R_r is a rotation of $\mathbb{R}^3$. The axis and angle of the rotation R_r are elegantly encoded in the four coordinates (a, b, c, d) in the following way, when r is a unit quaternion. If $r = \pm 1$, it is easy to see that R_r is the identity mapping on $\mathbb{R}^3$. Otherwise, R_r is a rotation about the axis determined by the vector (b, c, d), with angle of rotation $\theta = 2\cos^{-1}(a) = 2\sin^{-1}(\sqrt{b^2 + c^2 + d^2})$. To appreciate how nice this is, have a friend write down a 3×3 orthogonal matrix, say, with no zero entries; now find the axis and angle of rotation. You will quickly appreciate the elegance that quaternions bring to this problem, as compared with matrix methods.

The facts stated in the preceding paragraph are not supposed to be obvious. The next investigation gives a sequence of exercises that outline the proof. For a detailed discussion, see Kuipers [**10**, § 5.15].

INVESTIGATION D. How does a unit quaternion encode geometric information about its corresponding rotation? Let $r = a + bi + cj + dk$ be a unit quaternion. Verify that if $r = \pm 1$, then R_r defined above is the identity mapping. Otherwise, show that R_r is the rotation about the axis vector (b, c, d) by the angle $\theta = 2\cos^{-1}(a) = 2\sin^{-1}(\sqrt{b^2 + c^2 + d^2})$, as follows.

1. Show that R_r preserves norm, that is, that $\|R_r(p)\| = \|p\|$ for any pure quaternion $p = xi + yj + zk$. (This follows from the fact that the norm of a quaternion product equals the product of the norms.)
2. Show that the linear map R_r has eigenvector (b, c, d) with eigenvalue 1.
3. Here is a strategy to compute the angle of rotation. Choose a vector $\mathbf{w}$ perpendicular to the eigenvector (b, c, d). This can be broken down into two cases: if at least one of b and c is nonzero, we may use $\mathbf{w} = ci - bj$. If $b = c = 0$, we may use $\mathbf{w} = i$. Now compute the angle of rotation by finding the angle between the vectors $\mathbf{w}$ and $R_r\mathbf{w}$ using the following formula from analytic geometry, where the multiplication in the numerator on the right-hand side is the dot product in $\mathbb{R}^3$.

$$\cos\theta = \frac{\mathbf{w} \cdot R_r\mathbf{w}}{\|\mathbf{w}\|^2}$$

In all cases the right-hand side equals $a^2 - b^2 - c^2 - d^2 = 2a^2 - 1$. Now apply a half-angle identity to get $a = \cos(\theta/2)$.

Here is the fact that illustrates how Hamilton accomplished his goal to make an algebra of rotations.

INVESTIGATION E. Let r and s be unit quaternions. Verify that

$$R_r \circ R_s = R_{rs}.$$

In words rather than symbols: the *composition* of rotations can be accomplished by the *multiplication* of quaternions. Now go back and try Investigation A.

The next investigation is appropriate for a student who has some experience with groups, or could be a motivating problem for an independent study in the basics of group theory. (Armstrong [**2**] gives an excellent introduction to group theory with a geometric point of view.)

INVESTIGATION F. The set of unit quaternions, S^3, with the operation of quaternion multiplication satisfies the axioms of a *group*. The set of rotations in 3-space, with the operation of composition, is also a group, called SO(3). The map $\varphi\colon S^3 \to \mathrm{SO}(3)$ given by $r \mapsto R_r$ is a *group homomorphism*. Each rotation R in SO(3) can be written in the form $R = R_r$ for some $r \in S^3$ (that is, the map φ is *surjective*), and each rotation R_r has precisely two preimages in S^3, namely r and $-r$. The *kernel* of φ is the subgroup $\{1, -1\}$, and we have an *isomorphism* of groups

$$S^3/\{1, -1\} \approx SO(3).$$

The 3-sphere, rotations, and the Hopf fibration

We now reformulate the Hopf map in terms of quaternions. First, fix a distinguished point, say, $P_0 = (1, 0, 0) = i$, on S^2. (Any other point would work as well, but this one makes the formulas turn out particularly nicely.) Given a point (a, b, c, d) on S^3, let $r = a + bi + cj + dk$ be the corresponding unit quaternion. The quaternion r then defines a rotation R_r of 3-space given by (2) above. The Hopf fibration maps this quaternion to the image of the distinguished point under the rotation; in formulas, this is

$$r \mapsto R_r(P_0) = rir^{-1} = ri\bar{r}. \tag{3}$$

INVESTIGATION G. Verify that the two formulas (1) and (3) for the Hopf fibration are equivalent.

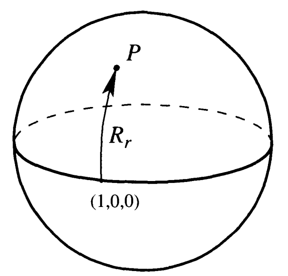


Figure 3 The unit quaternion r moves $(1, 0, 0)$ to P via R_r. The Hopf map takes r to P.

Consider the point $(1, 0, 0)$ in S^2. One can easily check that the set of points

$$C = \{(\cos t, \sin t, 0, 0) \mid t \in \mathbb{R}\}$$

in S^3 all map to $(1, 0, 0)$ via the Hopf map h. In fact, this set C is the *entire* set of points that map to $(1, 0, 0)$ via h. In other words, C is the preimage set $h^{-1}((1, 0, 0))$. You may recognize that C is the unit circle in a plane in $\mathbb{R}^4$. As we shall see, this is typical: for *any* point P in S^2, the preimage set $h^{-1}(P)$ is a circle in S^3. We will also refer to the preimage set $h^{-1}(P)$ as the *fiber* of the Hopf map over P.

We devote the remainder of this article to study one aspect of the geometry of the Hopf fibration, namely, the configuration of its fibers in S^3. Using stereographic projection (to be explained below) we get a particularly elegant decomposition of 3-space into a union of disjoint circles and a single straight line. Because this arrangement is fun to think about, we cast it first in the form of a puzzle.

INVESTIGATION H. (LINKED CIRCLES PUZZLE) Using disjoint circles and a single straight line, can you fill up 3-space in such a way that each pair of circles is linked, and the line passes through the interior of each circle?

It is the linked nature of the circles that makes this puzzle interesting. If the circles are not required to be linked, there are easy solutions. For example, just take stacks of concentric circles whose centers lie on the given line (see FIGURE 4). We will show that the Hopf fibers themselves give rise to a solution to this puzzle, but see if you can think of your own solution first.

We begin with an observation, presented in the form of an Investigation, on how to find rotations that take a given point A to a given point B.

INVESTIGATION I. Given two points A and B on S^2 that are not antipodal, how can we describe the set of *all* possible rotations that move A to B? First, choose an arc of a great circle joining A to B and call this arc $\overline{AB}$; note that the choice of arc is not unique, although the great circle is. Convince yourself that if R is a rotation taking A to B, then the axis of R must lie somewhere along the great circle bisecting $\overline{AB}$ (see FIGURE 5). Along this great circle there are two axes of rotation for which the angle of rotation is easy to compute.

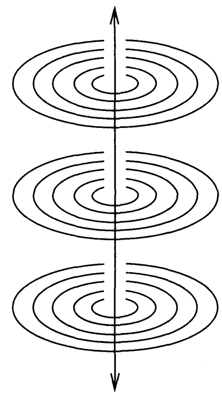


Figure 4 One way to fill $\mathbb{R}^3$ with disjoint circles and a line. Now try to arrange for every pair of circles to be linked!

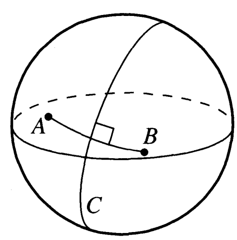


Figure 5 The axis of any rotation taking A to B must pass through the great circle C that bisects $\overline{AB}$

1. When the axis of rotation passes through the midpoint M of $\overline{AB}$, the angle of rotation θ is π radians or 180 degrees. Let us call this rotation R_1 (see the drawing on the left in FIGURE 6).
2. When the axis of rotation is perpendicular to the vectors $\mathbf{v} = \vec{OA}$ and $\mathbf{w} = \vec{OB}$, the angle of rotation is (plus or minus) the angle between $\mathbf{v}$ and $\mathbf{w}$ and is given by $\cos(\theta) = \mathbf{v} \cdot \mathbf{w}$. We will call this rotation R_2 (see the drawing on the right in FIGURE 6).

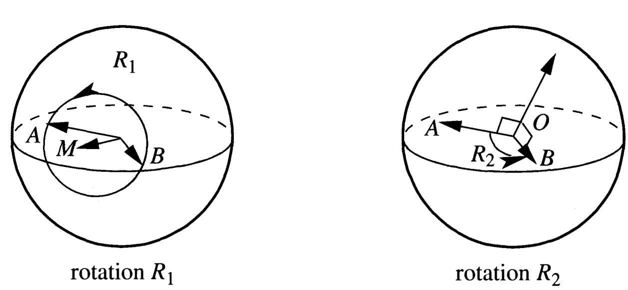


Figure 6 Two rotations taking A to B

If a point r in S^3 is sent by the Hopf map to the point P in S^2, then by Investigation G we know that the rotation R_r moves the point $(1, 0, 0)$ to P. We can use Investigation I to find the axis and angle of rotation for two rotations that map $(1, 0, 0)$ to P.

Once we have axes and angles of rotation for the rotations R_1 and R_2 of Investigation I, we can use Investigation D to find the quaternions r_1 and r_2 that map to R_1 and R_2 under the map φ, that is, $R_1 = R_{r_1}$ and $R_2 = R_{r_2}$.

INVESTIGATION J. What are explicit formulas for the quaternions r_1 and r_2 described above? For the point $P = (p_1, p_2, p_3)$ on S^2, verify that the quaternions r_1 and r_2 are given by

$$r_1 = \frac{1}{\sqrt{2(1+p_1)}}\left((1+p_1)i + p_2 j + p_3 k\right),$$
$$r_2 = \sqrt{\frac{1+p_1}{2}}\left(1 + \frac{-p_3 j}{1+p_1} + \frac{p_2 k}{1+p_1}\right).$$

Let us write e^{it} for $\cos t + i \sin t$. The fiber $h^{-1}(P)$ is given as a parametrically defined circle in $\mathbb{R}^4$ by either of the following.

$$h^{-1}(P) = \{r_1 e^{it}\}_{0\le t\le 2\pi}$$
$$h^{-1}(P) = \{r_2 e^{it}\}_{0\le t\le 2\pi}$$

The point $P = (-1, 0, 0)$ is a special case, and $h^{-1}((-1, 0, 0))$ is given by

$$h^{-1}((-1,0,0)) = \{k e^{it}\}_{0\le t\le 2\pi}.$$

Seeing the Hopf fibration

Next we demonstrate a method that allows us to *see* a little of what is going on with the Hopf fibration. Our aim is to show pictures of fibers. We do this by means of *stereographic projection*, which may be familiar to readers from an article by Delman and Galperin **[6]** in the previous issue of the MAGAZINE.

We begin by describing the stereographic projection of the 2-sphere to the x, y-plane. Imagine a light source placed at the "north pole" (0, 0, 1). Stereographic projection sends a point P on S^2 to the intersection of the light ray through P with the plane as in FIGURE 7.

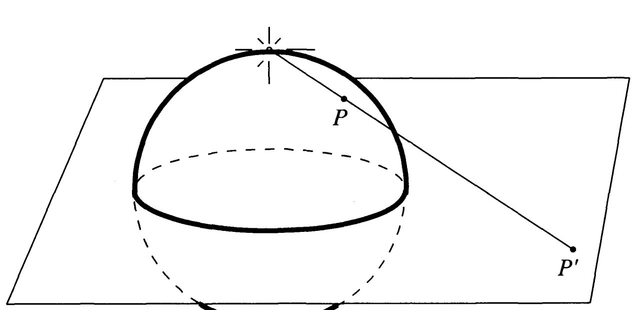


Figure 7 Stereographic projection

The alert reader will notice that the point $(0, 0, 1)$ has no sensible image under this projection. Therefore we restrict the stereographic projection to $S^2 \setminus (0, 0, 1)$.

INVESTIGATION K. Verify that the stereographic projection described above is given by

$$(x, y, z) \mapsto \left(\frac{x}{1-z}, \frac{y}{1-z}\right).$$

Write out the *inverse* map $\mathbb{R}^2 \to S^2 \setminus (0, 0, 1)$. That is, given a point (a, b) in the plane, what are the (x, y, z) coordinates of the point on S^2 sent to (a, b) by the stereographic projection? Show that a circle on S^2 that contains $(0, 0, 1)$ is mapped to a straight line in the plane. Prove that a circle on S^2 that does not pass through the point of projection $(0, 0, 1)$ is mapped by the stereographic projection to a circle in the plane. (Ahlfors [**1**, Ch. 1 § 2.4] gives a proof that stereographic projection preserves circles based on elementary geometry of complex numbers.)

Like the definition of the sphere, stereographic projection generalizes to all dimensions, and in particular, it provides a projection map $S^3 \setminus (1, 0, 0, 0) \to \mathbb{R}^3$ given by

$$(w, x, y, z) \mapsto \left(\frac{x}{1-w}, \frac{y}{1-w}, \frac{z}{1-w}\right). \tag{4}$$

The point $(1, 0, 0, 0)$ on S^3 from which we project is an arbitrary choice, but it does make the formulas simple.

The real power of stereographic projection is this: it allows us to see all of the 3-sphere (except one point) in familiar 3-space. This is remarkable because S^3 is a curved object that resides in 4-space.

The last property in Investigation K above—that stereographic projection preserves circles—holds in all dimensions [**4**, Chapter 18]. We know from the previous section that fibers of the Hopf map are circles in S^3. It follows that stereographic projection sends them to circles (or a line, if the fiber contains the point $(1, 0, 0, 0)$) in $\mathbb{R}^3$. We conclude with two Investigations that show how the stereographic images of the Hopf fibers solve the linked circles puzzle.

INVESTIGATION L. Let us denote by s the stereographic projection $s : S^3 \setminus (1, 0, 0, 0) \to \mathbb{R}^3$ given in (4). Then $s \circ h^{-1}((1, 0, 0))$ is the x-axis, $s \circ h^{-1}((-1, 0, 0))$ is the unit circle in the y, z-plane, and for any other point $P = (p_1, p_2, p_3)$ on S^2 not equal to $(1, 0, 0)$ or $(-1, 0, 0)$, $s \circ h^{-1}(P)$ is a circle in $\mathbb{R}^3$ that intersects the y, z-plane in exactly two points A and B, one inside and one outside the unit circle in the y, z-plane. This establishes that $s \circ h^{-1}(P)$ is linked with the unit circle in the y, z-plane. The points A and B lie on a line through the origin containing the vector $(0, p_3, -p_2)$. The plane of the circle $s \circ h^{-1}(P)$ cannot contain the x-axis (if it did, $s \circ h^{-1}(P)$ would intersect $s \circ h^{-1}((1, 0, 0))$, but fibers are disjoint). From these observations we can conclude that the x-axis passes through the interior of the circle $s \circ h^{-1}(P)$. See FIGURE 8.

INVESTIGATION M. To show the linked nature of *any two* circles C and D that are projections of fibers, we exhibit a continuous one-to-one map $\psi : \mathbb{R}^3 \to \mathbb{R}^3$ that takes C to the unit circle in the y, z-plane, and takes D to some other projected fiber circle. Since the image of D is linked with the unit circle in the y, z-plane, as in FIGURE 8, C and D must also be linked. (Students who have never studied topology may accept the intuitively reasonable statement that the linked nature of circles cannot be altered by a continuous bijective map, aided by FIGURE 9. Students with experience in topology may enjoy trying to prove this.)

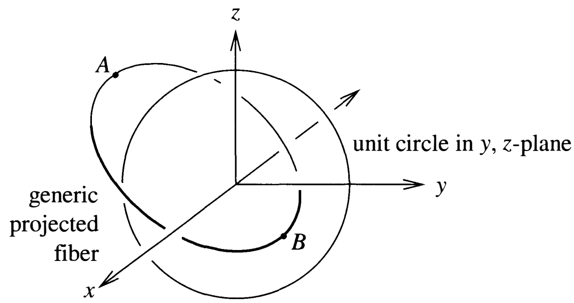


Figure 8 A generic projected Hopf fiber. A and B mark the intersections of the fiber with the y, z-plane.

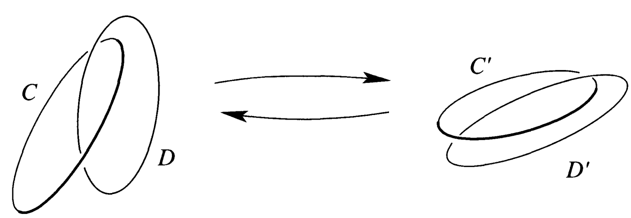


Figure 9 If the continuous bijective images C', D' of circles C, D are linked, then C and D must also be linked.

Here is how to construct the map ψ. Let P be any point on the circle C, and let $r = s^{-1}(P)$. Define $f: \mathbb{R}^4 \to \mathbb{R}^4$ by $f(x) = kr^{-1}x$ (quaternion multiplication). The map ψ is the composition $s \circ f \circ s^{-1}$.

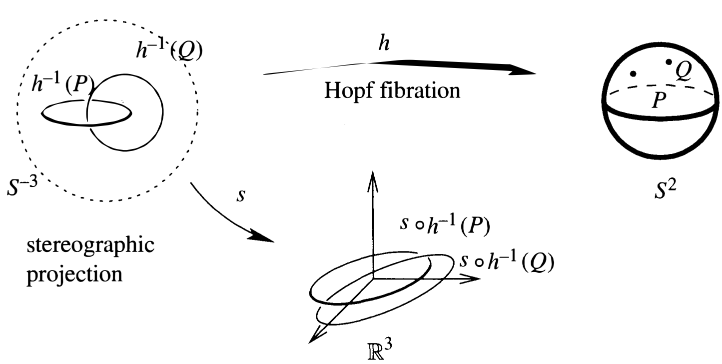


Figure 10 Stereographic projections of Hopf fibers. Any two projected fibers are linked circles, except $s \circ h^{-1}(1, 0, 0)$, which is a line.

Conclusion

We have explained how to understand the Hopf fibration in terms of quaternions. In the process, we showed how the algebra of rotations in 3-space is built into the workings of the Hopf map.

Topics raised in the Investigations suggest many lines of inquiry for independent student research. For example, making computer animations of linked Hopf fibers has been an independent study research project for two of our undergraduate students. FIGURE 11 shows an image from the software written by Nick Hamblet (see Acknowledgment below). The left panel shows a set of points lying on a circle in the codomain S^2 of the Hopf fibration. The right panel shows, via stereographic projection, the fibers corresponding to those points. An ongoing project is to build a web tutorial site featuring the animations. The reader who finds topics in this article appealing will enjoy a related article **[18]**. For general inspiration, and more on the geometry of $\mathbb{R}^3$ and rotations, see Hermann Weyl's lovely book *Symmetry* **[16]**.

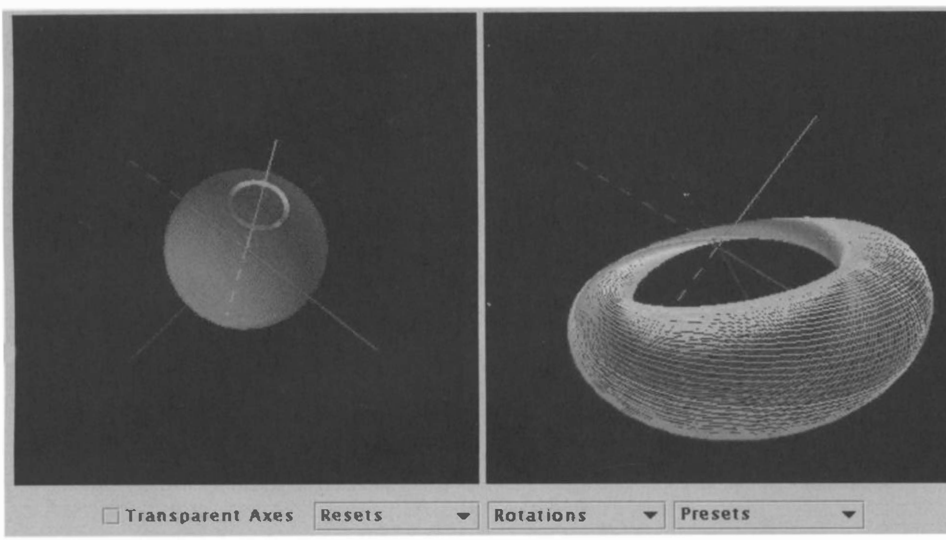


Figure 11 Screenshot of Hopf fiber software

Acknowledgment. We are grateful to Lebanon Valley College for summer support for Nick Hamblet's software development project. Nick Hamblet is a student at Lebanon Valley College, class of 2004. His work builds on joint work of the author with Paul Hemler, Professor of Computer Science, and Keely Chorn, class of 1999, both at Wake Forest University.

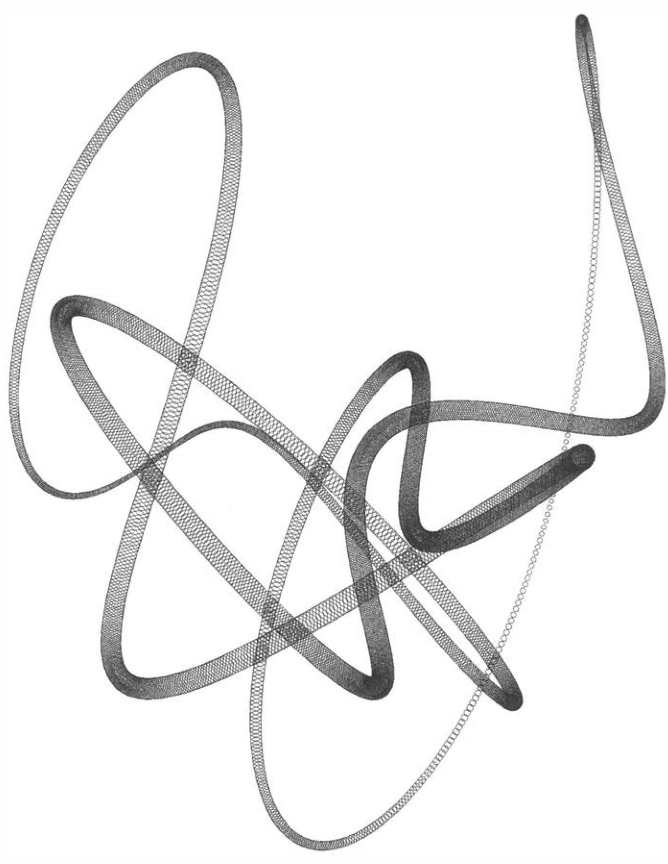
REFERENCES

1. L. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 3rd ed., 1979.
2. M. A. Armstrong, *Groups and Symmetry*, Springer-Verlag, New York, 1988.
3. T. Banchoff and J. Wermer, *Linear Algebra Through Geometry*, Springer-Verlag, New York, 1983.
4. M. Berger, *Geometry*, Springer-Verlag, New York, 1994.
5. H. S. M. Coxeter, *Introduction to Geometry*, John Wiley and Sons, New York, 2nd ed., 1989.
6. C. Delman and G. Galperin, A Tale of Three Circles, this MAGAZINE **76**:1 (2002), 15–32.
7. T. L. Hankins, *Sir William Rowan Hamilton*, Johns Hopkins University Press, Baltimore, 1980.
8. D. Hilbert, *Geometry and the Imagination*, Chelsea Publishing, New York, 1952.
9. H. Hopf, Über die Abbildungen der dreidimensionalen Sphäre auf die Kugelfläche, *Math. Ann.* **104** (1931), 637–665.
10. J. B. Kuipers, *Quaternions and Rotation Sequences*, Princeton University Press, Princeton, NJ, 1999.
11. J. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1994.
12. G. Martin, *Transformation Geometry*, Springer-Verlag, New York, 1982.

13. R. Mosseri and R. Dandoloff, Geometry of entangled states, Bloch spheres and Hopf fibrations, *J. Phys. A* **34** (2001), 10243–10252.
14. M. Nakahara, *Geometry, Topology and Physics*, Institute of Physics Publishing, Philadelphia, 1990.
15. S. O'Donnell, *William Rowan Hamilton: Portrait of a Prodigy*, Boole Press, Dublin, 1983.
16. H. Weyl, *Symmetry*, Princeton University Press, Princeton, NJ, 1980.
17. P. Zenor, E. Slaminka, and D. Thaxton, *Calculus with Early Vectors*, Prentice Hall, Upper Saddle River, New Jersey, 1999.
18. L. Zulli, Charting the 3-sphere—an exposition for undergraduates, *Amer. Math. Monthly* **103**:3 (1996), 221–229.

Mathematics Awareness Month: Mathematics and Art

Bob Brill makes art by composing simple computer algorithms that generate imagery. There are worlds of order and beauty lying dormant in our various mathematical systems, waiting to be made visible by these algorithmic processes. This is the beauty of pattern, rhythm, symmetry, asymmetry, balance, and movement. These are the worlds he explores in his art. "Mathematics," Brill says, "more than any other human activity, seems to offer connections to the underlying order of the world. This is a great inspiration for an artist and a great challenge."



"Beyond Lissajous," by Bob Brill. Visit the artist's website: see `http://users.migate.net/~bobbrill`.

For more about Math Awareness Month, see page 118.

C. S. Peirce and the Bell Numbers

STEPHEN POLLARD
Division of Social Science
Truman State University
Kirksville, MO 63501
spollard@truman.edu

So little time, so much to read! Why waste a single precious minute on a fellow who gives every indication of being a nut (and a rather cranky and unpleasant nut at that)? Sample the writings of the eccentric American polymath Charles Sanders Peirce and you will be rewarded with passage after passage that is either incomprehensible or absurd (crack-brained trash, to use two of Peirce's favorite terms of abuse). For absurdity, consider Peirce's assessment of *A Guess at the Riddle*, a book he never managed to finish, much less publish: "The undertaking which this volume inaugurates is ... to outline a theory so comprehensive that, for a long time to come, the entire work of human reason ... shall appear as the filling up of its details." Mind you: *The entire work of human reason*! As for incomprehensibility, we shall encounter a prime example in a moment. We may then feel sympathy for Simon Newcomb, the distinguished astronomer and long-time editor of the *American Journal of Mathematics*, who once complained that Peirce employed an expository style that "the human mind cannot follow."

So why read Peirce? Well, as everyone who looks into the matter eventually concedes, Peirce was ferociously smart. If he turned his mind to a problem, there was always a good chance he would dig deep. Furthermore, Peirce's very crankiness and perversity stem, in part, from an outlook and disposition that may endear him to readers of this MAGAZINE. Peirce was the Outsider (his occasional pseudonym), Emerson's American Scholar who "cannot be fed on the sere remains of foreign harvests." The literary historian David Porter's ruminations on Emily Dickinson include an observation that fits Peirce beautifully: "The solitary figure, giving the world names as if for the first time: that is quintessentially the American voice." In this special sense, Peirce was thoroughly American. He would not just dig deep, he would dig here when everyone else dug there; he would dig slant when everyone else dug straight (for example, read Dauben [**7**] for Peirce's approach to infinite sets). The many volumes of his writings are littered with insights stimulating as much for their oddness as for their acuity. When Peirce is at his best, the oddness stimulates much more than it dismays.

This essay reviews one of Peirce's successful and suggestive slantwise attacks on a classic problem: the number of partitions of a finite set. For readers who do not live and breathe combinatorics, here is a little background.

Animal teams and rhyming schemes

Last year, my daughter Emily, then seven years old, brought home an assignment that particularly attracted her father's interest. She was to select several items and then describe various ways of sorting them into families or teams. Emily chose four animals from her farm set and was happily, but haphazardly, trying out one combination after another, when her spoilsport father recommended a more systematic approach. How many ways are there of sorting the four animals into four teams? Emily quickly recognized that there is one way:

Team A: Cow
Team B: Horse
Team C: Sheep
Team D: Rooster.

How many ways are there of sorting the four animals into one team? That's easy. Again, the answer is one:

Team A: Cow, Horse, Sheep, Rooster.

How about two teams? Not quite so easy, but a little fiddling produced the answer seven. And three teams? That would be six. Now the numbers $S(n, k)$, the Stirling numbers of the second kind, give precisely the number of ways of sorting n farm animals into k teams. So we had determined that: $S(4, 1) = 1$; $S(4, 2) = 7$; $S(4, 3) = 6$; $S(4, 4) = 1$. If you sort four farm animals into teams, you have to end up with either one, two, three, or four teams. So if you add up our Stirling numbers, you obtain the total number of ways of sorting four farm animals into teams. This number is B_4, the fourth Bell number. So $B_4 = 1 + 7 + 6 + 1 = 15$. Leaving aside the farm animals, we say that there are 15 partitions of any set with exactly four members. More generally, there are B_n partitions of any set with n members.

It would be a poor sort of combinatorial object that had only one interpretation and, indeed, there are other ways of thinking about the Bell numbers. Consider the following arrangement of the farm animals:

Team A: Cow
Team B: Horse, Rooster
Team C: Sheep.

We could communicate the same information by writing

Cow: Team A
Horse: Team B
Sheep: Team C
Rooster: Team B.

Or, having fixed the order Cow, Horse, Sheep, Rooster, we could just write ABCB, which we can recognize as a rhyme scheme for a stanza of four lines.

Faith is a fine invention
For Gentlemen who see—
But Microscopes are prudent
In an Emergency!

——Emily Dickinson [**8**]

Since we can always represent a partition as a rhyme scheme and vice versa, we see that the Bell number B_n is both the number of partitions of an n-membered set and the number of rhyme schemes for a stanza of n lines. This will turn out to be important as we struggle to understand Peirce, to whom we now return.

The Bell numbers in Peirce

Peirce did not originate the Bell numbers any more that E. T. Bell did. Christian Kramp, a Strasbourg physician to whom we owe the notation $n!$ for the factorial function, beat him out by nearly a century. (For a taste of the history, see Bell [**4**], Gould [**9**], and Rota [**12**].) Our thesis is that Peirce offers a new perspective on what is old and familiar. Peirce's discussion of the Bell numbers occurs in one of his best known papers: "On the Algebra of Logic" of 1880 [**10**]. Readers of this work might, however, be forgiven for having overlooked this. Peirce introduces the topic as follows.

> A relative is a term whose definition describes what sort of a system of objects that is whose first member (which is termed the relate) is denoted by the term; and names for the other members of the system (which are termed the correlates) are usually appended to limit the denotation still further. [**10**, p. 43]

Say what? Perhaps this is the sort of prose we should expect from a stylist who once responded to an editor's insistence on a word limit by cleansing his manuscript of pronouns and articles. Further along in "Algebra of Logic," Peirce announces, without any explanation or argument, that the number of "individual forms for the $(n+2)$-fold relative" is

$$\begin{aligned}
&2 + (2^n - 1)\cdot 3 + \frac{1}{2!}\{(3^n - 1) - 2\,(2^n - 1)\}\cdot 4\\
&+ \frac{1}{3!}\{(4^n - 1) - 3\,(3^n - 1) + 3\,(2^n - 1)\}\cdot 5\\
&+ \frac{1}{4!}\{(5^n - 1) - 4\,(4^n - 1) + 6\,(3^n - 1) - 4\,(2^n - 1)\}\cdot 6\\
&+ \frac{1}{5!}\{(6^n - 1) - 5\,(5^n - 1) + 10\,(4^n - 1) - 10\,(3^n - 1) + 5\,(2^n - 1)\}\cdot 7 + \text{etc.}
\end{aligned}$$

What sense can we make of this? We are to consider certain "terms" called "relatives." These relatives come in various flavors (unary, binary, ternary, ...). The number of "forms" of the relative of a given flavor is expressed by rather a long formula. Beyond these scraps of meaning, little is to be discerned—at least, at first. Once we suppress our initial exasperation, we notice that it could be fun, even rewarding, to puzzle this out. Start with one term of Peirce's formula.

$$\frac{1}{4!}\{(5^n - 1) - 4\,(4^n - 1) + 6\,(3^n - 1) - 4\,(2^n - 1)\}\cdot 6.$$

Rewrite it as

$$\frac{6}{4!}\left\{\binom{4}{0}(5^n-1) - \binom{4}{1}(4^n-1) + \binom{4}{2}(3^n-1) - \binom{4}{3}(2^n-1) + \binom{4}{4}(1^n-1)\right\}.$$

Distribute the binomial coefficients $\binom{4}{i}$ and rearrange the terms:

$$\begin{aligned}
\frac{6}{4!}\Bigg\{&\binom{4}{0}5^n - \binom{4}{1}4^n + \binom{4}{2}3^n - \binom{4}{3}2^n + \binom{4}{4}1^n\\
&-\binom{4}{0} + \binom{4}{1} - \binom{4}{2} + \binom{4}{3} - \binom{4}{4}\Bigg\}.
\end{aligned}$$

A familiar identity with binomial coefficients is

$$-\binom{4}{0} + \binom{4}{1} - \binom{4}{2} + \binom{4}{3} - \binom{4}{4} = 0.$$

So we obtain

$$\frac{6}{4!}\left\{\binom{4}{0}5^n - \binom{4}{1}4^n + \binom{4}{2}3^n - \binom{4}{3}2^n + \binom{4}{4}1^n\right\}.$$

Rewrite again as

$$\frac{6}{4!}\sum_{t=0}^{4}(-1)^t\binom{4}{t}(4+1-t)^n.$$

Call this the *Peirce number* $P(n, 4)$ and in general define

$$P(n,k) = \frac{k+2}{k!}\sum_{t=0}^{k}(-1)^t\binom{k}{t}(k+1-t)^n.$$

Finally, notice that the number of "individual forms for the $(n+2)$-fold relative" is

$$\sum_{i=0}^{n} P(n,i).$$

Students of combinatorics may now have a feeling of *déjà vu*. Peirce's approach is reminiscent of the development of the Bell numbers via the Stirling numbers of the second kind.

$$S(n,k) = \frac{1}{k!}\sum_{t=0}^{k}(-1)^t\binom{k}{t}(k-t)^n$$

$$B_n = \sum_{i=0}^{n} S(n,i)$$

To connect the two approaches note that

$$\frac{k+2}{k!}\sum_{t=0}^{k}(-1)^t\binom{k}{t}(k+1-t)^n = \frac{k+2}{(k+1)!}\sum_{t=0}^{k+1}(-1)^t\binom{k+1}{t}(k+1-t)^{n+1}.$$

So

$$P(n,k) = (k+2)S(n+1,k+1).$$

We can make combinatorial sense of this equation. $S(n, k)$ is the number of rhyme schemes for a stanza of n lines with k rhyming syllables. So, for example, $S(3, 2) = 3$ since there are 3 rhyme schemes for a stanza of 3 lines with 2 rhyming syllables: AAB, ABA, ABB. Suppose we want to lengthen these schemes by adding a fourth line. There are two ways to do this. We could add a new occurrence of one of the 2 letters that already occur (AAB, for example, could become AABA or AABB) or we could add an occurrence of the next available letter (AAB could become AABC). The number of schemes we could obtain in this way is $3 \cdot S(3, 2)$, that is, $P(2, 1)$. More generally, $P(n, k)$ is the number of rhyme schemes we can obtain if we start with the schemes for $n+1$ lines with $k+1$ rhyming syllables and lengthen them by one line (by adding letters at the end). Every stanza of $n+2$ lines has at least 1 but no more than $n+1$ rhyming syllables in its first $n+1$ lines. So if we take the sum of all the Peirce numbers $P(n, k)$ from $k = 0$ to $k = n$, we obtain the number of rhyme schemes for a stanza of $n+2$ lines. As we have seen, this is the Bell number B_{n+2}. So

$$B_{n+2} = \sum_{i=0}^{n} P(n, i)$$

and the number of "individual forms for the $(n+2)$-fold relative" is the same as the number of rhyme schemes for a stanza of $n+2$ lines.

Perhaps a form for the $(n+2)$-fold relative is just a rhyme scheme for a stanza of $n+2$ lines (as H. W. Becker asserted long ago [**3**]). Peirce says as much in a manuscript only recently published [**11**], though thought to date from around 1889. Peirce also helps us out by "setting down a few forms."

Monads	Dyads	Triads	Tetrads
A	AA	AAA	AAAA
	AB	$\underline{\text{AAB}}$	$\underline{\text{AAAB}}$
		ABA	AABA
		ABB	AABB
			ABAA
			ABAB
			ABBA
			ABBB
		ABC	AABC
			ABAC
			$\underline{\text{ABBC}}$
			ABCA
			ABCB
			ABCC
			ABCD

Note the three lines (one in the Triad column, two in the Tetrad column) that Peirce's editors have conscientiously reproduced. We can understand them completely: they separate schemes enumerated by distinct Peirce numbers. Look at the two lines in the column of tetrads. Above the first are the $2 = P(2, 0)$ tetrads whose initial triads consist entirely of As. Between the first and second lines are the $9 = P(2, 1)$ tetrads whose initial triads consist of As and Bs. Below the second line are the $4 = P(2, 2)$ tetrads whose initial triads consist of As, Bs, and Cs. There are no other tetrads. So $2 + 9 + 4 = B_4$ is the total number of tetrads.

We can also understand Peirce's spacing convention. The spaces in each column separate schemes enumerated by distinct Stirling numbers. Above the first space in the column of tetrads is the 1 ($= S(4, 1)$) scheme consisting entirely of As. Between the first and the second spaces are the $7 = S(4, 2)$ schemes consisting of As and Bs. Between the second and the third are the $6 = S(4, 3)$ schemes consisting of As, Bs, and Cs. Below the third is the $1 = S(4, 4)$ scheme consisting of As, Bs, Cs, and Ds. Perhaps this system of lines and spaces was intended to help us see the relationship between Stirling numbers of the second kind and Peirce numbers. As we pass from column to column, the scheme above the first space doubles to form the schemes above the first line; the schemes between the first and second spaces triple to form the schemes between the first and second lines; the schemes between the second and third spaces quadruple to form the schemes between the second and third lines; and so on.

Although Peirce does not mention it (and may not have realized it), the Peirce numbers satisfy a simple recurrence relation. To put the matter somewhat differently, the Peirce numbers are the unique solution of the following system of equations (where n

and k are natural numbers):

$$P(n,0) = 2$$
$$P(0,k) = 0 \text{ if } k > 0$$
$$P(n+1,k) = \frac{k+2}{k+1}P(n,k-1) + (k+1)P(n,k).$$

To make combinatorial sense of this last equation, first rewrite it as

$$P(n+1,k) = (k+2)\left(\frac{1}{k+1}P(n,k-1) + \frac{k+1}{k+2}P(n,k)\right).$$

We know that

$$P(n+1,k) = (k+2)S(n+2,k+1).$$

So we want to show that

$$\left(\frac{1}{k+1}P(n,k-1) + \frac{k+1}{k+2}P(n,k)\right) = S(n+2,k+1).$$

The right-hand side is the number of rhyme schemes for a stanza of $n+2$ lines with $k+1$ rhyming syllables. Divide up these schemes according to the number of rhyming syllables in their first $n+1$ lines. There are only two possibilities.

- Case 1: in the first $n+1$ lines, k rhyming syllables occur. Then a brand new rhyming syllable must occur in line $n+2$. This gives us the numerator of the fraction $1/(k+1)$. The denominator $k+1$ is just the number of syllables that can normally be used to lengthen a stanza that already features k syllables. The Peirce number $P(n,k-1)$ is the number of rhyme schemes we could have obtained if we had been able to employ all those $k+1$ syllables.
- Case 2: in the first $n+1$ lines, $k+1$ rhyming syllables occur. Then one of the $k+1$ rhyming syllables already present must occur in line $n+2$. This gives us the numerator of the fraction $(k+1)/(k+2)$. The denominator $k+2$ is the number of syllables that can normally be used to lengthen a stanza that already features $k+1$ rhyming syllables. The number of rhyme schemes we could have obtained if we had been able to employ all those $k+2$ syllables is $P(n,k)$.

Our recurrence relation makes it easy to generate Peirce numbers. Here is a table with a few values of $P(n,k)$.

n \ k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	B_{n+2}
0	2	0	0	0	0	0	0	0	0	0	2
1	2	3	0	0	0	0	0	0	0	0	5
2	2	9	4	0	0	0	0	0	0	0	15
3	2	21	24	5	0	0	0	0	0	0	52
4	2	45	100	50	6	0	0	0	0	0	203
5	2	93	360	325	90	7	0	0	0	0	877
6	2	189	1204	1750	840	147	8	0	0	0	4140
7	2	381	3864	8505	6300	1862	224	9	0	0	21147
8	2	765	12100	38850	41706	18522	3696	324	10	0	115975
9	2	1533	37320	170525	255150	159789	47040	6750	450	11	678570

By summing each row, we obtain the Bell numbers B_2 through B_{11}. Peirce also noticed a much simpler way to generate Bell numbers. He recognized that if we write down Bell numbers B_1 through B_{n+1} and construct their difference table, we obtain B_n. For example, B_1 through B_6 yield B_5:

$$\begin{array}{ccccccccccc}
1 & & 2 & & 5 & & 15 & & 52 & & 203 \\
& 1 & & 3 & & 10 & & 37 & & 151 & \\
& & 2 & & 7 & & 27 & & 114 & & \\
& & & 5 & & 20 & & 87 & & & \\
& & & & 15 & & 67 & & & & \\
& & & & & 52 & & & & &
\end{array}$$

(The history of the Bell numbers is a tricky business. Comtet [**6**] attributes this result to A. C. Aitken [**1**] even though Peirce reached it five decades earlier.) Peirce observed that we can now obtain, by reflection, a table of sums in which the first n Bell numbers yield B_{n+1}:

$$\begin{array}{ccccccccccc}
1 & & 1 & & 2 & & 5 & & 15 & & 52 \\
& 2 & & 3 & & 7 & & 20 & & 67 & \\
& & 5 & & 10 & & 27 & & 87 & & \\
& & & 15 & & 37 & & 114 & & & \\
& & & & 52 & & 151 & & & & \\
& & & & & 203 & & & & &
\end{array}$$

Since the binomial coefficient function gives the number of times that a number in row one contributes to the final sum, this is equivalent to the well-known formula

$$B_{n+1} = \sum_{t=0}^{n} \binom{n}{t} B_t.$$

Peirce noted that this summation property provides the following algorithm for generating the Bell numbers. Write down two 1s and record their sum below them.

$$\begin{array}{ccc}
1 & & 1 \\
& 2 &
\end{array}$$

Copy down the new term in row one and again take sums.

$$\begin{array}{ccccc}
1 & & 1 & & 2 \\
& 2 & & 3 & \\
& & 5 & &
\end{array}$$

Repeat this process to obtain as many Bell numbers as you wish. (Emily entertained herself by filling up an entire sheet of paper in this way.) Peirce may not have been the first mathematician to recommend this technique. He was certainly not the last: it was rediscovered as recently as 1962 [**2, 5**].

Conclusion

Peirce seems to have thought that his formula for the Bell numbers was original. He notes with apparent pride [**11**] that the number of forms of a given plurality (that is, the number of rhyme schemes for a stanza of a given length), "... has the value given by me." He then cites his article in the *American Journal of Mathematics* [**10**]. While we know the Bell numbers were studied before Peirce, I have not uncovered evidence that the Peirce numbers were anticipated by anyone. Since the sequence of Peirce numbers (2, 2, 3, 2, 9, 4, 2, 21, 24, 5, ...) appeared only recently (April 10, 2002) on N. J. A. Sloane's list of more than 67,000 integer sequences **[13]**, with no references to any literature, one suspects that they have not been the objects of much study. Really, though, questions of priority are not our main interest here.

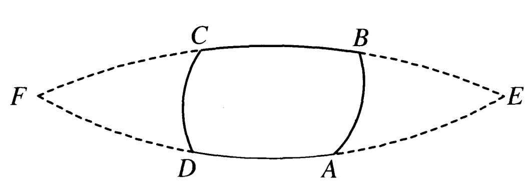
Peirce offers an approach to the Bell numbers that is probably new to most students of combinatorics. As a bonus he sets us some puzzles that are stimulating and, once we start to make some progress, even fun. Fresh, stimulating, diverting: I suspect this is how most mathematicians would view Peirce if only they could overcome their initial feelings of disorientation. There are many gems to be mined. I have given just one example. See what you can find!

REFERENCES

1. A. C. Aitken, A problem on combinations, *Edinburgh Math. Notes* **28** (1933), 18–33.
2. H. W. Becker, Solution of problem E 461, *Amer. Math. Monthly* **48** (1941), 701–703.
3. H. W. Becker, Combinatory interpretations of Bell's numbers, *Bull. Amer. Math. Soc.* **52** (1946), 415 (Abstract # 106).
4. E. T. Bell, The Iterated Exponential Integers, *Ann. of Math.* **39** (1938), 539–557.
5. M. Cohn, S. Even, K. Menger, and P. K. Hooper, On the number of partitionings of a set of n distinct objects, *Amer. Math. Monthly* **69** (1962), 782–785.
6. L. Comtet, *Advanced Combinatorics*, D. Reidel, Dordrecht/Boston, 1974.
7. J. W. Dauben, C. S. Peirce's philosophy of infinite sets, this MAGAZINE **50** (1977), 123–135.
8. Emily Dickinson, *The Manuscript Books of Emily Dickinson*, vol. 1, R.W. Franklin, ed., The Belknap Press of Harvard University Press, Cambridge, MA, 1981, pp. 186 & 230.
9. H. W. Gould, *Research Bibliography of Two Special Number Sequences*, Combinatorial Research Institute, Morgantown, West Virginia, 1976.
10. C. S. Peirce, On the algebra of logic, *Amer. J. Math.* **3** (1880), 15–57. Reprinted in *Writings of Charles S. Peirce*, Vol. 4, Indiana Univ. Pr., Bloomington, Indiana, 1986, pp. 163–209.
11. C. S. Peirce, On the numbers of forms of sets, *Writings of Charles S. Peirce*, Vol. 6, Indiana Univ. Pr., Bloomington, Indiana, 2000, pp. 360–362.
12. G.-C. Rota, The number of partitions of a set, *Amer. Math. Monthly* **71** (1964), 498–504.
13. N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, `http://www.research.att.com/~njas/sequences/`.

Additional solution to Q929 on page 152. No peeking!

A929. *Solution II.* Let the vertices of the quadrilateral be labeled $ABCD$. Extend the great circle arcs containing AD and BC so they meet in points E and F, as shown in the accompanying figure. Because $\angle BEA \cong \angle DFC$, $\angle ABE \cong \angle CDF$, and $\angle BAE \cong \angle DCF$, it follows from the AAA congruence theorem for spherical triangles that $\triangle ABE \cong \triangle CDF$. Thus $AB = CD$. Because the two arcs of the lune $EBCFDA$ are equal in length, it follows that $AD = CB$.



The Dinner-Diner Matching Problem

BARBARA H. MARGOLIUS
Cleveland State University
Cleveland, Ohio 44115
b.margolius@csuohio.edu

At the Spring 2000 meeting of the Ohio section of the MAA in Huntington, West Virginia, there was a banquet for NExT fellows, untenured faculty in the section's faculty development program. Each attendee had selected a dinner from among four entrees. The following orders had been placed by the twenty-two guests: 1 pasta vegetarian, 8 chicken cordon bleu, 6 prime rib, and 7 filet of sole dinners. None of the guests could say with certainty what they had ordered for dinner, leaving the server greatly distressed.

DINNER-DINER MATCHING PROBLEM. If the dinners are served randomly, what is the probability distribution of the number of diners who are served what they ordered?

This is a matching problem, or a problem with restricted position. Such problems have a long history in classical probability and are often described as card-matching problems. Barton [**2**] gives an extensive history of these through 1958. Also see the article by Knudsen and Skau [**11**], appearing recently in this MAGAZINE, and references therein. Generalizations of this problem were solved by Greville in 1941 [**8**] and Anderson in 1943 [**1**].

Rook numbers and the exact distribution

It is possible to compute the exact distribution for this matching problem. Each possible assignment of dinners to diners may be represented uniquely on a square chessboard with the entrees (or dinners) represented by columns and the diners represented by rows. Assignments of dinners to diners are represented in the square below by 22 dots, no two in the same row or column. These dots may be thought of as rooks. Since a rook in chess may capture (or take) a piece only in its own row or column, the rooks arranged in this way are referred to as *nontaking* or *nonattacking rooks*. The small circles in diagonal blocks of cells represent assignments that would match a diner to the dinner that he ordered.

The configuration shown has no diners receiving the dinners that they ordered. Let N be the total number of diners, and n_i be the number who ordered the ith entree of m entrees, so that $\sum_{i=1}^{m} n_i = N$. In the example, $n_1 = 1$, $n_2 = 8$, $n_3 = 6$, and $n_4 = 7$. There are $m = 4$ entrees and the total number of dinners ordered is $N = \sum_{i=1}^{m} n_i = 22$. Of the eight diners who ordered chicken, one is served pasta, three are served prime rib, and four filet of sole. There are 22! ways of placing 22 nonattacking rooks on this board. If the dinners are assigned randomly, then we consider each of these arrangements of rooks on the board (of assignments of dinners to diners) to be equally likely.

Denote by $X(N)$ the number of matches of dinners to the diners who ordered them. If A_i denotes the event that there is a match of the ith diner to his order

TABLE 1: Placement of $\sum_{i=1}^{m} n_i = N$ nonattacking rooks

	Type of Entree Ordered																					
Diners by Entree Ordered	Pasta	Chicken								Prime Rib						Filet of Sole						
Pasta	○			●																		
Chicken		○	○	○	○	○	○	○	○								●					
	●	○	○	○	○	○	○	○	○													
		○	○	○	○	○	○	○	○									●				
		○	○	○	○	○	○	○	○												●	
		○	○	○	○	○	○	○	○				●									
		○	○	○	○	○	○	○	○											●		
		○	○	○	○	○	○	○	○			●										
		○	○	○	○	○	○	○	○	●												
Prime Rib										○	○	○	○	○	○				●			
								●		○	○	○	○	○	○							
										○	○	○	○	○	○	●						
							●			○	○	○	○	○	○							
									●	○	○	○	○	○	○							
										○	○	○	○	○	○							●
Filet of Sole					●											○	○	○	○	○	○	○
			●													○	○	○	○	○	○	○
						●										○	○	○	○	○	○	○
		●														○	○	○	○	○	○	○
											●					○	○	○	○	○	○	○
														●		○	○	○	○	○	○	○
															●	○	○	○	○	○	○	○

($i = 1, 2, \ldots, N$), then by the inclusion-exclusion theorem, we have

$$\mathbf{P}\{X(N) = j\} = \sum_{s=j}^{N} (-1)^{s-j} \binom{s}{j} B_s(N) \tag{1}$$

where

$$B_s(N) = \sum_{1 \le j_1 < j_2 < \cdots < j_s \le N} \mathbf{P}\{A_{j_1} A_{j_2} \ldots A_{j_s}\}. \tag{2}$$

We can derive a formula for the numbers $B_s(N)$ using the chessboard analogy. Suppose that we are able to count the number of ways that j of the N nonattacking rooks may be placed in the diagonal blocks with circles without regard to the placement of the other $N - j$ rooks, and that this number, called a rook number, is r_j. In the example, $r_1 = 1 + 8^2 + 6^2 + 7^2 = 150$ because there are 150 squares with circles in them that indicate a diner has been served the dinner that he ordered. If j rooks are placed in the diagonal blocks that represent matches of dinners to diners, then there are $(N - j)!$ ways that the other rooks may be placed without regard to whether their placement results in a match or not. Because there is no restriction on the placement of the remaining $N - j$ rooks, the number $r_j(N - j)!$ is an overcount of the number of ways of placing j nonattacking rooks in the restricted positions on the chessboard that represent matches of diners to the dinners that they ordered. Since there are a total of

$N!$ possible placements of the rooks so that they are nonattacking, the quantity $B_s(N)$ can be computed in terms of these rook numbers and is given by

$$B_s(N) = \frac{r_s(N-s)!}{N!}.$$

Substituting in to (1) the probability of exactly j matches is

$$\mathbf{P}\{X(N) = j\} = \sum_{s=j}^{N}(-1)^{s-j}\binom{s}{j} r_s(N-s)!/N!. \tag{3}$$

If we know what the rook numbers r_j are, then we can compute the probability distribution exactly. Consider an $n_i \times n_i$ chessboard. There are n_i^2 ways of placing the first rook, which may go anywhere on the board, $(n_i - 1)^2$ ways of placing the second rook which may go anywhere on the board except the row and column the first rook was placed in, and continuing in this manner, there are $(n_i - (j-1))^2$ ways of placing the jth nonattacking rook. There are $j!$ orders in which the rooks may be placed, so for an $n_i \times n_i$ board, the jth rook number, which we will designate $r_j^{(n_i)}$, is

$$r_j^{(n_i)} = \frac{n_i^2(n_i-1)^2\cdots(n_i-(j-1))^2}{j!} = \binom{n_i}{j}^2 j!.$$

We are not interested in an $n_i \times n_i$ board, but rather in one like the example. The $n_i \times n_i$ diagonal blocks indicate restricted positions. The rook numbers for this more complicated board will give the number of ways of placing j nonattacking rooks in any of the $\sum_{i=1}^{m} n_i^2$ restricted positions. Because placement of a rook in one of the blocks does not limit our ability to place a rook in one of the other blocks (the separate blocks of restricted positions do not share any rows or columns), we may compute the rook numbers by summing all possible products of the rook numbers for the individual blocks such that the indices sum to j, so

$$r_j = \sum_{s_1+s_2+\cdots+s_m=j} \prod_{i=1}^{m} r_{s_i}^{(n_i)},$$

where s_i is the number of rooks placed in restricted positions in the ith block, and the s_i form an ordered partition of j. To get these numbers, we may write down the rook polynomial

$$R(x) = \prod_{i=1}^{m}\left(\sum_{j=1}^{n_i}\binom{n_i}{j}^2 j!\,x^j\right). \tag{4}$$

The rook number, r_j, is the jth coefficient of this polynomial. More background on rook polynomials is available from a variety of sources **[19, 21, 22, 26]**. We now have enough information to answer the problem posed at the beginning of this article. We used *Maple* to calculate the generating function for the probabilities, where the coefficient of t^k is the probability of k matches.

$$\frac{1}{7682154480}t^{22} + \frac{167}{7682154480}t^{20} + \frac{241}{1920538620}t^{19} + \frac{12161}{7682154480}t^{18}$$
$$+\frac{71}{7034940}t^{17} + \frac{1549}{24387792}t^{16} + \frac{2801}{9145422}t^{15} + \frac{92681}{73163376}t^{14} + \frac{118843}{27436266}t^{13}$$
$$+\frac{4569883}{365816880}t^{12} + \frac{1660069}{54872532}t^{11} + \frac{6153893}{99768240}t^{10} + \frac{117257}{1110780}t^{9}$$

$$+\frac{386783941}{2560718160}t^8+\frac{23978}{133705}t^7+\frac{44741581}{256071816}t^6+\frac{21988283}{160044885}t^5$$

$$+\frac{326784389}{3841077240}t^4+\frac{1826989}{45727110}t^3+\frac{7339561}{548725320}t^2+\frac{78025}{27436266}t+\frac{2881}{9976824}$$

We use these probabilities to produce FIGURE 1. The dinner probability distribution is shown together with a normal distribution with the same mean and variance, and a Poisson distribution with the same mean.

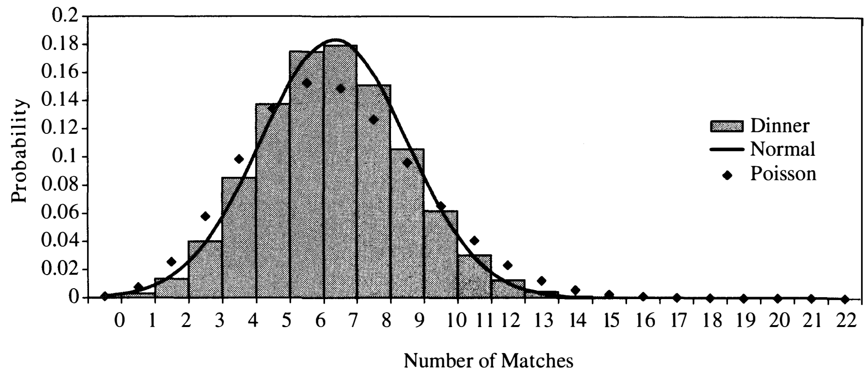


Figure 1 Ohio NExT dinner-diner matching probabilities

Approximating distributions: Normal and Poisson

The close match between the exact distribution and a normal distribution suggests that the dinner-diner probability distribution may be well approximated by the normal distribution. In the present section, we explore the asymptotic behavior of these distributions.

We have been considering the Ohio NExT dinner problem in which the number of entrees ordered varies with the entree. Suppose we are at a dinner party with m entrees to choose from, and each entree has been ordered by k diners (so $n_i = k$ for each type of entree i). We can model the cases where m is four and k is thirteen (or k is four and m is thirteen) with two decks of playing cards. Lay the first deck out in order by suits and within suits by rank. Shuffle the second deck and lay it out next to the first. A match occurs when two cards of the same suit (same rank) are side by side. (See FIGURES 3 and 4 at the end of this note for probabilities related to this example.) Knudsen and Skau **[11]** considered this type of matching problem and obtained the limiting distribution as m, the number of ranks (the number of types of entrees), tends to infinity. In the theorem below, $p_j^{(k,m)}$ is the probability of exactly j matches when there are m ranks and k of each rank and a match occurs when adjacent cards are of the same rank. The theorem shows that this distribution is asymptotically Poisson with mean k equal to the number of cards of each rank.

THEOREM 1. (KNUDSEN-SKAU) *If we have two decks of cards each with mk cards, m ranks and k cards of each rank, the limit of the number of matches is Poisson with parameter $\lambda = k$, that is*

$$\lim_{m\to\infty} p_j^{(k,m)} = \frac{k^j}{j!}e^{-k}, \quad j = 0, 1, 2 \ldots .$$

The previous theorem is a special case of their results. They showed that the limiting distribution is Poisson when the restricted positions are $k \times l$ rectangles. Barton [**2, 5**] proved a more general result.

THEOREM 2. (BARTON) *If we have two decks of cards each with N cards and m suits, the first deck with n_i cards of suit i and the second deck with m_i cards of suit i so that $\sum_{i=1}^{m} n_i = \sum_{i=1}^{m} m_i = N$, all n_i and m_i bounded, and we let m and N tend to infinity so that*

$$\lambda = \frac{1}{N}\sum_{i=1}^{m} n_i m_i$$

is constant, then the limit of the number of matches is Poisson with parameter λ.

The proof of this theorem is based on the moment convergence theorem [**25**]. In the particular case of the Poisson distribution, this was proved in 1921 by Mises [**16**]. Barton shows that the factorial moments of this distribution

$$\mu_{[j]} = E(X(X-1)\cdots(X-j+1)) \tag{5}$$

tend to λ^j as $m, N \to \infty$. These are the factorial moments of the Poisson distribution, so this card-matching distribution is asymptotically Poisson.

We are concerned with the case where the restricted positions are $k \times k$ squares, which we will generalize to the case where the restricted positions are squares of varying sizes. Knudsen and Skau consider the limit as the number of squares (number of choices of entrees) tends to infinity. We consider the limit where the number of squares is fixed, but their size (the number of guests who have chosen each entree) tends to infinity. They conjecture that "$k = m$ represents the actual breakpoint for the distribution to be *asymptotically* Poisson," [**11**]. David and Barton [**5**] address this question as well.

The answer to this question is contained in the following theorem.

THEOREM 3. *The random variable*

$$Y(N) = \frac{X(N) - k}{k\sqrt{(m-1)/(mk-1))}}$$

has asymptotically normal distribution as $k \to \infty$ with parameters $(0, 1)$, where $X(N)$ is a random variable representing the number of matches for the dinner problem with a choice of m entrees with k diners having selected each entree. The expectation of $X(N)$ is

$$\mathbf{E}\{X(N)\} = \mu = k,$$

and its variance is

$$\sigma^2 = \frac{k^2(m-1)}{(mk-1)}.$$

Remark

Note that a Poisson random variable has equal mean and variance, but the variance for the dinner distribution is strictly less than the mean for $k > 1$. When k is large,

$$\sigma^2 \approx k\left(1 - \frac{1}{m}\right). \tag{6}$$

For large m, $(1 - \frac{1}{m})$ is close to one so the mean and variance are nearly equal, and the Poisson distribution, which is approximately normal for large values of its parameter, is a good approximation to the dinner distribution for large k. If m is small, then the Poisson distribution is a poor approximation for the dinner distribution. See `http://academic.csuohio.edu/bmargolius/dinner/cards.htm` for graphs for various dinner-diner matching distributions that illustrate this idea.

To prove this theorem, we first find the mean and variance of the distribution, and then use an idea from statistics to show that the distribution is asymptotically normal. By (1) and (2) we have

$$\mathbf{E}\left\{\binom{X(N)}{s}\right\} = B_s(N) \tag{7}$$

for $s = 0, 1, 2, \ldots, N$. $B_s(N)$ is the sth binomial moment of $X(N)$. If we know the binomial moments of $X(N)$, then the moments of $\mathbf{E}\{[X(N)]^s\}$ (see, for example, K. Jordan [**10**, Ch. II, sec. 17]) are given by

$$\begin{aligned} \mathbf{E}\{[X(N)]^s\} &= \sum_{k=1}^{s} S^*(s,k)k!\, B_k(N) \\ &= \sum_{k=1}^{s} S^*(s,k)k!\,(N-k)!\, r_k/N! \end{aligned} \tag{8}$$

for $s = 1, 2, \ldots, N$ where $S^*(s, k)$ are Stirling's numbers of the second kind. First, we will compute the mean and variance of $X(N)$ using this formula. The number of ways of placing one rook in a restricted position is equal to the number of restricted positions, so $r_1 = mk^2$, and

$$\mu = r_1/(mk) = k.$$

To find the variance, we need r_2. Observe that there are m ways of placing two rooks in the same block, and $\binom{m}{2}$ ways of placing one rook in each of two blocks, so the second rook number,

$$\begin{aligned} r_2 &= 2!\,m\binom{k}{2}^2 + 1!\binom{m}{2}\binom{k}{1}^4 \\ &= \frac{mk^2}{2}(mk^2 - 2k + 1). \end{aligned} \tag{9}$$

From equations (8) and (9) we have

$$\sigma^2 = \mathbf{E}\left\{[X(N)]^2\right\} - \mu^2 = \frac{k^2(m-1)}{(mk-1)}.$$

To complete the proof, we use an idea from statistics. Recall that a bivariate contingency table is a table in which cell entries represent frequencies (or counts) corresponding to two variables. In our case, columns represent diners categorized by what they ordered, and rows represent diners categorized by what they are served. The data from the chessboard at the beginning of this note would be displayed in a contingency table as shown below.

More generally, the dinner-diner assignments can be displayed on an $m \times m$ bivariate contingency table with fixed marginals (that is, fixed row and column totals). The

TABLE 2: Contingency table of assignment of dinners to diners

	Pasta	Chicken	Prime Rib	Filet of Sole	Total
Pasta	0	1	0	0	1
Chicken	1	0	3	4	8
Prime Rib	0	3	0	3	6
Filet of Sole	0	4	3	0	7
Total	1	8	6	7	22

row and column totals are fixed because we know how many guests ordered each entree, and we know how many there are of each entree. When we have m entrees each ordered by k diners, the row totals and column totals are k for each of m rows and each of m columns. The probability of having $n_{i,j}$ diners who ordered entree i but receive entree j, for $i = 1, \ldots, m$, $j = 1, \ldots, m$ is

$$\frac{k!^{2m}}{(mk)! \prod_{i,j} n_{i,j}!}, \tag{10}$$

the same as the probability of having entries of $n_{i,j}$, for $i = 1, \ldots, m$, $j = 1, \ldots, m$, in a contingency table with fixed row and column totals equal to the number of diners who ordered each entree. The distribution of the cell entries is known to be asymptotically jointly normal when row and column totals are fixed proportions of the grand total, N, and $N \to \infty$; see, for example, Roy and Mitra **[20]** for a discussion of the asymptotic distribution of cell entries in contingency tables with various combinations of fixed row and column sums, or Lancaster **[12, 13]**. Because the distribution of cell entries in a contingency table is jointly normally distributed in the limit as the grand total $N \to \infty$, and the row and column totals remain a fixed proportion of N, the sum of the entries in the diagonal cells will also be normally distributed. The sum of the entries in the diagonal cells is the number of matches of dinners to diners. This completes the proof.

The asymptotic behavior of cell entries in a contingency table is not restricted to tables with equal row and column totals, so our proof extends to the more general dinner-diner distribution where there are different numbers of guests ordering each entree. So long as the proportion of guests ordering each entree remains fixed as total orders tends to infinity, we have the following result:

THEOREM 4. *The random variable*

$$Y(N) = \frac{X(N) - \mu}{\sigma}$$

has asymptotically normal distribution as $k \to \infty$ *with parameters* $(0, 1)$, *where* $X(N)$ *is a random variable representing the number of matches for the dinner problem with a choice of* m *entrees with* kn_i *diners having selected the* i*th entree, and* $N = k\sum_{i=1}^{m} n_i$. *The expectation of* $X(N)$ *is*

$$\mu = \frac{k\sum_{i=1}^{m} n_i^2}{\sum_{i=1}^{m} n_i}, \tag{11}$$

and the variance is

$$\sigma^2 = \frac{k^2\left(\left(\sum_{i=1}^{m} n_i^2\right)^2 - 2\left(\sum_{i=1}^{m} n_i\right)\left(\sum_{i=1}^{m} n_i^3\right) + \sum_{i=1}^{m} n_i^2 \left(\sum_{i=1}^{m} n_i\right)^2\right)}{\left(\sum_{i=1}^{m} n_i\right)^2 \left(k\sum_{i=1}^{m} n_i - 1\right)}.$$

The idea of extending the contingency table approach to more general card-matching problems is applied by David and Barton [**5**], and by Anderson [**1**] who prove more general results.

The data in Tables 1 and 2 for the Ohio NExT dinner problem are related as follows: The probability for the assignment shown in Table 1, the chessboard, was $1/22! \approx 8.9 \times 10^{-22}$, but the probability of the assignment shown in the contingency table is greater and is given by the product of factorials of the row and column totals divided by the product of the factorials of the cell entries and the grand total factorial, (this is formula (10) generalized to unequal row and column totals):

$$\frac{1!^2\, 6!^2\, 7!^2\, 8!^2}{22!\, 3!^4\, 1!^2\, 4!^2} = 350/13718133 \approx 2.55 \times 10^{-5}.$$

The probability is greater because there are 28,677,390,336,000,000 different chessboard arrangements of nonattacking rooks that result in this particular contingency table.

The expected number of entries in row i, column j, is

$$E(n_{i,j}) = \frac{n_i n_j}{N},$$

where n_i is the fixed sum for row i and n_j is the fixed sum for column j. The expected number of matches of dinners to diners is the sum of the expectations of the diagonal cells. In the Ohio NExT dinner example, we have $(1^2 + 8^2 + 6^2 + 7^2)/22 = 150/22 \approx 6.82$. More generally, the expected number of matches is given in theorem 4:

$$\mu = \frac{\sum_{i=1}^{m} n_i^2}{\sum_{i=1}^{m} n_i},$$

(substitute 1 for the scaling factor k in equation (11) from theorem 4). In the dinner-diner problem, we have considered only the case where the sum of the ith row is equal to the sum of the ith column. This restriction is not necessary. See [**1**] and [**5**] for the more general case.

Examples

In the examples that follow reference is made where applicable to *The on-line encyclopedia of integer sequences* (EIS) [**23**]. This is an amazing research tool that provides easy access to an extensive database of integer sequences. Sequences can be looked up in this reference by entering at least three consecutive terms. The citation provides more terms of the sequence, a name for the sequence and often references, a generating function, a formula for the nth term, hyperlinks, computer code for generating the sequence and more. Many sequences appear in multiple contexts. The EIS helps researchers to make those connections. (In what follows, A followed by 6 digits, e.g. A059056, refers to EIS sequence numbers.)

Distributions for some particular values for m and k are well studied. When $k = 1$, we have Montmort's [**14**] hat-matching problem with solution

$$p_j^{(1,m)} = \binom{m}{j} \frac{1}{j!} \sum_{i=0}^{n-j} \frac{(-1)^i}{i!}.$$

When $k = 2$, we have what Penrice [**17**] (also see Sequences A059056, A059057 in the EIS [**23**]) has described as the married version of the Christmas gift problem where adult children and their spouses draw names to exchange gifts for Christmas. Any drawing in which a person draws his own name or his spouse's is invalid. Penrice showed that this probability tends to e^{-2} as the number of married couples, $m \to \infty$. More generally, he showed that for fixed k,

$$\lim_{m\to\infty} p_0^{(k,m)} = e^{-k}.$$

This is a particular case of Knudsen and Skau [**11**] and David and Barton's [**5**] more general results. Earlier work on this problem that suggests these results was done by Riordan [**19**].

For general k, when $m = 2$, we can write an explicit formula for the matching probabilities. Consider the probability of zero matches. There are $(2k)!$ permutations of which $(k!)^2$ are permutations that do not result in a match, so

$$p_0^{(2,k)} = \frac{(k!)^2}{(2k)!}.$$

The formula for arbitrary i is

$$p_i^{(k,2)} = \begin{cases} \frac{\binom{k}{i/2}^2 (k!)^2}{(2k)!}, & k \text{ even} \\ 0 & k \text{ odd,} \end{cases}$$

since for there to be i matches, $k - i/2$ of the k guests who ordered chicken must have been switched with $k - i/2$ of the k guests who ordered prime rib and the k guests who are served prime rib and the k guests who are served chicken may be in any of $k!$ orders. The mode of the probability distribution is $i = \lfloor k/2 \rfloor$, and using Stirling's formula we can show

$$p_{\lfloor k/2 \rfloor}^{(2,k)} \sim \frac{2}{\sqrt{\pi k}}.$$

This distribution is a discrete approximation of a normal distribution (if one considers only even numbers of matches) with mean $\mu = k$ and variance $\sigma^2 = k^2/(2k-1)$. See the graph below, which compares the dinner-diner matching probabilities to the normal distribution with the same mean and variance, and to the Poisson distribution with the same mean. The numerators of these probabilities appear in [**23**] as Sequences A059064 and A059065.

For $m = 3$, the probability of no matches is given by

$$p_0^{(3,k)} = \frac{(k!)^3 \sum_{j=0}^{k} \binom{k}{j}^3}{(3k)!}.$$

If a random permutation has no matches, then for some $j = 0, \ldots, k$, j of the guests who ordered chicken will be served prime rib and $k - j$ will be served filet of sole. This means that $k - j$ of the guests who ordered filet of sole will be served prime rib and j will be served chicken and the remaining chicken and filet of sole orders will be served to those who ordered prime rib. There are $\binom{k}{j}^3$ ways of making such selections. Each of the three partitions may be in any of $k!$ orders. So there are $(k!)^3 \sum_{j=0}^{k} \binom{k}{j}^3$ possible permutations that correspond to no matches. The numbers $\sum_{j=0}^{k} \binom{k}{j}^3$ are called Franel numbers; see Cusick [**4**] and Sequence A000172 in the EIS [**23**]. Other sequences related to $m = 3$ are Sequence A059066 and A059067.

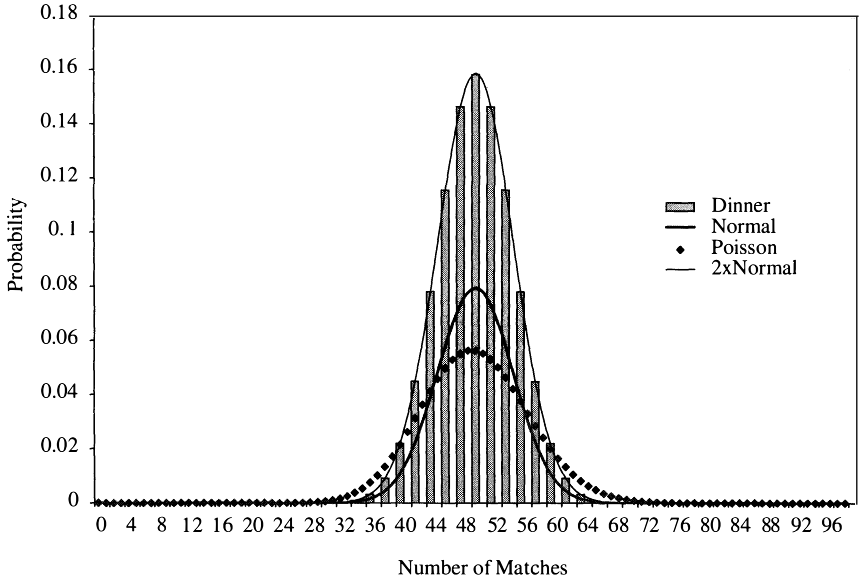


Figure 2 Dinner-diner matching probabilities, choice of 2 entrees

Sequences A059056-A059074, A008290, A000316, A000459, and A000166 are all related to these dinner-diner matching probabilities. More details on how they are computed are provided in the *On-line Encyclopedia of Integer Sequences* [**23**]. *Maple* code for generating the numerators of these probabilities is also provided there.

Conclusion: Matching probabilities for two standard decks of cards

We close by illustrating graphically that the normal approximation is superior to the Poisson if we let the number of diners increase while holding the number of choices of entrees constant, although the Poisson approximation is better if we hold the number of diners choosing each entree constant, but let the number of choices increase.

Instead of thinking of dinners and diners, consider a standard deck of playing cards. The graphs below show that, as expected, the Poisson distribution provides a better approximation for the rank-matching problem, but the normal distribution provides a better approximation for the suit-matching problem. For rank matching, the variance is about 94% of the mean, but for suit matching, it is only 76%. (The variance ap-

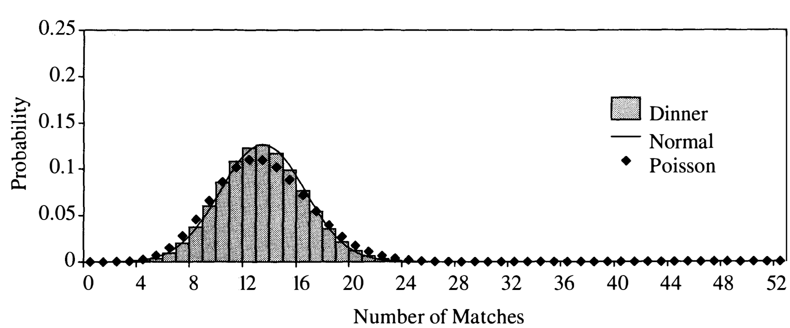


Figure 3 Standard deck matching suit

proximation using formula (6) from the Remark gives 92% and 75%, respectively, $(1 - 1/(mk))$ of the true variance.)

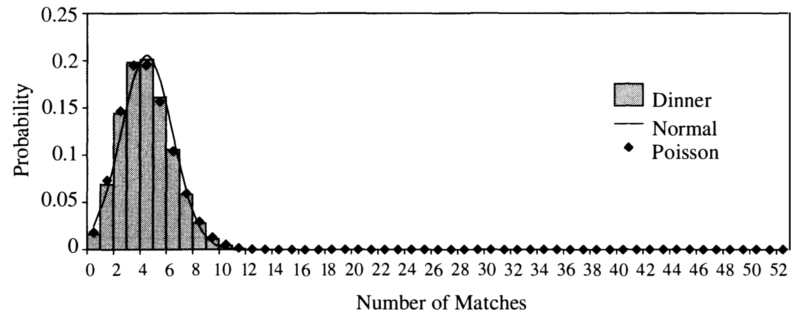


Figure 4 Standard deck matching rank

We leave the following as exercises for the interested reader (J. A. Greenwood [**7**]):

1. Show that the third central moment of the dinner-diner matching distribution is:

$$E\left([X(N)]^3\right) = \frac{m-2}{mk-2}\sqrt{\frac{mk-1}{m-1}},$$

for $N = km$ as in theorem 3 and $k, m > 1$. Hence,

$$\lim_{k\to\infty} E\left([X(N)]^3\right) = 0,$$

as is required for the third moment of a centered normal random variable.

2. Show that the fourth central moment of the dinner-diner matching distribution is (when $N = km$ as in theorem 3):

$$E\left([X(N)]^4\right) = \frac{3m^2(m-1)k^3 + (m^3 - 21m^2 + 21m)k^2 + 6(4m-3)k - m - 6}{m^2(m-1)k^3 - 5m(m-1)k^2 + 6(m-1)k},$$

and hence,

$$\lim_{k\to\infty} E\left([X(N)]^4\right) = 3,$$

as is required for a random variable that is asymptotically standard normal.

Acknowledgments. I am indebted to Felipe Martins and to three anonymous referees for additional references, comments, suggestions, and corrections. In particular, one of the referees suggested finding the moments of $X(N)$ using the binomial moments. This approach is more elegant and efficient than the generating function approach I had initially employed.

REFERENCES

1. T. W. Anderson, On card matching, *The Annals of Mathematical Statistics* **14** (1943), 426–435.
2. D. E. Barton, The matching distributions, *Journal of the Royal Statistical Society (B)* **20** (1958), 73–92.
3. R. J. Clarke, Derangements and Bell numbers, this MAGAZINE **66** (1993), 299–303.
4. T. W. Cusick, Recurrences for sums of powers of binomial coefficients, *J. Combin. Theory A* **52** (1989), 77–83.
5. F. N. David and D. E. Barton, *Combinatorial Chance*, Hafner, NY, 1962, ch. 7 and ch. 12.
6. W. Feller, *An introduction to probability theory and its applications*, second edition, John Wiley and Sons, New York, NY, 1971.

7. J. A. Greenwood, The first four moments of a general matching problem, *Annals of Eugenics* **10** (1940), 290–292.
8. T. N. E. Greville, The frequency distribution of a general matching problem, *The Annals of Mathematical Statistics* **12** (1941) 350–354.
9. A. Hald, *A History of Probability and Statistics and Their Applications Before 1750*, John Wiley and Sons, New York, NY, 1990.
10. K. Jordan, *Chapters on the Practical Calculus of Probability*, Akadémiai Kiadó, Budapest, 1972.
11. F. F. Knudsen and I. Skau, On the Asymptotic Solution of a Card-Matching Problem, this MAGAZINE **69** (1996), 190–197.
12. H. O. Lancaster, "The derivation and partition of χ^2 in certain discrete distributions," *Biometrika* **36**, (1949) 117–129.
13. H. O. Lancaster, *The Chi-squared Distribution*, John Wiley and Sons, New York, NY, 1969, Ch. XI.
14. P. R. de Montmort, *Essay d'Analyse sur les Jeux de Hazard*, Quillau, Paris, published anonymously, 1708.
15. P. R. de Montmort, *Essay d'Analyse sur les Jeux de Hazard*, Seconde Edition, Revûe et augmentée de plusiers Lettre, Quillau, Paris, 1713. Reprinted 1714. Published anonymously. Reprinted by Chelsea, New York, NY, 1980.
16. R. v. Mises, Das problem der iterationem, *Zeitschrift für angewandte Mathematik und Mechanik* **1** (1921) 298–307.
17. S. G. Penrice, Derangements, permanents and Christmas presents, *Amer. Math. Monthly* **98** (1991), 617–620.
18. S. Petit, *Encyclopedia of combinatorial structures*, published electronically at http://pauillac.inria.fr/algo/encyclopedia/, 2000.
19. J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley and Sons, Inc., New York, NY, 1958, p. 177.
20. S. N. Roy and S. K. Mitra, An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis, *Biometrika*, **43** (1956), 361–376.
21. F. S. Roberts, *Applied Combinatorics*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1984.
22. R. P. Stanley, *Enumerative Combinatorics Volume I*, Cambridge University Press, New York, NY, 1997.
23. N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, `http://www.research.att.com/~njas/sequences/`.
24. L. Takács, A moment convergence theorem, *Amer. Math. Monthly*, **98** (1991), 742–746.
25. L. Takács, *Archiv Hist. Exact Sci.*, **21** (1980), 229–244.
26. A. Tucker, *Applied Combinatorics*, 4th Ed., John Wiley & Sons, New York, NY, 2002.

Mathematics Awareness Month

The Joint Policy Board for Mathematics has selected Mathematics and Art as the theme for Mathematics Awareness Month 2003 (April). Essays, links, recommended books, and a list of speakers on the theme are available at `www.mathforum.org/mam/03`.

Activities for Mathematics Awareness Month are generally organized by college and university departments and student groups. They include a wide variety of workshops, competitions, exhibits, festivals, lectures, and symposia.

To advertise Mathematics Awareness Month the Joint Policy Board for Mathematics will send posters to most college and university mathematics departments. Individuals may order posters by visiting the Math Awareness web site. The sponsor for Mathematics Awareness Month for 2003 is the National Security Agency.

NOTES

A Natural Generalization of the Win-Loss Rating System

CHARLES REDMOND
Mercyhurst College
Erie, PA 16546
chad@mercyhurst.edu

How are teams in a tournament usually ranked? For most of the popular sports in the U.S., the percentage of wins is all-important, and the team with the highest winning percentage earns first place. Although this system is here to stay (and for many reasons, such as its simplicity, this is desirable), all sports fans recognize a nagging flaw: in a tournament that is not round-robin, so that each team plays only a subset of the others, a team with a weak schedule may have a considerable advantage over one facing strong opponents.

The sports world offers many possible remedies. Any reader of the sports section of *USA Today* is familiar with Jeff Sagarin. Sagarin's ratings, applied to numerous professional and collegiate sports, have enlightened fans since the mid-1980s and have officially guided both the NCAA basketball tournament selection committee and the college football Bowl Championship Series commission. It is clear that strength of schedule undoubtedly plays a role in this ranking scheme. For example, in his final ranking of NFL teams for the 2001 season, Sagarin [**8**] places Kansas City at 6-10 above Washington at 8-8. According to Sagarin, Kansas City faced the 16th most difficult schedule, whereas Washington opposed the 28th. Other ranking systems abound, including those by Richard Billingsley and Kenneth Massey, both employed by the Bowl Championship Series commission, and the time-honored Dunkel Index, which has been around since the 1920s.

The mathematical community has also tackled the problem, leaving a trail of research going back as far as Zermelo [**14**]. Driven by mathematical interest, as opposed to applicability, mathematicians tend to focus only on team performance and avoid building into their models factors such as home-field advantage, game location, recent team performance, and so on. The simplicity of this approach is preferred by many mathematicians, rather than the more complicated models used by sports professionals. Well-known authors in this genre are Keener [**4**] and Minton [**5**]. Keener uses his results to cast doubt on Brigham Young University's 1984 national football title, and Minton argues that Colorado should have stood alone in 1990, the year the Buffaloes shared the championship with Georgia Tech. For an introduction to this area of research, the reader should consult these articles as well as Stob [**10**], which is an excellent survey of some previous advances in the area.

In the spirit of Barbeau [**1**], Keener [**4**], and Saaty [**7**], at the heart of whose ranking schemes is a limiting process, and Minton [**5**], who stresses point spread over point ratio, I wish to share with you my own system. It may not be applicable to the world of sports, but it does seem simple and natural. In my opinion, the scheme and its underlying theory are among the easiest to grasp in the literature. I hope therefore to provide a window into this mathematically intriguing subject for undergraduate students, especially those of linear algebra, as well as for anyone curious about by these questions. I am deeply indebted to the works of the authors mentioned above, especially Minton, with whom, although my approach is different, my rankings agree.

An illustrative example For the sake of clarity and to avoid tedious notation, I will limit my discussion to a simple example and leave the general formulation to the interested reader. Consider the following tournament of four teams in which each team has played two games.

Opponents	**Result**
A vs. B	5-10
A vs. D	57-45
B vs. C	10-7
C vs. D	3-10

B is 2-0, A is 1-1, D is 1-1, and C is 0-2. In other words, B's winning percentage is 1.00, A's and D's are each 0.50, and C's is 0.00. The traditional win/loss method of ranking places B in first place followed by A and D in second place and C in last place. Note that in the calculation of the winning percentages it is as if each team has been given 1 point for each win and 0 points for each loss. We have, however, ignored the possibility of a tie. The first revision, then, that we will suggest is that teams be given a score of 1 for each win, a score of 0 for each tie, and a score of -1 for each loss. Each team's rating would then be determined by the sum of its scores divided by the number of games played (2 in the case of our example). B's rating is still 1.00, but A's and D's are each now 0.00, and C's is -1.00.

With this amendment in place, we can now define the *dominance* of one team over another. Because B defeated A, we will say that B's dominance over A is 1. Conversely, A's dominance over B will be said to be -1. The average dominance of A over its opponents is its rating, 0.00, and so forth. There is still a flaw in this approach, however, since there is now an artificial limitation on one team's dominance over another. In our example, B defeats A by 5 points, but defeats C by only 3 points. Thus B's dominance of 1 over each team reflects imperfectly what has really happened. We will therefore make another change to our proposed ranking scheme by redefining one team's dominance over another to be its score in the game played minus its opponents score. B's dominance over A is then 5, whereas A's dominance over B is -5. We list each team's average dominance over the field of its competitors.

Team	**Average Dominance**
A	3.5
B	4
C	-5
D	-2.5

These new ratings are perhaps more descriptive, but they still don't factor in strength of schedule. Before tackling this problem, we need to make yet one more minor, but important, modification of our proposed ranking scheme. We will consider each team as having played one game against itself. Each team will receive a score of zero, of course, for this game. Though this requirement seems strange at first, the reason for it will be made apparent. Under this latest revision, our modified initial standings are listed in the table below.

Team	**Rating**
A	$7/3 = 2.33$
B	$8/3 = 2.67$
C	$-10/3 = -3.33$
D	$-5/3 = -1.67$

The fundamental idea Consider team A. It does not play C in the tournament, so we do not have any direct way of comparing the two teams. However, A plays D, and D plays C, so there is a path from A to C. A defeats D by 12 points, and D defeats C by 7 points. The fundamental idea is to consider an imaginary game to have been played between A and C where A defeats C by $12 + 7 = 19$ points. C could be called a *second-generation opponent* of A. Let us enumerate all second-generation games "played" by A by constructing a tree where each first-generation opponent of A emanates from A, and each second-generation opponent of A emanates from the first-generation opponent it plays. The appropriate dominance is assigned to each edge.

We are now considering A to have played nine games instead of three, and the scores of the games are obtained by adding down each branch of the tree as in FIGURE 1. The second-generation dominance of A is then determined by the average of these scores. We list the average second-generation dominances of all our teams.

Team	**Average Second Generation Dominance**
A	3.44
B	3.22
C	-4.11
D	-2.56

Note that A has moved into first place. Strength of schedule now plays a role in ranking teams since the ranking also depends on the performance of their first-generation opponents. Though a team might not have a stellar first-generational record, teams with difficult schedules will reap the benefits of their opponents' success. One interesting point is that some of these nine games are identical to an original real game. For example, the tree in FIGURE 1 lists the second-generation game $A \xrightarrow{0} A \xrightarrow{-5} B$, which is identical to the first-generation game $A \xrightarrow{-5} B$. This is one of the reasons that we require a team to be an opponent of itself. The original real comparisons are not lost.

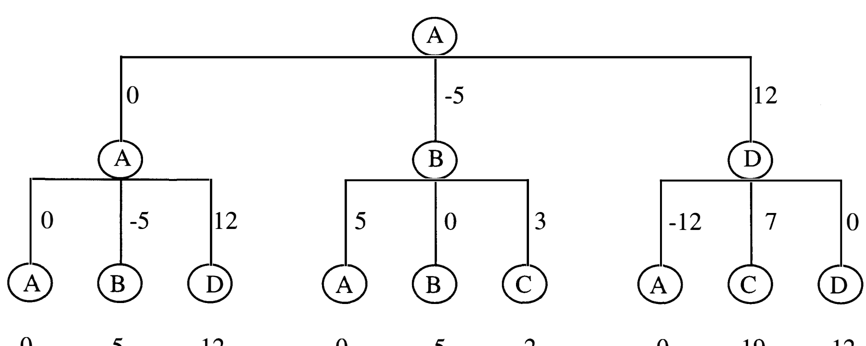


Figure 1 The average dominance of A over its second-generation opponents = 31/9 = 3.44

This process can be continued through any number of generations, and our intuition suggests that as the number of generations increases, our ranking becomes more accurate. Let $r_n(A)$ denote the average n-generation dominance of A or, equivalently, A's

n-generation rating. A natural and important question to ask at this stage is whether or not $\lim_{n\to\infty} r_n(A)$ exists. This limit, if it exists, is the ultimate rating we intend to assign to A. Here is where the mathematics becomes interesting and, I feel, surprising.

The mathematical formulation We illustrate the structure of our tournament with the graph in FIGURE 2 (note the edges from each team to itself).

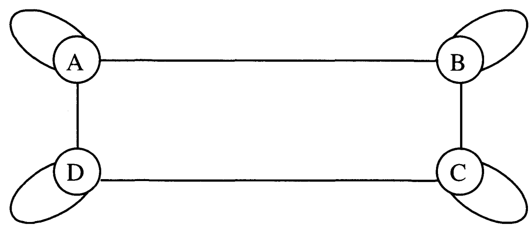


Figure 2

The incidence matrix for this graph is a 4 x 4 matrix with a row and column for each node. Each entry is a 0 or a 1, depending on whether or not the node for the entry's row is connected to the node for the entry's column. Our incidence matrix (where A corresponds to the first row and column, B to the second, etc.) is

$$M = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

Now consider the square of the incidence matrix, M^2. It turns out that the (i, j) entry of M^2 is the number of distinct paths of length 2 between the node corresponding to row i and the node corresponding to column j. (The reader should think this through.)

$$M^2 = \begin{bmatrix} 3 & 2 & 2 & 2 \\ 2 & 3 & 2 & 2 \\ 2 & 2 & 3 & 2 \\ 2 & 2 & 2 & 3 \end{bmatrix}$$

Compare the first row of M^2 to FIGURE 1. The entries in the first row of M^2 indicate that A appears as a second-generation opponent of itself three times, whereas the other teams appear twice. This is exactly what FIGURE 1 shows. Other powers of M behave the same way: the (i, j) entry of M^n is the number of times the team corresponding to column j appears as an n-generation opponent of the team corresponding to row i.

Now define the vector

$$S = \begin{bmatrix} 7 \\ 8 \\ -10 \\ -5 \end{bmatrix}.$$

The coordinates of S are the net points scored by the teams in the tournament (for instance, the net number of points A scored is 7). Then the coordinates of the vector

$$\frac{1}{3}M^0 \cdot S = \frac{1}{3}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 7 \\ 8 \\ -10 \\ -5 \end{bmatrix} = \begin{bmatrix} 2.33 \\ 2.67 \\ -3.33 \\ -1.67 \end{bmatrix}$$

are the first-generation ratings of our teams.

Using FIGURE 1, we can compute the second-generation ratings also. Observe that to calculate the second-generation rating of A, the second-generation dominances have to be added in once, whereas the first-generation dominances each have to be added in three times. The first coordinate of $3M^0 \cdot S$ will give us the latter sum, and the first coordinate of $M^1 \cdot S$ will give us the former. (The reader should pause to think this through.) Thus, the coordinates of the vector

$$\frac{1}{3^2}(3M^0 \cdot S + M^1 \cdot S) = \frac{1}{3}M^0 \cdot S + \frac{1}{3^2}M^1 \cdot S$$

yield the second-generation ratings for our teams. Likewise, the vector yielding the n-generation ratings has the formula

$$\sum_{j=1}^{n} \frac{1}{3}\left(\frac{M}{3}\right)^{j-1} \cdot S.$$

Does the limiting vector

$$\lim_{n\to\infty}\left(\sum_{j=1}^{n} \frac{1}{3}\left(\frac{M}{3}\right)^{j-1} \cdot S\right)$$

necessarily exist?

The limit To evaluate this limit (and show that it exists), we use an eigenvector decomposition of $M/3$. The eigenvalues of $M/3$ are 1, $-1/3$, and $1/3$ occurring with multiplicity 2. Because M is symmetric, we may choose a set of four linearly independent orthonormal eigenvectors corresponding to these eigenvalues. We may choose eigenvectors

$$\vec{v}_0 = \begin{bmatrix} 1/2, \\ 1/2, \\ 1/2, \\ 1/2 \end{bmatrix}, \vec{v}_1 = \begin{bmatrix} -1/\sqrt{2}, \\ 0, \\ 1/\sqrt{2}, \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 0, \\ -1/\sqrt{2}, \\ 0, \\ 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \vec{v}_3 = \begin{bmatrix} -1/2, \\ 1/2, \\ -1/2, \\ 1/2 \end{bmatrix}$$

corresponding to eigenvalues 1, $1/3$, $1/3$, and $-1/3$, respectively.

Since the eigenvectors form a basis, we may now express S as a linear combination of these eigenvectors. The coefficient of each eigenvector may be obtained by computing its scalar product with S:

$$S = \begin{bmatrix} 7 \\ 8 \\ -10 \\ -5 \end{bmatrix} = (S \cdot \vec{v}_0)\vec{v}_0 + (S \cdot \vec{v}_1)\vec{v}_1 + (S \cdot \vec{v}_2)\vec{v}_2 + (S \cdot \vec{v}_3)\vec{v}_3$$

$$= 0\begin{bmatrix} 1/2, \\ 1/2, \\ 1/2, \\ 1/2 \end{bmatrix} - \frac{17}{\sqrt{2}}\begin{bmatrix} -1/\sqrt{2}, \\ 0, \\ 1/\sqrt{2}, \\ 0 \end{bmatrix} - \frac{13}{\sqrt{2}}\begin{bmatrix} 0, \\ -1/\sqrt{2}, \\ 0, \\ 1/\sqrt{2} \end{bmatrix} + 3\begin{bmatrix} -1/2, \\ 1/2, \\ -1/2, \\ 1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 17/2 \\ 0 \\ -17/2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 13/2 \\ 0 \\ -13/2 \end{bmatrix} + \begin{bmatrix} -3/2 \\ 3/2 \\ -3/2 \\ 3/2 \end{bmatrix}.$$

Note that the three vectors above are still eigenvectors of $M/3$ with eigenvalues $1/3$, $1/3$, and $-1/3$, respectively. Let us call these new eigenvectors $\vec{s_1}$, $\vec{s_2}$, and $\vec{s_3}$. In fact, for any positive integer j they are also eigenvectors of $(M/3)^j$ with associated eigenvalues $(1/3)^j$, $(1/3)^j$ and $(-1/3)^j$, respectively. Thus we can write

$$\begin{aligned}
\lim_{n\to\infty}\sum_{j=1}^{n}\frac{1}{3}\left(\frac{M}{3}\right)^{j-1}\cdot S &= \frac{1}{3}\lim_{n\to\infty}\sum_{j=0}^{n-1}\left(\frac{M}{3}\right)^{j}\cdot(\vec{s_1}+\vec{s_2}+\vec{s_3}) \\
&= \frac{1}{3}\lim_{n\to\infty}\sum_{j=0}^{n-1}\left[\left(\frac{1}{3}\right)^{j}\vec{s_1}+\left(\frac{1}{3}\right)^{j}\vec{s_2}+\left(-\frac{1}{3}\right)^{j}\vec{s_3}\right] \\
&= \frac{1}{3}\sum_{j=0}^{\infty}\left(\frac{1}{3}\right)^{j}\vec{s_1}+\frac{1}{3}\sum_{j=0}^{\infty}\left(\frac{1}{3}\right)^{j}\vec{s_3}+\frac{1}{3}\sum_{j=0}^{\infty}\left(-\frac{1}{3}\right)^{j}\vec{s_3} \\
&= \frac{1}{3}\left(\frac{1}{1-1/3}\right)\vec{s_1}+\frac{1}{3}\left(\frac{1}{1-1/3}\right)\vec{s_2}+\frac{1}{3}\left(\frac{1}{1+1/3}\right)\vec{s_3} \\
&= \begin{bmatrix} 3.875 \\ 3.625 \\ -4.625 \\ -2.875 \end{bmatrix}.
\end{aligned}$$

The keys The reader can now easily see that the keys to establishing the limit are:

1. $M/3$ has an eigenvalue of 1 with a corresponding eigenvector having identical coordinates; this is because the rows of $M/3$ add to 1. In other words, the sum of each row in the incidence matrix is the number of games played by each team.
2. The coordinates of S add to zero, which is necessarily true of any vector obtained by our rankings, since the coordinates are the net number of points scored by the teams in a tournament. This fact and the key mentioned above guarantee that the coefficient of the eigenvector with eigenvalue 1 in the decomposition of S will be zero.
3. The absolute value of each of the other eigenvalues is strictly less than 1. This guarantees that the infinite series above converge. Why is this true of the other eigenvalues? This follows because $M/3$ is a *Markov matrix*: a matrix with non-negative entries and each column adding to 1. The eigenvalues of Markov matrices are very special. One eigenvalue is equal to 1, and the absolute value of each of the other eigenvalues is ≤ 1 . This inequality becomes strict if any power of the Markov matrix has all positive entries (see Strang **[11]** for more on Markov matrices). Because our tournament graph is connected (it wouldn't really make sense to consider a tournament for which this is not the case), given any team, every other team must appear as its opponent in some generation. Once a team appears as an opponent in some generation, it will appear in every generation after that because it is an opponent of itself. Consequently, some power of $M/3$ will have all positive entries. This is the primary reason we make this requirement at the outset.

Concluding remarks When I first worked on this problem, the notion of the teams opposing themselves had not occurred to me. When I used *Mathematica* to run the ranking algorithm on several tournaments, I did not always notice limiting behavior. I understand the mathematical necessity now for having teams oppose themselves, but a further explanation still eludes me.

The rankings produced by the scheme described in this note are identical to those produced by Minton's method [**5**], which requires the solution of a system of linear equations; in ours we calculate eigenvalues and eigenvectors. Both tasks could become difficult if the tournament is large. Ours, however, has the advantage of offering approximate rankings. One has just to run the algorithm out to a specified generation. This is not as accurate as calculating the limits, but the results would be perhaps fairer than those given by the win/loss ranking system.

What if the tournament is unbalanced in the sense that the teams do not all play the same number of games? By allowing the convention that teams may play any number of games against themselves, the number of games played can be made equal. For instance, suppose the tournament is just two games, A against B, and A against C. In this case, A has three opponents (B, C, and itself), whereas each of the others has two. To even matters, B and C can be required to play two games against themselves instead of one.

Some readers may object to the diminished significance of a win and the heightened importance of the score in our scheme. I admit that winning is part of the excitement of a game. Allotting a certain number of points to the victor just for prevailing, however, may preserve the value of a win. If many points are awarded for a victory, winning will drive the ranking of the teams.

Every point scored, though, is potentially important in the end. This is why teams, even with victory assured, must press on in each game to score as many points as possible. One may object to this feature of our system since it appears to advocate the humiliation of the loser. When only the win counts, which is the case in most tournaments today, a team losing by a large margin is embarrassed, because those extra points do not count. In our system, every point gained or lost could potentially make or break a team in the end. Even if the winner is decided early on, both teams must play as hard as possible for the entire game. A team might just qualify for the playoffs because of a heroic goal-line stand preventing a touchdown in a game it lost by 50 points. I see this as a positive aspect in that a game is never over or meaningless before its conclusion.

Finally, I would like to comment on the additive nature of the dominances we use. Since, in our example, A defeats D by a score of 57-45, we say that A is 12 points better than D. This is the additive approach, used also by Minton. A is then considered to be $(57 - 45) + (10 - 3) = 19$ points better than C, and so on. Other authors such as Barbeau [**1**], Keener [**4**], and Saaty [**7**] use a multiplicative approach; in such a ranking system, A is considered to be $57/45 = 1.27$ times better than D. A would then be $(57/45) \cdot (10/3) = 4.22$ times better than C. For contests in which each team plays both offense and defense, I much prefer the additive approach. In our tournament, although the scores of the respective games appear to be very different, the additive approach evaluates the strength of A against D (12 points) similarly to the way it evaluates the strength of D against C (7 points). To me this is reasonable. The game between A and D may be high scoring because they both have powerful offenses and weak defenses. On the other hand, the game between D and C may be low scoring because C is strong defensively and weak offensively. It seems unfair to allow the actual number of points scored to play a major role in comparing two teams. In the multiplicative approach, A is rated as being just 1.27 times better than D, but D is rated as being significantly better than C (3.33 times).

For teachers of introductory linear algebra and their students, I think what we've described is an interesting problem that highlights some of the topics encountered towards the end of the course. I hope that you find the time spent on it as rewarding as I have and that it enriches your enthusiasm for the subject.

Acknowledgments. The author thanks the referees for all their helpful suggestions. The author also thanks Don Platte for his valuable comments and Jessica Dillabough for her technical assistance.

REFERENCES

1. E. Barbeau, Perron's result and a decision on admission tests, this MAGAZINE **59**: 1 (1986), 12–22.
2. D. Harville, The use of linear-model methodology to rate high school or college football teams, *J. Amer. Statist. Assoc.* **72** (1977), 278.
3. T. A. Jech, The ranking of incomplete tournaments: a mathematician's guide to popular sports, *Amer. Math. Monthly* **90** (1983), 246–266.
4. J. P. Keener, The Perron-Frobenius theorem and the ranking of football teams, *SIAM Review* **35**:1 (1993), 80–93.
5. R. Minton, A mathematical rating system, *UMAP Journal* **13**:4 (1992), 313–334.
6. I. Peterson, Who's Really No. 1?, (1998). Available at `http://www.maa.org/mathland/mathtrek/_12_14_98.html`.
7. T. L. Saaty, Rank according to Perron: a new insight, this MAGAZINE **60** (1987), 211–213.
8. J. Sagarin, Jeff Sagarin NFL ratings, (2002). Available at http://www.usatoday.com/sports/sagarin/nfl01.htm.
9. J. H. Smith, Adjusting baseball standings for strength of teams played, *Amer. Statis.* **10** (1956), 23–24.
10. M. Stob, A supplement to a mathematician's guide to popular sports, *Amer. Math. Monthly* **91** (1984), 277–281.
11. G. Strang, *Linear Algebra and Its Applications*, 3rd edition, Harcourt Brace Jovanovich, San Diego, CA, 1988.
12. M. Thompson, On any given Sunday: fair competitor orderings with maximum likelihood methods, *J. Amer. Statist. Assoc.* **70** (1975), 536–541.
13. A. Tucker, *Linear Algebra: an introduction to the theory and use of vectors and matrices*, MacMillan, New York, NY, 1993.
14. E. Zermelo, Die berechnung der turnier-ergebnisse als ein maximumproblem der warscheinlichkeitsrechnung, *Math. Z.* **29** (1926), 436–460.

Volume, Surface Area, and the Harmonic Mean

PAUL FJELSTAD
32991 Dresden Ave.
Northfield, MN 55057
fjelstap@stolaf.edu

IVAN GINCHEV
Department of Mathematics
Technical University of Varna
9010 Varna, Bulgaria
ginchev@ms3.tu-varna.acad.bg

The volume-area-derivative relationship $dV/dr = A$ for a ball of radius r is a rather striking one when one first comes across it. Accordingly, it is natural to consider generalizations, as Emert and Nelson did in a Note in this MAGAZINE [**1**], first for n-dimensional spheres, and then for n-dimensional polyhedra that circumscribe n-dimensional spheres. When the circumscription fails, they suggest saving the relationship by pushing out the faces of the polyhedron a distance of ϵ, while noting that this does not produce a family of similar objects. We introduce an alternative parameter, called h because the harmonic mean plays a role. This parameter indexes a family of similar objects and differentiating with respect to h rescues the volume-area-derivative relationship, which, in the vernacular, might whimsically be rephrased as "deriving something superficial from something that has depth."

Review To understand the essentials of the relationship, consider a ball of radius r. If one extends r to $r + dr$, say, by painting the ball with a thickness dr, then, neglecting the higher order powers of dr, the increase in volume will be the area A of the ball times dr: $dV = A\,dr$. Expressing the volume V as a function of r gives the celebrated formula $dV/dr = A$.

It is certainly possible to express the volume as a function of other parameters. If instead one expresses V as a function of, say, the half-radius $r' = r/2$, then $dr = 2dr'$ and $dV/dr' = 2A$. That is, extending the radius by dr corresponds to painting 2 coats of thickness dr' (FIGURE 1, left). Similarly, expressing V as a function of the double-radius $r'' = 2r$, means $dV/dr'' = A/2$. Because the possible parameters form a continuum, one can derive any multiple of A one likes.

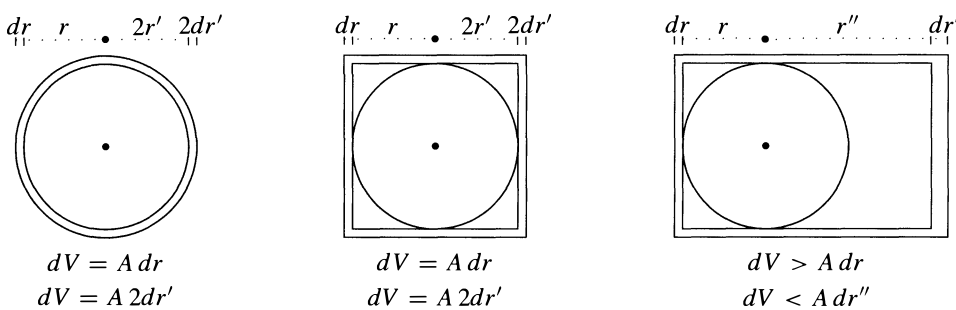


Figure 1 Here $r = 2r' = \frac{1}{2}r''$, so $dr = 2dr' = \frac{1}{2}dr''$

If one can inscribe a sphere of radius r in a polyhedron, the same conclusions result for the volume and area of the polyhedron, since extending r to $r + dr$ means adding a thickness dr to each face of the polyhedron (FIGURE 1, middle).

If one can't inscribe a sphere of radius r in a polyhedron, meaning at least one face is farther than r from the sphere's center, say $r'' > r$, then as r is increased to $r + dr$, one adds a thickness dr'' greater than dr to this face in order to get a polyhedron similar to the original one and $dV/dr > A$, while $dV/dr'' < A$ (FIGURE 1, right).

These results for polyhedra, generalized to n-dimensional polyhedra, also called polytopes (with V always the n-dimensional content and A the $(n-1)$-dimensional content), were presented in a more formal manner by authors Emert and Nelson in this MAGAZINE, after which they said the following [**1**, p. 369]:

> In general it is not possible to pick an "inner radius r" so that $dV/dr = A$. But is there a different variable that can replace r to achieve this "volume-area" relationship?

Their answer to this question is to pick an interior point from which to push out each face of the polytope a distance ϵ, so that V and A, as functions of ϵ, result in $dV/d\epsilon = A$. The members of the family parameterized by ϵ, however, are not similar, and different interior points result in different families. Further, ϵ is considered as being small, as opposed to r, which can be any positive number. These characteristics of ϵ can be avoided.

An alternative There is a parameter that produces similar family members and which seems to be the natural one to replace r when the polytope fails to circumscribe a sphere. As an instructive example, consider a family of similar 3-dimensional solid boxes parameterized by s, and let the distances from the centerpoint to the six faces be as, as, bs, bs, cs, cs. Then

$$V = 2as\,2bs\,2cs = 8abcs^3, \text{ and}$$
$$A = 2(2as\,2bs + 2as\,2cs + 2bs\,2cs) = 8(ab+ac+bc)s^2.$$

To determine h, it must be some multiple of s, so let $h = ks$. Then

$$V = 8abc\left(\frac{h}{k}\right)^3, \quad \frac{dV}{dh} = \frac{24abc}{k^3}h^2, \quad A = \frac{8(ab+bc+ac)}{k^2}h^2.$$

Setting $dV/dh = A$, we find that $k = 3abc/(ab+ac+bc)$, so

$$h = \frac{3abc}{ab+ac+bc}s = \frac{6}{\dfrac{1}{as}+\dfrac{1}{as}+\dfrac{1}{bs}+\dfrac{1}{bs}+\dfrac{1}{cs}+\dfrac{1}{cs}}.$$

Thus h is the harmonic mean of the centerpoint-to-face distances, and this result extends to the general case of n-dimensional boxes. This suggests that h, besides standing for *height*, might also stand for *harmonic*.

To be more general (working now in dimension n), let A_i be the area of the polytope's ith face so that $\sum A_i = A$, the total area. Also, let h_i be the distance from an interior point of the polytope to this face so $V_i = h_iA_i/n$ is the volume of the pyramid built on the ith face and $\sum V_i = V$, the total volume. For h_s and h_g the shortest and greatest of the h_i, with $h_s < h_g$, it follows that $dV/dh_s > A$, as considered above and depicted in FIGURE 1. Using the same kind of reasoning, it follows that $dV/dh_g < A$. There is accordingly some mean distance h between h_s and h_g such that $dV/dh = A$.

More generally, for a family of similar polytopes, the volumes will be proportional to h^n, so

$$V = Ch^n, \quad A = \frac{dV}{dh} = Cnh^{n-1} = \frac{nV}{h}, \quad \text{from which} \quad h = \frac{nV}{A} = \frac{\sum h_iA_i}{\sum A_i}.$$

Thus the mean distance h is the weighted mean of the h_i with the areas A_i being the weights. A generalization of the box example occurs when an interior point exists such that all V_i are equal. Then $V_i = V/m$ for a polytope with m faces and h is a harmonic mean:

$$h = \frac{nV}{A} = \frac{m}{\sum A_i/(nV/m)} = \frac{m}{\sum A_i/(nV_i)} = \frac{m}{\sum A_i/(h_iA_i)} = \frac{m}{\sum 1/h_i}.$$

The term nV/A shows, however, that h is independent of which interior point was used.

In fact, the result $h = nV/A$ suddenly opens up the possibility for a sweeping generalization, for it applies not only to families of similar polytopes, but to any family of similar objects which have some n-dimensional content V and some $(n-1)$-dimensional content A. Phrased in this way, the problem of how content is determined is sidestepped. That different methods can give different results (see Gelbaum and Olmsted [**2**, p. 150]; Schneider [**3**] treats more esoteric aspects of the subject) is not a problem, since whatever methods provide values for V and A as functions of some length parameter s, these values will be proportional to s^n and s^{n-1}

because of similarity. The parameter h for a given object can be thought of as the radius of a sphere that has the same ratio of V to A as the object. It will henceforth be referred to as the object's *harmonic parameter.*

To check how $h = nV/A$ works on some examples, consider the three simple cases of FIGURE 1, plus that of a right circular cone.

Circles: radius r, $h = 2(\pi r^2)/(2\pi r) = r$.

Squares: side $2r$, $h = 2(2r)^2/4(2r) = r$.

Rectangles: length $3r$, width $2r$, $h = 2(3r2r)/2(3r + 2r) = 6r/5$.

Cones: radius r, height ar,

$$h = 3\left(\frac{1}{3}\pi r^2 ar\right) \Big/ (\pi r^2 + \pi r\sqrt{r^2 + (ar)^2}) = ar/(1 + \sqrt{1 + a^2}).$$

If you can assign a volume and an area to your coffee cup, it will have a value for its harmonic parameter h. The magnitude of h is somewhat indicative of function. For a lung and a balloon of the same volume, the lung will have a much smaller h than the balloon, and the same holds in comparing a brain and a stomach. Returning to the vernacular, one might say that increasing superficiality provides for higher order functioning.

As might be expected, the generalization invoked by the idea of the harmonic parameter can be extended to higher derivatives and also to multi-parameter objects. It must be confessed that the authors found it quite entertaining to do so, for the harmonic mean kept showing up in unexpected ways, which they believe readers might now enjoy discovering by themselves.

REFERENCES

1. John Emert and Roger Nelson, Volume and surface area for polyhedra and polytopes, this MAGAZINE **70** (1997), 365-371.
2. B. R. Gelbaum and J. M. Olmsted, *Counterexamples in Analysis*, Holden-Day, San Francisco, CA, 1964.
3. R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory*, Cambridge University Press, Cambridge, 1993.

Markov Chains for the *RISK* Board Game Revisited

JASON A. OSBORNE
North Carolina State University
Raleigh, NC 27695
osborne@stat.ncsu.edu

Probabilistic reasoning goes a long way in many popular board games. Abbott and Richey [**1**] and Ash and Bishop [**2**] identify the most profitable properties in *Monopoly*, and Tan [**3**] derives battle strategies for *RISK*. In *RISK*, the stochastic progress of a battle between two players over any of the 42 countries can be described using a Markov chain. Theory of Markov chains can be applied to address questions about the probabilities of victory and expected losses in battle.

Tan addresses two interesting questions:

> *If you attack a territory with your armies, what is the probability that you will capture this territory? If you engage in a war, how many armies should you expect to lose depending on the number of armies your opponent has on that territory?* [**3**, p. 349]

A mistaken assumption of independence leads to a slight misspecification of the transition probability matrix for the system, which leads to incorrect answers to these questions. Correct specification is accomplished here using enumerative techniques. The answers to the questions are updated and recommended strategies are revised and expanded. Results and findings are presented along with those from Tan's article for comparison.

TABLE 1: An example of a battle

Roll #	No. of armies		No. of dice rolled		Outcome of the dice		No. of losses	
	attacker	defender	attacker	defender	attacker	defender	attacker	defender
1	4	3	3	2	5, 4, 3	6, 3	1	1
2	3	2	3	2	5, 5, 3	5, 5	2	0
3	1	2	1	2	6	4, 3	0	1
4	1	1	1	1	5	6	1	0
5	0	1						

The Markov chain The object for a player in *RISK* is to conquer the world by occupying all 42 countries, thereby destroying all armies of the opponents. The rules of *RISK* are straightforward and many readers may need no review. Newcomers are referred to Tan's article where a clear and concise presentation can be found. Tan's Table 1 is reproduced here for convenience. It shows the progress of a typical battle over a country, with the defender prevailing after five rolls. This table also serves as a reminder of the number of dice rolled in various situations—never more than three for the attacker, and never more than two for the defender.

Following Tan's notation, let A denote the number of attacking armies and D the number of defending armies at the beginning of a battle. The state of the battle at any time can be characterized by the number of attacking and defending armies remaining. Let $X_n = (a_n, d_n)$ be the state of the battle after the nth roll of the dice, where a_n and d_n denote the number of attacking and defending armies remaining. The initial state is $X_0 = (A, D)$. The probability that the system goes from one state at turn n to another state at turn $n+1$, given the history before turn n, depends only on (a_n, d_n), so that $\{X_n : n = 0, 1, 2, \ldots\}$ forms a Markov chain:

$$\Pr\left[X_{n+1} = (a_{n+1}, d_{n+1}) \mid x_n, x_{n-1}, \ldots, x_1, x_0\right] = \Pr\left[X_{n+1} = (a_{n+1}, d_{n+1}) \mid x_n\right]$$

The AD states where both a and d are positive are transient. The $D + A$ states where either $a = 0$ or $d = 0$ are absorbing. Let the $AD + (D + A)$ possible states be ordered so that the AD transient states are followed by the $D + A$ absorbing states. Let the transient states be ordered

$$\{(1, 1), (1, 2), \ldots, (1, D), (2, 1), (2, 2), \ldots, (2, D), \ldots, (A, D)\}$$

and the absorbing states

$$\{(0, 1), (0, 2), \ldots, (0, D), (1, 0), (2, 0), \ldots, (A, 0)\}.$$

Under this ordering, the transition probability matrix takes the simple form

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix},$$

where the $(AD) \times (AD)$ matrix Q contains the probabilities of going from one transient state to another, and the $(AD) \times (D + A)$ matrix R contains the probabilities of going from a transient state into an absorbing state.

The transition probability matrix, *P* It turns out that P contains only 14 distinct probabilities, having to do with how many dice are rolled and how many armies lost as a result. Let π_{ijk} denote the probability that the defender loses k armies when rolling j dice against an attacker rolling i dice, as given in Table 2. To obtain the π_{ijk}, the joint probabilities associated with the best and second best roll from 2 or 3 six-sided dice need to be quantified. Let Y_1, Y_2, Y_3 denote the unordered outcome for an attacker when rolling three dice and let W_1, W_2 denote the unordered outcome for an attacker when rolling two dice. Let Z_1, Z_2 denote the unordered outcome for a defender rolling two dice. Then Y_1, Y_2, Y_3 and W_1, W_2 and Z_1, Z_2 are random samples from the discrete uniform distribution on the integers 1 through 6:

$$\Pr(Y_j = y) = \begin{cases} \frac{1}{6} & \text{for } y = 1, 2, 3, 4, 5, 6 \\ 0 & \text{else.} \end{cases}$$

When order is taken into account and denoted using superscripts, as in $Y^{(1)} \geq Y^{(2)} \geq Y^{(3)}$, the ordered random variables are called *order statistics.* The joint distributions of order statistics needed for specification of π_{ijk} can be obtained using straightforward techniques of enumeration. When two dice are rolled, the joint distribution of $(Z^{(1)}, Z^{(2)})$ is

$$\Pr(Z^{(1)} = z^{(1)}, Z^{(2)} = z^{(2)}) = \begin{cases} \frac{1}{36} & \text{for } z^{(1)} = z^{(2)} \\ \frac{2}{36} & \text{for } z^{(1)} > z^{(2)} \\ 0 & \text{else,} \end{cases}$$

and the *marginal* distribution of the best roll $Z^{(1)}$ is obtained by summing the joint distribution over values of $z^{(2)}$:

$$\Pr(Z^{(1)} = z^{(1)}) = \left\{ \frac{2z^{(1)} - 1}{36} \quad \text{for } z^{(1)} = 1, 2, 3, 4, 5, 6. \right.$$

When three dice are rolled, the pertinent distribution of the best two rolls is

$$\Pr(Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}) = \begin{cases} \frac{3y^{(1)}-2}{216} & \text{for } y^{(1)} = y^{(2)} \\ \frac{6y^{(2)}-3}{216} & \text{for } y^{(1)} > y^{(2)} \\ 0 & \text{else,} \end{cases}$$

and the marginal distribution of the best roll is

$$\Pr(Y^{(1)} = y^{(1)}) = \left\{ \frac{1 - 3y^{(1)} + 3(y^{(1)})^2}{216} \quad \text{for } y^{(1)} = 1, 2, 3, 4, 5, 6. \right.$$

All of the probabilities are 0 for arguments that are not positive integers less than or equal to 6. The joint distribution of $W^{(1)}$ and $W^{(2)}$ is the same as that for $Z^{(1)}$ and $Z^{(2)}$.

The marginal distributions given in Tan's article can be obtained directly from these joint distributions. However, the marginal distributions alone are not sufficient to correctly specify the probabilities of all 14 possible outcomes. In obtaining probabilities such as π_{322} and π_{320}, Tan's mistake is in assuming the independence of events such as $Y^{(1)} > Z^{(1)}$ and $Y^{(2)} > Z^{(2)}$. Consider π_{322}. Tan's calculation proceeds below:

$$\begin{aligned}\pi_{322} &= \Pr(Y^{(1)} > Z^{(1)} \cap Y^{(2)} > Z^{(2)}) \\ &= \Pr(Y^{(1)} > Z^{(1)}) \Pr(Y^{(2)} > Z^{(2)}) \\ &= (0.471)(0.551) \\ &= 0.259.\end{aligned}$$

The correct probability can be written in terms of the joint distributions for ordered outcomes from one, two, or three dice. For example,

$$\begin{aligned}\pi_{322} &= \Pr(Y^{(1)} > Z^{(1)}, Y^{(2)} > Z^{(2)}) \\ &= \sum_{z_1=1}^{5} \sum_{z_2=1}^{z_1} \Pr(Y^{(1)} > z_1, Y^{(2)} > z_2) \Pr(Z^{(1)} = z_1, Z^{(2)} = z_2) \\ &= \sum_{z_1=1}^{5} \sum_{z_2=1}^{z_1} \sum_{y_1=z_1+1}^{6} \sum_{y_2=z_2+1}^{y_1} \Pr(Y^{(1)} = y_1, Y^{(2)} = y_2) \Pr(Z^{(1)} = z_1, Z^{(2)} = z_2) \\ &= \frac{2890}{7776} \\ &= 0.372.\end{aligned}$$

Note that those events in this quadruple sum for which an argument with a subscript of 2 exceeds an argument with the same letter and subscript 1 have probability zero.

The probabilities π_{ijk} that make up the transition probability matrix P can be obtained similarly using the joint distributions for $Y^{(1)}, Y^{(2)}$, for $Z^{(1)}, Z^{(2)}$, and for $W^{(1)}, W^{(2)}$. The probabilities themselves, rounded to the nearest 0.001, are listed in Table 2.

TABLE 2: Probabilities making up the transition probability matrix

i	j	Event	Symbol	Probability	Tan's value
1	1	Defender loses 1	π_{111}	$15/36 = 0.417$	0.417
1	1	Attacker loses 1	π_{110}	$21/36 = 0.583$	0.583
1	2	Defender loses 1	π_{121}	$55/216 = 0.255$	0.254
1	2	Attacker loses 1	π_{120}	$161/216 = 0.745$	0.746
2	1	Defender loses 1	π_{211}	$125/216 = 0.579$	0.578
2	1	Attacker loses 1	π_{210}	$91/216 = 0.421$	0.422
2	2	Defender loses 2	π_{222}	$295/1296 = 0.228$	0.152
2	2	Each lose 1	π_{221}	$420/1296 = 0.324$	0.475
2	2	Attacker loses 2	π_{220}	$581/1296 = 0.448$	0.373
3	1	Defender loses 1	π_{311}	$855/1296 = 0.660$	0.659
3	1	Attacker loses 1	π_{310}	$441/1296 = 0.340$	0.341
3	2	Defender loses 2	π_{322}	$2890/7776 = 0.372$	0.259
3	2	Each lose 1	π_{321}	$2611/7776 = 0.336$	0.504
3	2	Attacker loses 2	π_{320}	$2275/7776 = 0.293$	0.237

The probability of winning a battle Given any initial state, the system will, with probability one, eventually make a transition to an absorbing state. For a transient state i, call $f_{ij}^{(n)}$ the probability that the first (and final) visit to absorbing state j occurs on the nth turn:

$$f_{ij}^{(n)} = \Pr(X_n = j, X_k \neq j \text{ for } k = 1, \ldots, n-1 \mid X_0 = i).$$

Let the $AD \times (D+A)$ matrix of these *first transition* probabilities be denoted by $F^{(n)}$. In order for the chain to begin at state i and enter state j at the nth turn, the first $n-1$ transitions must be among the transient states and the nth must be from a transient state to an absorbing state so that $F^{(n)} = Q^{n-1}R$. The system proceeds for as many turns as are necessary to reach an absorbing state. The probability that the system goes from transient state i to absorbing state j is just the sum $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$. The $AD \times (D+A)$ matrix of probabilities for all of these $D+A$ absorbing states can be obtained from

$$F = \sum_{n=1}^{\infty} F^{(n)} = \sum_{n=1}^{\infty} Q^{n-1}R = (I-Q)^{-1}R.$$

If the system ends in one of the last A absorbing states then the attacker wins; if it ends in one of the first D absorbing states, the defender wins. Since the initial state of a battle is the $i = AD$th state using the order established previously, the probability that the attacker wins is just the sum of the entries in the last (or ADth) row of the submatrix of the last A columns of F:

$$\Pr\big(\text{Attacker wins} \mid X_0 = (A, D)\big) = \sum_{j=D+1}^{D+A} f_{AD,j}$$

and

$$\Pr\big(\text{Defender wins} \mid X_0 = (A, D)\big) = \sum_{j=1}^{D} f_{AD,j}.$$

The row sums of F are unity, which confirms that the system always ends in one of the $D+A$ absorbing states.

TABLE 3: Probability that the attacker wins

D \ A	1	2	3	4	5	6	7	8	9	10
1	0.417	0.106	0.027	0.007	0.002	0.000	0.000	0.000	0.000	0.000
2	0.754	0.363	0.206	0.091	0.049	0.021	0.011	0.005	0.003	0.001
3	0.916	0.656	0.470	0.315	0.206	0.134	0.084	0.054	0.033	0.021
4	0.972	0.785	0.642	0.477	0.359	0.253	0.181	0.123	0.086	0.057
5	0.990	0.890	0.769	0.638	0.506	0.397	0.297	0.224	0.162	0.118
6	0.997	0.934	0.857	0.745	0.638	0.521	0.423	0.329	0.258	0.193
7	0.999	0.967	0.910	0.834	0.736	0.640	0.536	0.446	0.357	0.287
8	1.000	0.980	0.947	0.888	0.818	0.730	0.643	0.547	0.464	0.380
9	1.000	0.990	0.967	0.930	0.873	0.808	0.726	0.646	0.558	0.480
10	1.000	0.994	0.981	0.954	0.916	0.861	0.800	0.724	0.650	0.568

The matrix F is used to obtain Table 3, a matrix of victory probabilities for a battle between an attacker with i armies and a defender with j armies for values of i and j

not greater than 10. Some of these are shown graphically in FIGURE 1, along with some for higher numbers of attacking armies.

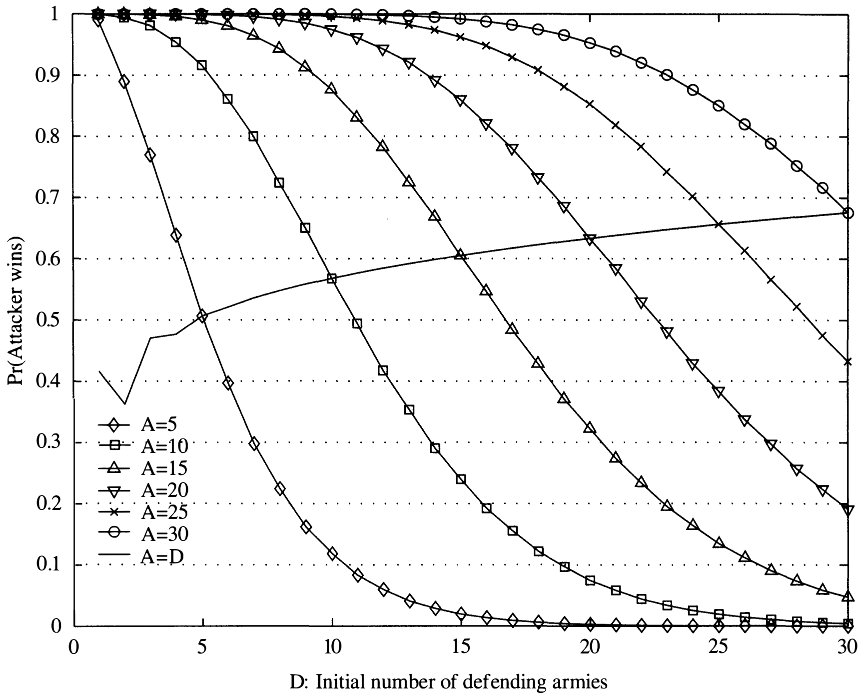


Figure 1 Attacker's winning probabilities at various strengths

Expected losses The expected values and variances for the losses that the attacker and defender will suffer in a given battle can also be found from F. Let L_A and L_D denote the respective losses an attacker and defender will suffer during a given battle given the initial state $X_0 = (A, D)$. Let $R_D = D - L_D$ and $R_A = A - L_R$ denote the number of armies remaining for the attacker and defender respectively. The probability distributions for R_D and R_A can be obtained from the last row of F:

$$Pr(R_D = k) = \begin{cases} f_{AD,k} & \text{for } k = 1, \dots, D \\ 0 & \text{else} \end{cases}$$

and

$$\Pr(R_A = k) = \begin{cases} f_{AD,D+k} & \text{for } k = 1, \dots, A \\ 0 & \text{else.} \end{cases}$$

For example, suppose $A = D = 5$. In this case, the 25th row of the 25×10 matrix F gives the probabilities for the $D + A = 10$ absorbing states:

$$F_{25,\cdot} = (0.068, 0.134, 0.124, 0.104, 0.064, 0.049, 0.096, 0.147, 0.124, 0.091).$$

The mean and standard deviation for the defender's loss in the $A = D = 5$ case are $E(L_D) = 3.56$ and $SD(L_D) = 1.70$. For the attacker, they are $E(L_A) = 3.37$ and $SD(L_A) = 1.83$. Plots of expected losses for values of A and D between 5 and 20 are given in FIGURE 2. This plot shows that the attacker has an advantage in the sense that expected losses are lower than for the defender, provided the initial number of attacking armies is not too small.

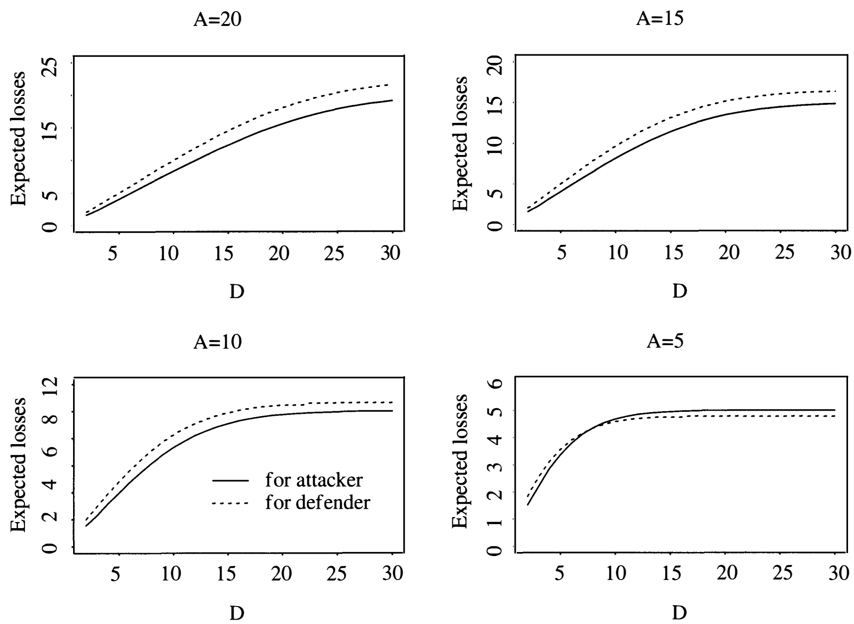


Figure 2 Expected losses for attacker and for defender

Conclusion and recommendations The chances of winning a battle are considerably more favorable for the attacker than was originally suspected. The logical recommendation is then for the attacker to be more aggressive. Inspection of FIGURE 1 shows that when the number of attacking and defending armies is equal ($A = D$), the probability that the attacker ends up winning the territory exceeds 50%, provided the initial stakes are high enough (at least 5 armies each, initially.) This is contrary to Tan's assertion that that this probability is less than 50% because "in the case of a draw, the defender wins" in a given roll of the dice. When $A = D$, FIGURE 2 indicates that the attacker also suffers fewer losses on average than the defender, provided A is not small. With the innovation of several new versions of *RISK* further probabilistic challenges have arisen. *RISK II* enables players to attack simultaneously rather than having to wait for their turn and involves single rolls of nonuniformly distributed dice. The distribution of the die rolled by an attacker or defender depends on the number of armies the player has stationed in an embattled country. The Markovian property of a given battle still holds, but the entries of the transition probability matrix P are different. Further, decisions about whether or not to attack should be made with the knowledge that attacks cannot be called off as in the original *RISK*.

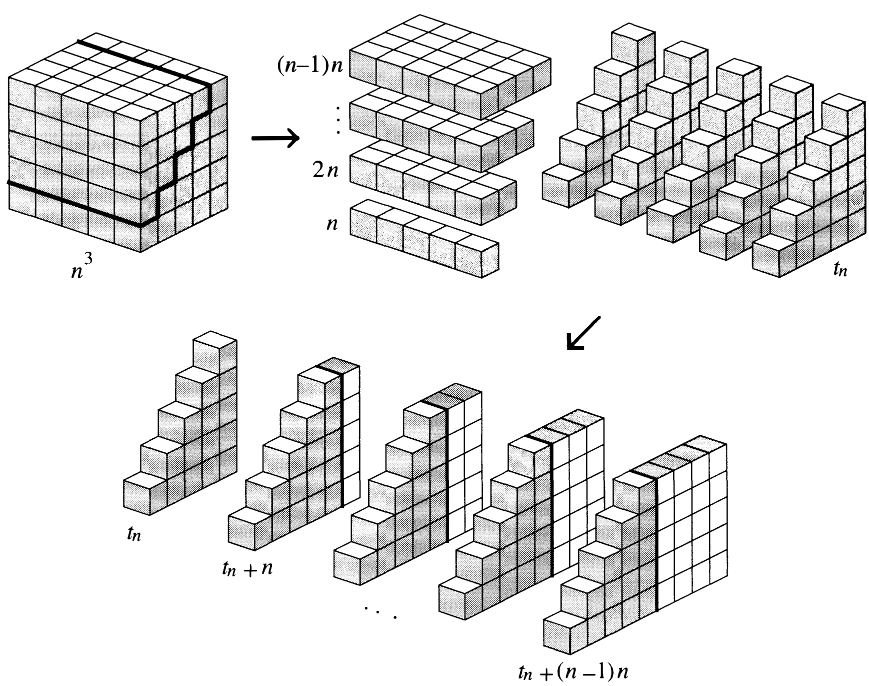
Acknowledgment. The author thanks Jerry Veeh and Dean Hoffman and two referees for comments and confirmations.

REFERENCES

1. Steve Abbott and Matt Richey, Take a Walk on the Boardwalk, *College Math. J.* **28** (1997), 162–171.
2. Robert Ash and Richard Bishop, Monopoly as a Markov Process, this MAGAZINE **45** (1972), 26–29.
3. Baris Tan, Markov Chains and the *RISK* Board Game, this MAGAZINE **70** (1997), 349–357.

Proof Without Words: The Cube as an Arithmetic Sum

$$1 = 1$$
$$8 = 3 + 5$$
$$27 = 6 + 9 + 12$$
$$64 = 10 + 14 + 18 + 22$$
$$\vdots$$
$$t_n = 1 + 2 + \cdots + n \to n^3 = t_n + (t_n + n) + (t_n + 2n) + \cdots + (t_n + (n-1)n)$$



—Roger B. Nelsen
Lewis and Clark College
Portland, OR 97219

Visualizing a Nonmeasurable Set

SALEEM WATSON
ARTHUR WAYMAN
California State University
Long Beach, CA 90840
saleem@csulb.edu
away@csulb.edu

In most real analysis textbooks, the standard example of a nonmeasurable set is a subset of the real line that is due to Vitali [**3**]. We describe a similar nonmeasurable subset of the torus (and hence the plane), where we can more easily visualize the set. In the process of constructing the set, students get an opportunity to experience how topics they learned in algebra and topology can be used in analysis.

The idea of Vitali's example is to express the unit interval I as a disjoint union of countably many mutually congruent sets A_k. The nonmeasurability of each A_k follows from the observation that $I = \bigcup_{k\in\mathbb{Z}} A_k$ and that countable additivity of measure implies that $1 = m(I) = \sum_{k\in\mathbb{Z}} m(A_k)$. Since each set A_k must have the same measure, the last equation shows that no nonnegative value can be assigned as the measure of each A_k. We will use this same idea with the square $[0, 1] \times [0, 1]$ in the plane $\mathbb{R}^2$. The advantage is that we will now have a more visual object than that of Vitali's example because the example will appear as a subset of a torus.

The torus In order to understand the example that we will eventually construct, we need to consider different ways of describing the torus. We will exploit topological and group theoretic properties associated with two different representations of a torus to obtain information that we can piece together to construct an interesting example of a nonmeasurable set.

Begin by considering the square $[0, 1] \times [0, 1]$ as a topological subspace of $\mathbb{R}^2$ endowed with the usual topology. After identifying opposite edges of the square, we obtain (via the identification topology) a space called the *torus,* denoted by $\mathbb{T}$. A convenient way to visualize the torus is as the surface of a doughnut. In fact, the mapping $\Omega : [0, 1] \times [0, 1] \to \mathbb{R}^3$ given by

$$\Omega(r, s) = \Big(\big[2 + \cos(2\pi s)\big]\cos(2\pi r), \big[2 + \cos(2\pi s)\big]\sin(2\pi r), \sin(2\pi s)\Big)$$

renders a concrete parametrization of the torus as a subspace of $\mathbb{R}^3$. The mapping Ω identifies the pair of edges $[0, 1] \times \{0\}$ and $[0, 1] \times \{1\}$ as well as the pair of edges $\{0\} \times [0, 1]$ and $\{1\} \times [0, 1]$, as in FIGURE 1. This latter pair of edges of the square, labelled M in the figure, corresponds to a circle on the torus called a *meridian.*

Another way to visualize the torus is as the topological product of two circles $\mathbb{S}^1 \times \mathbb{S}^1$, where $\mathbb{S}^1 = \{e^{2\pi ir} \mid 0 \le r \le 1\}$ is the unit circle in the complex plane $\mathbb{C}$. Viewed in this way, the torus is a topological group under componentwise multiplication. The mapping

$$\Psi : \mathbb{R}^2 \to \mathbb{T}$$

given by $\Psi(r, s) = (e^{2\pi ir}, e^{2\pi is})$ satisfies

$$\Psi\big((a, b) + (c, d)\big) = \Psi\big((a, b)\big)\Psi\big((c, d)\big)$$

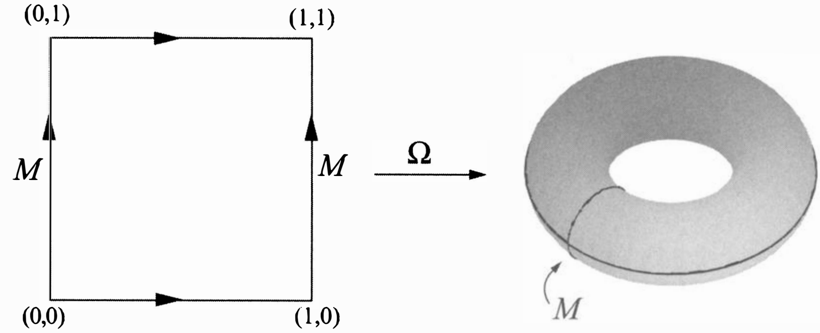


Figure 1 A torus is a square with opposite edges identified

where (a, b) and (c, d) are points in $\mathbb{R}^2$. Indeed, Ψ is a continuous surjective group homomorphism (actually it is a covering map) from the additive group $\mathbb{R}^2$ onto the multiplicative group $\mathbb{T}$. Moreover, the points (a, b) and (c, d) are identified via the mapping Ψ if $(e^{2\pi ia}, e^{2\pi ib}) = (e^{2\pi ic}, e^{2\pi id})$. Thus each unit square in $\mathbb{R}^2$ is wrapped once around the torus by Ψ. Note also that each of the vertical lines labelled M_k in FIGURE 2 corresponds to the meridian M.

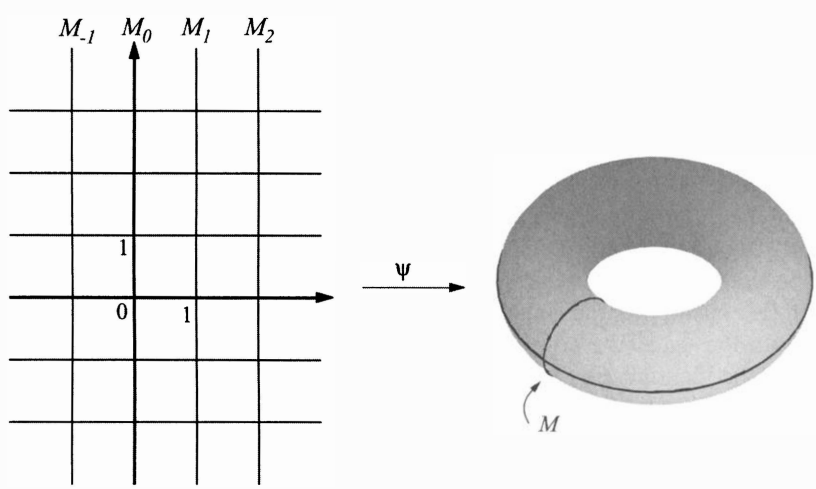


Figure 2 Each unit square in the plane is identified with the torus by Ψ

The mapping Ψ suggests yet another way to describe the torus. It is the quotient space of $\mathbb{R}^2$ relative to the following equivalence relation: two points (a, b) and (c, d) in $\mathbb{R}^2$ are identified if $c = a + k$ and $d = b + l$ for some integers k and l. When this is the case we will write $(a, b) \equiv (c, d) \bmod 1$.

One-parameter subgroups of $\mathbb{T}$ Our goal in this section is to describe a family of continuous group homomorphisms from $\mathbb{R}$ into the torus $\mathbb{T}$. Such a map would wrap the real line on the torus. To visualize such a map, we will first send $\mathbb{R}$ into $\mathbb{R}^2$ and then identify $\mathbb{R}^2$ with the torus via the map Ψ.

Let α and β be fixed real numbers. A mapping $\varphi : \mathbb{R} \to \mathbb{R}^2$ given by $\varphi(t) = (\alpha t, \beta t)$ is a continuous group homomorphism between the additive groups $\mathbb{R}$ and $\mathbb{R}^2$. The image of φ is called a *one-parameter subgroup* of $\mathbb{R}^2$. This image is the line whose Cartesian equation is $y = (\beta/\alpha)x$. It is easy to see that if two points $(\alpha s, \beta s)$

and $(\alpha t, \beta t)$ corresponding to $s \neq t$ in $\mathbb{R}$ are equivalent mod 1, then β/α is a rational fraction.

Now let's suppose that the fixed real numbers α and β have an irrational ratio. In this case, if $(\alpha s, \beta s) \equiv (\alpha t, \beta t)$ mod 1, then $s = t$. Thus the mapping

$$\Psi \circ \varphi : \mathbb{R} \to \mathbb{T}$$

is injective, it is also a continuous group homomorphism whose image is a one-parameter subgroup of $\mathbb{T}$. Let L denote the image of φ in $\mathbb{R}^2$. Then L is a line in the plane. We can visualize the image $\Psi(L)$ as a coil on the torus, a piece of which is shown in bold in FIGURE 3. Proving that this coil is dense in the torus makes a nice exercise, though we will not do so here.

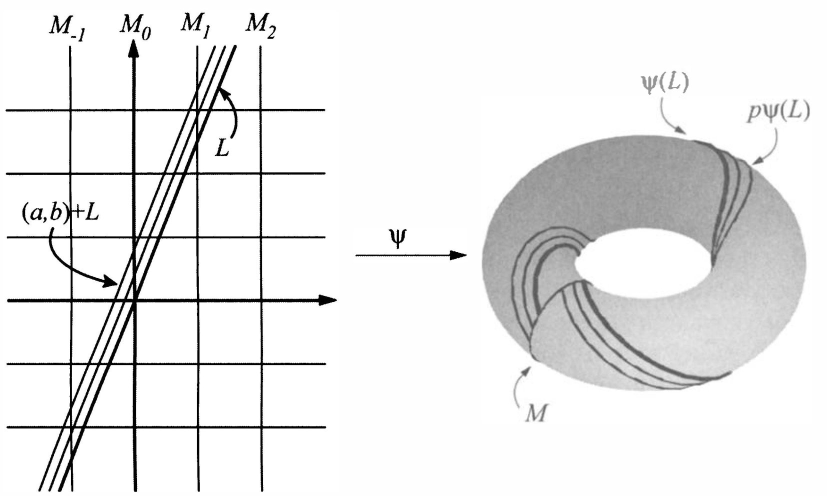


Figure 3 The line L and its translates correspond to parallel coils on the torus

The line L is a subgroup of the additive group $\mathbb{R}^2$. For $(a, b) \in \mathbb{R}^2$ the coset $(a, b) + L$ is a line parallel to L, often referred to as a *translate* of L. Since Ψ is a homomorphism the coil $\Psi(L)$ is a subgroup of the multiplicative group $\mathbb{T}$. For $p \in \mathbb{T}$, the coset $p\Psi(L)$ is a coil parallel to $\Psi(L)$, because it is the image of a line $(a, b) + L$ parallel to L. Indeed, if $p = \Psi((a, b))$, then

$$\Psi\big((a, b) + L\big) = \Psi\big((a, b)\big)\Psi(L) = p\Psi(L).$$

In other words, under the map Ψ, parallel lines of the form $(a, b) + L$ in $\mathbb{R}^2$ correspond to parallel coils of the form $p\Psi(L)$ in $\mathbb{T}$. FIGURE 3 shows three such lines and the corresponding coils. It is important to note that the lines $(a, b) + L$ and $(c, d) + L$ correspond to the same coil on the torus if and only if $(a, b) \equiv (c, d)$ mod 1.

Now, each coil $p\Psi(L)$ intersects the meridian M in $\mathbb{T}$ infinitely many times; this is easily seen in each of the representations of the torus. So by the Axiom of Choice there is a subset Λ of M such that each coset of $\Psi(L)$ is represented by a unique point in Λ. Thus the sets $p\Psi(L)$, $p \in \Lambda$, form a complete set of cosets of $\Psi(L)$ in $\mathbb{T}$, so we obtain the disjoint union

$$\bigcup_{p\in\Lambda}(p\Psi(L)) = \mathbb{T}.$$

In other words, we can visualize the torus as the disjoint union of uncountably many parallel coils, each one a coset of $\Psi(L)$.

A nonmeasurable subset of $\mathbb{T}$ Now we will construct a nonmeasurable set by partitioning the torus into a disjoint union of countably many geometrically congruent sets A_k. We begin with the subsets in the plane that consist of those parts of the translates (cosets) of L that lie in the strip between M_k and M_{k+1}. The sets A_k are the corresponding sets in the torus. They can be visualized as the portions of each coil $p\Psi(L)$ starting and ending at M (see FIGURE 3).The details are as follows.

For each integer k in $\mathbb{Z}$, define the set

$$L^{(k)} = \left\{(\alpha t, \beta t) \,\middle|\, \frac{k}{\alpha} \le t < \frac{k+1}{\alpha}\right\}.$$

Observe that the sets $L^{(k)}$ are pairwise disjoint; they are merely half-open intervals on the line L. In fact, they are the parts of the line L between consecutive vertical lines M_k in FIGURE 3. The corresponding sets on the torus are $\Psi(L^{(k)})$. Now for each integer k in $\mathbb{Z}$, set $A_k = \bigcup_{p\in\Lambda} p\Psi(L^{(k)})$ and observe that

$$\mathbb{T} = \bigcup_{k\in\mathbb{Z}} A_k.$$

Finally, the subsets A_k of the torus are pairwise disjoint by construction, pairwise congruent via translation (a multiplication in $\mathbb{T}$), and there are countably many of them. The Lebesgue measure of the torus $\mathbb{T}$ is its surface area, a positive number. Recall that Lebesgue measure is translation invariant and countably additive. Therefore, if the sets A_k are measurable, then they have the same positive measure. Since the torus $\mathbb{T}$ is the countable union of such sets, the sets A_k cannot be measurable.

REFERENCES

1. Stephan C. Carlson, *Topology of Surfaces, Knots, and Manifolds: A First Undergraduate Course*, John Wiley & Sons, Inc., New York, 2001.
2. Serge Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1993.
3. H. L. Royden, *Real Analysis*, 2nd ed., Macmillian, New York, 1968.

Nondifferentiability of the Ruler Function

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

A fixture of any introductory course in mathematical analysis is the pathological function, one whose intuition-defying behavior serves to crystallize our understanding of analytic concepts. Among the more accessible of these is the so-called ruler function, defined on (0, 1) by

$$r(x) = \begin{cases} 1/q & \text{if } x = p/q \quad \text{(lowest terms)} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

An $\epsilon - \delta$ proof shows that, if $a \in (0, 1)$, then $\lim_{x\to a} r(x) = 0$. Thus r is continuous at each irrational point and discontinuous at each rational point of (0, 1). But in spite of its being so wildly discontinuous, r is (Riemann) integrable over the unit interval, with

$$\int_0^1 r(x)\,dx = 0.$$

(These basic properties of the ruler function appear, for instance, in Abbott [**1**, p. 102, p. 203].) Students at the beginning of their mathematical careers find this pathological indeed.

Although an introductory course may treat continuity and even integrability of the ruler function, it is less likely to address that third pillar of analysis: differentiability. In fact, the ruler function is nowhere differentiable on (0, 1), but proofs of this seem hard to come by (see [**2, 3, 4**]). What follows is an argument that is short, straightforward, and—as an added attraction—features a cameo appearance by no less a mathematician than Euclid himself.

THEOREM. *The ruler function is nowhere differentiable on* (0, 1).

Proof. Being discontinuous at each rational, the ruler function could be differentiable only at irrational points, so for the sake of contradiction we assume that

$$r'(a) = \lim_{x\to a} \frac{r(x) - r(a)}{x - a}$$

exists for some irrational a in (0, 1).

Letting $\{x_n\}$ be a sequence of irrationals in (0, 1) for which $x_n \neq a$ for all n but where $\lim_{n\to\infty} x_n = a$, we see that

$$r'(a) = \lim_{n\to\infty} \frac{r(x_n) - r(a)}{x_n - a} = 0.$$

Consequently, for $\epsilon = 1$, there exists a $\delta > 0$ so that, if $0 < |x - a| < \delta$ then

$$\left|\frac{r(x) - r(a)}{x - a}\right| < 1 \quad \text{and therefore}$$

$$r(x) < |x - a|. \tag{1}$$

Now $1/\delta > 0$, and because $0 < a < 1$ we know that $1/a > 0$ and $1/(1 - a) > 0$ as well. Recalling Euclid's proof that there is no largest prime number, we choose a prime $P > \max\{1/\delta, 1/a, 1/(1 - a)\}$. It is clear that the interval $(a - 1/P, a + 1/P)$ is contained in (0, 1). In addition, because this interval has width $2/P$, there exists a whole number k with $1 \leq k < P$ for which the rational k/P belongs to it. We observe that $k/P \neq a$, because the latter is irrational, and that k/P is in lowest terms because P is prime. It follows that $0 < |k/P - a| < 1/P < \delta$ and so, by (1) above,

$$1/P = r(k/P) < |k/P - a| < 1/P,$$

a contradiction. Consequently, r is nowhere differentiable on (0, 1). ∎

One is tempted to conclude that, when it comes to differentiability, the ruler function just doesn't measure up.

REFERENCES

1. S. Abbott, *Understanding Analysis*, Springer-Verlag, New York, 2001.
2. G. A. Heuer, Functions continuous at the rationals and discontinuous at the irrationals, *Amer. Math. Monthly*, **72** (1965), 370–373.
3. J. E. Nymann, An application of diophantine approximation, *Amer. Math. Monthly*, **76** (1969), 668–671.
4. G. J. Porter, On the differentiability of a certain well-known function, *Amer. Math. Monthly*, **69** (1962), 142.

Math Bite: Convergence of p-series

We show the convergence of

$$\sum_{k=1}^{\infty} \frac{1}{k^p}$$

for $p > 1$.

Let $p = 1 + q, q > 0$. The sequence of partial sums, whose nth term is $S_n = \sum_{k=1}^{n} 1/k^p$, is monotone increasing. It is also bounded, as follows. Let $n = 10^j - 1$, then

$$\sum_{k=1}^{10^j-1} \frac{1}{k^p} = 1 + \frac{1}{2^p} + \cdots + \frac{1}{k^p} + \cdots + \frac{1}{(10^j-1)^p}$$

$$= 1 + \underbrace{\frac{1}{2^p} + \cdots + \frac{1}{9^p}}_{\text{9 terms}} + \underbrace{\frac{1}{10^p} + \frac{1}{11^p} + \cdots + \frac{1}{99^p}}_{\text{90 terms}} + \underbrace{\frac{1}{100^p} + \frac{1}{101^p} + \cdots + \frac{1}{999^p}}_{\text{900 terms}} + \cdots$$

$$< 1 + \cdots + 1 + \frac{1}{10^p} + \cdots + \frac{1}{10^p} + \frac{1}{100^p} + \cdots + \frac{1}{100^p} + \cdots + \frac{1}{(10^{j-1})^p}$$

$$= 9 + \frac{90}{10^p} + \frac{900}{100^p} + \cdots = 9\left(1 + \frac{1}{10^q} + \frac{1}{10^{2q}} + \cdots + \frac{1}{10^{(j-1)q}}\right) < \frac{9}{1 - 10^{-q}}.$$

Readers may wish to adapt the argument to show divergence in the case where $p < 1$. The first step is

$$\sum_{k=1}^{10^j} \frac{1}{k^p} = 1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p} + \cdots + \frac{1}{(10^j)^p}$$

$$= 1 + \underbrace{\frac{1}{2^p} + \cdots + \frac{1}{10^p}}_{\text{9 terms}} + \underbrace{\frac{1}{11^p} + \frac{1}{12^p} + \cdots + \frac{1}{100^p}}_{\text{90 terms}} + \underbrace{\frac{1}{101^p} + \frac{1}{102^p} + \cdots + \frac{1}{1000^p}}_{\text{900 terms}} + \cdots$$

$$> \frac{10}{10^p} + \frac{90}{100^p} + \frac{900}{1000^p} + \cdots.$$

—SIDNEY H. KUNG
UNIVERSITY OF NORTH FLORIDA
JACKSONVILLE, FL 32224
SHKUNG@TU.INFI.NET

[*Editor's note: Another manuscript received at about the same time, from Eugene Boman and Richard Brazier of Penn State University, Dubois Campus, presented this* *same idea using powers of* 2. *A referee pointed out that both methods amount to the Cauchy Condensation Test. See Konrad Knopp's classic books* Infinite Sequences and Series *or* Theory and Application of Infinite Series.]

A Classification of Matrices of Finite Order over $\mathbb{C}$, $\mathbb{R}$, and $\mathbb{Q}$

REGINALD KOO
University of South Carolina Aiken
Aiken, SC 29801
regk@usca.edu

A matrix A is said to have *finite order* $n \geq 1$ if $A^n = I$ and $A^r \neq I$ for $1 \leq r < n$. Otherwise we say that A has infinite order. An elementary exercise in abstract algebra asks for 2×2 matrices A, B over $\mathbb{R}$ each of finite order such that AB has infinite order. The matrices

$$A_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{and} \quad B_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

represent rotation about the origin through the signed angle θ and reflection in the line $y = x\tan(\theta/2)$. The matrix A_θ has finite order if and only if θ is a rational multiple of 2π, whereas every matrix B_θ has order 2. Moreover $A_\theta A_\phi = A_{\theta+\phi}$, $A_\theta B_\phi = B_{\theta+\phi}$, $B_\phi A_\theta = B_{\phi-\theta}$, and $B_\theta B_\phi = A_{\theta-\phi}$. Now let θ be an irrational multiple of 2π. Then the reflection matrices B_θ and B_0 have finite order, and their product $B_\theta B_0 = A_\theta$ has infinite order.

Are there examples other than reflections? To answer this it is natural to consider the matrices of finite order in $GL(2, \mathbb{R})$, the multiplicative group of nonsingular 2×2 matrices. The purpose of this note is to classify the matrices of finite order in $GL(k, F)$ for the fields $F = \mathbb{C}$, $\mathbb{R}$, and $\mathbb{Q}$, and to provide further examples of finite order matrices whose product has infinite order. The solution to this classification problem involves the factorization of $x^n - 1$ over F, and an application of the cyclic decomposition theorem of linear algebra. In this connection, we mention the paper [**3**] in which Robert Hanson determines, for a given n, the minimum k for which there is a $k \times k$ matrix A over F of order n, when F is $\mathbb{C}$, $\mathbb{R}$, or $\mathbb{Q}$.

Note that when F is a finite field with q elements then $GL(k, F)$ is a finite group of order $(q^k - 1)(q^k - q)\cdots(q^k - q^{k-1})$ [**6**, p. 178], so that each $k \times k$ matrix over F has finite order.

Minimal polynomials The text Blyth & Robertson [**1**] contains a concise account, with proofs, of the results of linear algebra stated here.

Having the same order (finite or infinite) is an equivalence relation in the multiplicative group $GL(k, F)$. We say that A is similar to B, denoted by $A \sim B$, if there exists a nonsingular matrix P such that $B = P^{-1}AP$. If A is similar to B, then A and B have the same order.

We denote the set of all $k \times k$ matrices over the field F, singular and nonsingular, by $M_k(F)$. If $A \in M_k(F)$ there is a polynomial $p \in F[x]$ for which $p(A) = 0$. One such polynomial is the *characteristic polynomial* of A defined by $\chi_A(x) = \det(xI - A)$.

The *minimal polynomial* of A is the monic polynomial $m_A \in F[x]$ of least degree satisfying $m_A(A) = 0$. The minimal and characteristic polynomials have the same zeroes over F and thus have the form

$$m_A = p_1^{\alpha_1} \dots p_r^{\alpha_r}, \qquad \chi_A = p_1^{\beta_1} \dots p_r^{\beta_r},$$

where the p_i are distinct monic irreducible polynomials over F, and α_i, β_i are integers with $1 \le \alpha_i \le \beta_i$. Note that, by considering dimensions, we have $\sum_{i=1}^{r} \beta_i \deg(p_i) = k$. A matrix is diagonalizable over F if and only if its minimal polynomial is a product of distinct linear factors over F. And, if p is a polynomial over F such that $p(A) = 0$, then the minimal polynomial m_A divides p. Thus if A has finite order n, then m_A divides $x^n - 1$. The irreducible factorization of the polynomial $x^n - 1$ over the fields $\mathbb{C}$, $\mathbb{R}$, and $\mathbb{Q}$ are well known, and we will consider these below.

Factorization of $x^n - 1$ Denote the complex nth roots of unity by $1, \omega, \dots, \omega^{n-1}$. Then the irreducible factorization over $\mathbb{C}$ can be expressed as

$$x^n - 1 = (x-1)(x-\omega)\cdots(x-\omega^{n-1}).$$

From this factorization over $\mathbb{C}$, we obtain the factorization of $x^n - 1$ over $\mathbb{R}$ by combining conjugate pairs of factors:

$$(x-\omega^j)(x-\omega^{n-j}) = x^2 - \left(2\cos\tfrac{2\pi j}{n}\right)x + 1.$$

We write $p_\theta = x^2 - (2\cos\theta)x + 1$ and $\theta_j = 2\pi j/n$. Then over $\mathbb{R}$ we have the factorization

$$\begin{aligned} x^n - 1 &= (x-1)(x+1)p_{\theta_1}\dots p_{\theta_r}, &&\text{when } n = 2r+2,\\ \text{or}\quad x^n - 1 &= (x-1)p_{\theta_1}\dots p_{\theta_r}, &&\text{when } n = 2r+1. \end{aligned}$$

We obtain the factorization of $x^n - 1$ over $\mathbb{Q}$ by combining factors involving the primitive mth roots of unity as m ranges over the divisors of n. Let U_m denote the multiplicative group of all mth roots of unity. An element $\xi \in U_m$ is called a *primitive mth root of unity* if ξ has order m. Denote the set of primitive mth roots of unity by Ω_m. Then $\Omega_m = \{\exp(2\pi ir/m) : 1 \le r \le m,\ (r, m) = 1\}$. The number of elements in Ω_m is denoted by $\phi(m)$, known as *Euler's phi-function*. The mth *cyclotomic polynomial* is the monic polynomial Φ_m whose roots in $\mathbb{C}$ are the primitive mth roots of unity, that is,

$$\Phi_m = \prod_{\xi\in\Omega_m}(x-\xi).$$

The first few cyclotomic polynomials are: $\Phi_1 = x - 1$, $\Phi_2 = x + 1$, $\Phi_3 = x^2 + x + 1$, $\Phi_4 = x^2 + 1$, $\Phi_5 = x^4 + x^3 + x^2 + x + 1$, $\Phi_6 = x^2 - x + 1$. Recursion formulae for Φ_m are set out in Lang [**5**, pp. 206–207], where in addition it is shown that Φ_m has integer coefficients and is irreducible over $\mathbb{Q}$. The group U_n is the disjoint union of the sets Ω_m for all divisors m of n. The irreducible factorization of $x^n - 1$ over $\mathbb{Q}$ is then

$$x^n - 1 = \prod_{m|n}\Phi_m.$$

We illustrate this by finding the factorization of $x^8 - 1$. Let $\omega = e^{2\pi i/8}$. The divisors of 8 are 1, 2, 4, 8. The primitive 1st root of unity is 1. The primitive 2nd root of unity is $-1 = \omega^4$. The primitive 4th roots of unity are $e^{2\pi i/4} = \omega^2$ and $e^{6\pi i/4} = \omega^6$. The primitive 8th roots of unity are ω, ω^3, ω^5, ω^7. The irreducible factorization over $\mathbb{Q}$ is then

$$\begin{aligned} x^8 - 1 &= \Phi_1\,\Phi_2\,\Phi_4\,\Phi_8 \\ &= (x-1)(x+1)(x^2+1)(x^4+1). \end{aligned}$$

Since these factorizations are products of *distinct* irreducible factors, we see that if the matrix A has finite order over $\mathbb{C}$, $\mathbb{R}$, or $\mathbb{Q}$, then m_A is a product $m_A = p_1 \cdots p_r$ of distinct irreducible polynomials.

Cyclic decomposition The *companion matrix* of a monic polynomial

$$f = a_0 + a_1 x + \cdots + a_{k-1} x_{k-1} + x^k$$

is the $k \times k$ matrix

$$C(f) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & 0 & \dots & 0 & -a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -a_{k-1} \end{bmatrix}.$$

The minimal polynomial of $C(f)$ is f (Herstein [**4**, p. 307]). The companion matrix of a linear polynomial $a_0 + x$ is simply the matrix $[-a_0]$.

Given matrices A_1 and A_2, their *direct sum* is the block diagonal matrix

$$A_1 \oplus A_2 = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}.$$

We denote the direct sum of j copies of a matrix C by $C^{[j]}$. The order of a direct sum is the least common multiple (lcm) of the orders of its summands.

We state the cyclic decomposition theorem in the special case when the minimal polynomial is a product of distinct irreducible polynomials.

CYCLIC DECOMPOSITION. *If the minimal and characteristic polynomials of A in $M_k(F)$ are*

$$m_A = p_1 \dots p_r, \qquad \chi_A = p_1^{e_1} \dots p_r^{e_r},$$

where the p_i are distinct monic irreducible polynomials in $F[x]$, then

$$A \sim C(p_1)^{[e_1]} \oplus \cdots \oplus C(p_r)^{[e_r]}.$$

As an illustration, if $A \in M_8(\mathbb{R})$ has

$$m_A = (x^2+1)(x-1) \quad \text{and} \quad \chi_A = (x^2+1)^3(x-1)^2,$$

then

$$A \sim C(x^2+1) \oplus C(x^2+1) \oplus C(x^2+1) \oplus C(x-1) \oplus C(x-1).$$

Matrices of finite order over $\mathbb{C}$ Let A in $GL(k, \mathbb{C})$ have finite order n. Then m_A divides $(x-1)(x-\omega)\dots(x-\omega^{n-1})$. It follows that m_A factors into distinct linear factors over $\mathbb{C}$, so that A is diagonalizable over $\mathbb{C}$. Thus A is similar to a diagonal matrix whose diagonal entries λ_i are members of $U_n = \{z \in \mathbb{C} \mid z^n = 1\}$. Let $G = \bigcup_{n=1}^{\infty} U_n$ be the multiplicative group of all roots of unity. The order of any $z \in G$ is the least positive n for which $z^n = 1$. If A is a diagonal $k \times k$ matrix whose diagonal entries λ_i belong to G, then A has finite order equal to the lcm of the orders of the λ_i.

Thus we obtain the classification of matrices of finite order over $\mathbb{C}$.

THEOREM. *A matrix A in $GL(k, \mathbb{C})$ has finite order if and only if A is similar to a diagonal matrix* $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ *for some $\lambda_1, \lambda_2, \ldots, \lambda_k$ in the multiplicative group G of all complex roots of unity.*

The order of such a matrix is the lcm of the orders of $\lambda_1, \lambda_2, \ldots, \lambda_k$.

Thus for a given $k \geq 1$, there exist $k \times k$ matrices over $\mathbb{C}$ of any finite order.

Matrices of finite order over $\mathbb{R}$ The characteristic polynomial of the 2×2 rotation matrix A_θ is the polynomial $p_\theta = x^2 - (2\cos\theta)x + 1$. If $0 < \theta < \pi$, then p_θ is irreducible over $\mathbb{R}$, so that $m_{A_\theta} = p_\theta$. It follows from the cyclic decomposition theorem that A_θ is similar over $\mathbb{R}$ to $C(p_\theta)$, for $0 < \theta < \pi$. Additionally one may see this by verifying that $QA_\theta = C(p_\theta)Q$, where

$$Q = \begin{bmatrix} \sin\theta & -\cos\theta \\ 0 & 1 \end{bmatrix}.$$

Now let $A \in GL(k, \mathbb{R})$ have finite order n, so that m_A divides $x^n - 1$. Then there exist $\epsilon_1, \epsilon_2 \in \{0, 1\}$, an integer $r \geq 0$, real numbers θ_i with $0 < \theta_1 < \cdots < \theta_r < \pi$, such that each θ_i is an integer multiple of $2\pi/n$ and

$$m_A = (x-1)^{\epsilon_1}(x+1)^{\epsilon_2} p_{\theta_1} \cdots p_{\theta_r}.$$

Moreover there are integers $k_1 \geq \epsilon_1, k_2 \geq \epsilon_2, d_1, \ldots, d_r \geq 1$ such that

$$\chi_A = (x-1)^{k_1}(x+1)^{k_2} p_{\theta_1}^{d_1} \cdots p_{\theta_r}^{d_r},$$

where if $\epsilon_1 = 0$ then $k_1 = 0$, and if $\epsilon_2 = 0$ then $k_2 = 0$. Since $C(p_\theta) \sim A_\theta$, we obtain the cyclic decomposition

$$A \sim I_{k_1} \oplus (-I_{k_2}) \oplus A_{\theta_1}^{[d_1]} \oplus \cdots \oplus A_{\theta_r}^{[d_r]},$$

where I_k is the $k \times k$ identity matrix. Note that $k = k_1 + k_2 + 2(d_1 + \cdots + d_r)$. If $\theta = 2\pi a/b$ is a rational multiple of 2π with a/b in lowest terms then A_θ has finite order b.

Thus we obtain a classification over $\mathbb{R}$.

THEOREM. *A matrix A in $GL(k, \mathbb{R})$ has finite order if and only if A is similar to*

$$I_{k_1} \oplus (-I_{k_2}) \oplus A_{\theta_1}^{[d_1]} \oplus \cdots \oplus A_{\theta_r}^{[d_r]},$$

where $k_1, k_2 \geq 0$, $r \geq 0$, $d_1, \ldots, d_r \geq 1$, $0 < \theta_1 < \cdots < \theta_r < \pi$, each θ_i is a rational multiple of 2π, and $k_1 + k_2 + 2(d_1 + \cdots + d_r) = k$.

Writing $\theta_i = 2\pi a_i/b_i$ with a_i/b_i in lowest terms, the order of such an A is $\operatorname{lcm}\{2, b_1, \ldots, b_r\}$ *or* $\operatorname{lcm}\{b_1, \ldots, b_r\}$ *according as $k_2 > 0$ or $k_2 = 0$ respectively.*

COROLLARY. *A matrix A in $GL(2, \mathbb{R})$ has finite order if and only if A is similar to a rotation matrix A_θ, where $0 \leq \theta \leq \pi$ is a rational multiple of 2π, or to the reflection matrix B_0. That is,*

$$A \sim \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{or} \quad A \sim \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus for $k \geq 2$ there exist $k \times k$ matrices over $\mathbb{R}$ of any finite order.

EXAMPLE. We give an example of matrices A, B of finite order whose product AB has infinite order, and A is neither a rotation nor reflection matrix. Let $A = P^{-1}JP$ where

$$P = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Let B be the rotation matrix corresponding to $\theta = 5\pi/4$. Thus

$$A = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}, \quad B = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}, \quad \text{and} \quad AB = \frac{1}{\sqrt{2}}\begin{bmatrix} -2 & 0 \\ 1 & 1 \end{bmatrix}.$$

Then $A^2 = I = B^8$. If $X = [0 \quad 1]^T$ then $(AB)X = 2^{-1/2}X$ and $(AB)^n X = 2^{-n/2}X$ for all $n \geq 1$. Thus AB has infinite order.

Matrices of finite order over $\mathbb{Q}$ Let A in $GL(k, \mathbb{Q})$ have finite order n. Then the irreducible factorizations of m_A and χ_A over $\mathbb{Q}$ are given by

$$m_A = \Phi_{m_1} \ldots \Phi_{m_r}, \quad \text{and} \quad \chi_A = \Phi_{m_1}^{d_1} \ldots \Phi_{m_r}^{d_r},$$

where $r \geq 1$, $m_1, \ldots, m_r$ are distinct and divide n, and $d_i \geq 1$. It follows that $d_1\phi(m_1) + \cdots + d_r\phi(m_r) = k$. The cyclic decomposition is given by

$$A \sim C(\Phi_{m_1})^{[d_1]} \oplus \cdots \oplus C(\Phi_{m_r})^{[d_r]}.$$

The order of $C(\Phi_m)$ is determined as follows. Let $\alpha \geq 1$ be the order of $C = C(\Phi_m)$. Since the minimal polynomial of C is Φ_m, and Φ_m divides $x^m - 1$, we deduce that $C^m = I$, so that m is a multiple of α. On the other hand $C^\alpha = I$ and so Φ_m divides $x^\alpha - 1 = \prod_{d|\alpha} \Phi_d$. Hence $\Phi_m = \Phi_d$ for some divisor d of α. As the cyclotomic polynomials are distinct, we deduce that $m = d$, that is, m is a divisor of α. Thus the order of $C(\Phi_m)$ is m. We have proved the following:

THEOREM. *A matrix A in $GL(k, \mathbb{Q})$ has finite order if and only if A is similar to*

$$C(\Phi_{m_1})^{[d_1]} \oplus \cdots \oplus C(\Phi_{m_r})^{[d_r]}$$

where $r \geq 1$, $m_1 < \cdots < m_r$, $d_1, \ldots, d_r \geq 1$ and $d_1\phi(m_1) + \cdots + d_r\phi(m_r) = k$. The order of such an A is $\operatorname{lcm}\{m_1, \ldots, m_r\}$.

To illustrate this theorem, let $C[\alpha, \ldots, \omega]$ denote the direct sum $C(\Phi_\alpha) \oplus \cdots \oplus C(\Phi_\omega)$. Then $C[\alpha, \ldots, \omega]$ has order $\operatorname{lcm}\{\alpha, \ldots, \omega\}$ and size $k \times k$, where $k = \phi(\alpha) + \cdots + \phi(\omega)$. For instance

$$C[3, 4] = C(\Phi_3) \oplus C(\Phi_4) = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

has order 12 and size $k \times k$ where $k = \phi(3) + \phi(4) = 4$.

We now specialize to the case $k=2$ of this theorem. This table of the function $\phi(n)$ is sufficient for our purposes:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\phi(n)$	1	1	2	2	4	2	6	4	6	4	10	4	12	6	8	8

The inequality $\sqrt{n}/2 \leq \phi(n)$ (Burton [**2**, p. 141]) implies that to find $\phi^{-1}(k)$ we need only check a table of $\phi(n)$ for $n \leq 4k^2$. There are two cases. Solving $\phi(m_1) + \phi(m_2) = 2$ gives $(m_1, m_2) = (1,1), (1,2), (2,2)$. Solving $\phi(m_1) = 2$ gives $m_1 = 3, 4, 6$. Hence $A \in GL(2, \mathbb{Q})$ has finite order if and only if A is similar over $\mathbb{Q}$ to one of the six matrices C[1,1], C[1,2], C[2,2], C[3], C[4], C[6] of orders 1, 2, 2, 3, 4, 6 respectively.

COROLLARY. *A matrix A in $GL(2, \mathbb{Q})$ has finite order if and only if A is similar to one of*

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}.$$

In contrast to the complex and real cases, the only possible finite orders for a rational 2×2 matrix are 1, 2, 3, 4, and 6. We can show, by a similar analysis, that there are ten similarity classes of finite order matrices in $GL(3, \mathbb{Q})$ with possible orders $\{1, 2, 3, 4, 6\}$, and twenty-four such classes in $GL(4, \mathbb{Q})$ with possible orders $\{1, 2, 3, 4, 5, 6, 8, 10, 12\}$.

EXAMPLE. The matrices

$$A = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

have orders 3, 2 respectively, and $AB = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ has infinite order. It is easy to see that AB is not similar over $\mathbb{Q}$ to any of the above finite order matrices by comparing characteristic polynomials.

Acknowledgments. I wish to thank a former colleague Dr. Ernest Eckert who, in casual conversation, mentioned the exercise in the introduction, and kindly brought the paper by Hanson [**3**] to my attention.

REFERENCES

1. T. S. Blyth and E. F. Robertson, *Linear Algebra*, Chapman and Hall, London, 1986.
2. David M. Burton, *Elementary Number Theory*, Revised printing, Allyn and Bacon, Boston, 1980.
3. Robert Hanson, Minimum dimension for a square matrix of order n, *College Math. J.* **21** (1990), 28–34.
4. I. N. Herstein, *Topics in Algebra*, 2nd ed., Wiley, New York, 1975.
5. Serge Lang, *Algebra*, Addison Wesley, Reading, MA, 1965.
6. C. W. Norman, *Undergraduate Algebra: A First Course*, Oxford Univ. Pr., New York, 1986.

Math Bite: The Extra Distance in an Outer Lane of a Running Track

ELLIOTT A. WEINSTEIN
Centers for Medicare & Medicaid Services
Baltimore, MD 21244
eweinstein2@cms.hhs.gov

The following question was asked recently in *Runner's World* [**2**], a widely circulated running magazine:

> At the track I use for speed workouts, I'm only allowed to use the outer lanes for training. But the distances of the outer lanes differ from those of the inner lanes. How can I find the exact distances of all the lanes?

This is a common question, since the inner lanes are yielded to faster runners by protocol, to slower runners and walkers who ignore or are ignorant of the protocol, and to the entire high school marching band, which just happens to be practicing on the track during your workout and is not bound by the protocol.

Of the two prominent coaches who answered the question, amazingly, the first went out to his rather standard 400-meter college track and actually measured (he didn't say how) the distances in lane 4 and lane 8 and suggested interpolating for the other lanes, while the second recommended running a lap in each lane while holding the pace steady throughout and noting the time differences [**2**]. Needless to say, these answers leave much to be desired.

A running track (as defined by its lane dividers) is composed of a pair of parallel, aligned straightaways connected at the ends by symmetrical turns. A turn is constructed in one of two ways: a single radius sweeps out concentric semicircles (a *quadrant track*, as in FIGURE 1); alternatively, equal radii sweep out circular arcs adjoining a straightaway from four centers, and these arcs are piecewise smoothly connected to central arcs swept out by another radius of longer length (a *double-bend track*, as in FIGURE 2) [**4**]. The first kind of track is by far the more common; the second looks more squarish and allows for a larger infield area. In either case, for any given track, every lane has the same width, constant (in a normal line sense) throughout the circuit.

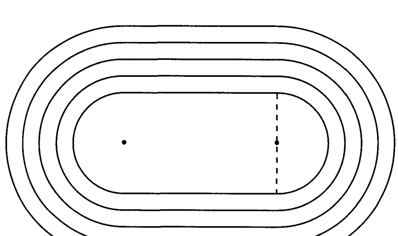


Figure 1 A 4-lane quadrant track

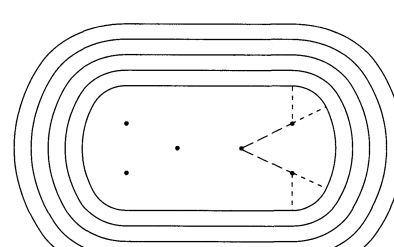


Figure 2 A 4-lane double-bend track

It is known [**3**] that the extra distance traveled around a path everywhere a distance d exterior to a piecewise-smooth simple closed convex curve (such as a lane divider!)

is simply $2\pi d$. Note that this result is independent of the distance around the interior curve. The same result holds more generally for a path exterior to a nonconvex curve, provided the exterior path does not develop any awkward kinks or loop back on itself [**1**]. So it follows from either reference that for a running track with lane width w, the extra distance around the track when running in lane n is simply $2\pi w(n-1)$. Of course, this result is trivially true for a quadrant track, where the two turns taken together form circles.

In practice, the distance around any lane (including lane 1) of a 400-meter track, the outdoor standard nowadays, is actually measured not around the outer edge of the lane's inner boundary as one might expect, but rather around an undrawn curve, called the *measure line*, which is everywhere 20 cm exterior to that outer edge [**4**]. (No doubt it has to do with the fact that one needs two feet with which to run.) Since this just shifts the reference while preserving the lane width, the extra distance sought after will not change.

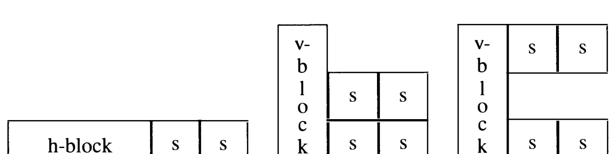
According to international standards, an outdoor track's lane width can vary from 36 to 48 inches [**2**]. Applying the result above gives a range of 5.75–7.66 meters per lane per lap of extra distance. The standard lane width for most U.S. high school and college outdoor tracks is 42 inches [**2**], thus resulting in an extra 6.70 meters per additional lane. So for a track with this lane width, lane 4 has a distance of 420 meters and lane 8 has a distance of 447 meters, rounded to the nearest meter. It turns out the first coach measured lane 4 one meter too short and lane 8 two meters too short. He may well have shortened his work considerably if he simply took the differences in the lanes' staggered starting marks for an appropriate track event. Regardless, he probably measured along the outer edge of a lane's interior divider, inconsistent with the measure line rule. We apply the result yet again and subtract $2\pi(20)$ cm, or 1.26 meters, from our figures, which indeed gives the distance the coach measured for lane 4, but his distance for lane 8 still comes out one meter too short. Apparently, it's safer and easier just to do the math!

REFERENCES

1. R. Bruce Crofoot, Running with Rover, this MAGAZINE **75**:4 (2002), 311–316.
2. *Runner's World*, December 2000, p. 34.
3. J. Sekino, The band around a convex set, *College Math. J.* **32** (2001), 110–114.
4. United States Tennis Court & Track Builders Association Guidelines for Track Construction, 1999, `http://www.ustctba.com/guidelines-track/contents.html`

(continued from page 161)

a block. Thus a vertical block is associated to at most four stamps.



Thus, if we count stamps block by block (plus the extra stamps in the two leftmost columns), the total number is $n^2 \le 2n + 3H + 3V + 2H + 4V = 2n + 5H + 7V \le 2n + 6H + 6V$, giving the desired bound.

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, Western Washington University; PAUL ZEITZ, The University of San Francisco

Proposals

To be considered for publication, solutions should be received by September 1, 2003.

1667. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, AB, Canada.*

Let a, b, and c be nonnegative constants. Determine the maximum and minimum values of

$$\sqrt{a^2x^2+b^2y^2+c^2z^2}+\sqrt{a^2y^2+b^2z^2+c^2x^2}+\sqrt{a^2z^2+b^2x^2+c^2y^2},$$

subject to $x^2+y^2+z^2=1$.

1668. *Proposed by Steve Butler, Provo, UT.*

Let f be a real valued function defined on an open interval I containing $[a, b]$. Assume that f has a continuous second derivative on I and that there is a single line tangent to the graph of $y=f(x)$ at $(a, f(a))$ and $(b, f(b))$. Prove that if $f''(x)$ is not identically zero on (a, b), then $f''(x)$ must change sign at least twice on (a, b).

1669. *Proposed by Ali Nabi Duman (student), Bilkent University, Turkey.*

Let ABC be a triangle and let E be the midpoint of $\overline{BC}$. A circle passing through A and C intersects $\overline{BA}$ and $\overline{BC}$ in points G and E respectively. Let D be the midpoint of $\overline{EC}$. A line through D and perpendicular to $\overline{BC}$ intersects $\overline{AC}$ at F, with $3AF = FC$. Prove that triangle FDG is similar to triangle ABC.

1670. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College of the City University of New York, Bronx, NY.*

Let $n \geq 3$ be an odd integer and let $\{a_1, a_2, \ldots, a_{\phi(n)}\}$ be the set of positive integers less than n and relatively prime to n. Prove that

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. Each solution should begin on a separate sheet.

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 should include the reader's name, full address, and an e-mail address and/or FAX number.

$$\left|\prod_{k=1}^{\phi(n)} \cos\left(\frac{a_k \pi}{n}\right)\right| = \frac{1}{2^{\phi(n)}}.$$

1671. *Proposed by M. N. Deshpande, Institute of Science, Nagpur, India.*

Let $\mathcal{T}$ be the set of triangles ABC for which there is a point D on BC such that segments AB, BD, AD, DC, and AC have integral length and $\angle ACD = \frac{1}{2}\angle ABC = \frac{1}{3}\angle ADB$.

(a) Characterize the sets $\{a, b, c\}$ that are sets of side lengths of triangles in $\mathcal{T}$.

(b) Find the triangle of minimum area in $\mathcal{T}$.

Quickies

Answers to the Quickies are on page 155.

Q929. *Proposed by G. Don Chakerian, University of California, Davis, CA, and Murray S. Klamkin, University of Alberta, Edmonton, AB, Canada.*

Let Q be a convex spherical quadrilateral contained in an open hemisphere. Show that if the opposite angles of Q are equal, then so are the opposite sides.

Q930. *Proposed by Norman Schaumberger (Emeritus), Bronx Community College of the City University of New York, Bronx, NY.*

Let x, y, z be real numbers with $0 < x, y, z < 1$ and $x + y + z = 2$. Prove that

$$x^{1-x}y^{1-y}z^{1-z} + x^{1-y}y^{1-z}z^{1-x} + x^{1-z}y^{1-x}z^{1-y} \leq 2.$$

Solutions

Bounds on a Sequence **April 2002**

1643. *Proposed by Árpád Bényi, University of Kansas, Lawrence, KS, and Ioan Caşu, West University of Timişoara, Timişoara, Romania.*

The sequence $(x_n)_{n\geq 0}$ of nonnegative real numbers satisfies the inequalities

$$x_{n-1}^2 \leq c x_{n-2} x_n, \qquad n \geq 2,$$

where c is a positive constant. Show that for integers n and k, with $0 \leq k \leq n$,

$$x_k \leq c^{k(n-k)/2} x_n^{k/n} x_0^{(n-k)/n}.$$

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

First observe that if $x_s = 0$ for some $s \geq 2$, then $x_{s-1} = 0$, and if $x_s = 0$ for some $s \geq 0$, then $x_{s+1} = 0$. It follows that either $x_k = 0$ for all $k \geq 1$, or $x_k > 0$ for all $k \geq 0$. In the first case it is easy to check that the desired inequalities hold, if we interpret $x^0 = 1$ when $x = 0$. For the second case, define $u_s = x_s/x_{s-1}$ for $s \geq 1$. Because $u_s \leq c u_{s+1}$ for $s \geq 1$, it follows that $u_s \leq c^t u_{s+t}$ for all $s \geq 1$ and $t \geq 0$. Hence,

$$\left(\frac{x_k}{x_0}\right)^{n-k} = \left(\prod_{s=1}^{k} u_s\right)^{n-k} \leq \left(\prod_{s=1}^{k} c^{k+1-s} u_{k+1}\right)^{n-k} = c^{(n-k)k(k+1)/2} u_{k+1}^{k(n-k)}$$

$$\leq c^{(n-k)k(k+1)/2}\left(\prod_{s=0}^{n-k-1} c^s u_{k+1+s}\right)^k = c^{nk(n-k)/2}\left(\frac{x_n}{x_k}\right)^k.$$

The desired inequality follows immediately.

II. *Solution by Knut Dale, Telemark University College, Telemark, Norway.*

As in the first solution, the case in which $x_n = 0$ for $n \geq 1$ is immediate, so we assume that $x_n > 0$ for $n \geq 0$. We define the sequence $(a_n)_{n\geq 0}$ by $x_n = x_0 e^{a_n} c^{-n^2/2}$. The condition $x_{n-1}^2 \leq c x_{n-2} x_n$, $n \geq 2$ is then equivalent to $2a_{n-1} \leq a_{n-2} + a_n$, $n \geq 0$, and the inequality to be proved is equivalent to

$$na_k \leq ka_n, \qquad 0 \leq k \leq n.$$

This inequality is trivial for $k = 0$ and $k = n$. The cases $0 < k < n$ follow from the inequality

$$(k+1)a_k \leq ka_{k+1}, \qquad k \geq 0. \tag{1}$$

This is true for $k = 0$. If inequality (1) holds for a given $k \geq 0$, then we have

$$(k+2)a_{k+1} = 2(k+1)a_{k+1} - ka_{k+1} \leq (k+1)(2a_{k+1} - a_k) \leq (k+1)a_{k+2}.$$

This establishes (1) by induction, and it now follows that for $0 < k < n$,

$$\frac{a_k}{k} \leq \frac{a_{k+1}}{k+1} \leq \cdots \leq \frac{a_n}{n}.$$

This completes the proof.

Also solved by Roy Barbara (Lebanon), Michel Bataille (France), Kenneth Bernstein, Jany C. Binz (Switzerland), Minh Can, Daniele Donini (Italy), Robert L. Doucette, Marty Getz and Dixon Jones, Elias Lampakis (Greece), Rolf Richberg (Germany), Li Zhou, and the proposer.

Functions of Two Variables **April 2002**

1644. *Proposed by Michael Golomb, Purdue University, West Lafayette, IN.*

Assume that the continuous, real valued functions f_i, $i = 1, 2$, are defined on the domain $\mathcal{D} = \{(x, y) : 0 \leq x \leq y \leq 1\}$ and satisfy the following:

(1) $f_i(x, x) = 0, \qquad 0 \leq x \leq 1,$

(2) $f_i(0, x) + f_i(x, 1) = 1, \qquad 0 \leq x \leq 1,$

(3) $f_i(x, y)$ is strictly decreasing in x and strictly increasing in y.

Show that there is a point $(x_0, y_0) \in \mathcal{D}$ such that $f_1(x_0, y_0) = f_2(x_0, y_0) = \frac{1}{2}$.

Solution by McDaniel College Problems Group, McDaniel College, Westminster, MD.

Because $f_i(0, 0) = 0$ and $f_i(0, 1) = 1$, for each i there is a t_i with $0 < t_i < 1$ such that $f_i(0, t_i) = f_i(t_i, 1) = 1/2$. If $t_1 = t_2$, then $(x_0, y_0) = (0, t_1)$ satisfies the desired conditions. If $t_1 \neq t_2$, then we may assume that $t_1 < t_2$. By condition (3), we have $f_i(0, x) < 1/2$, and hence $f_i(x, 1) > 1/2$, for $0 \leq x < t_i$. Thus, if $0 \leq x < t_1$, the vertical segment from (x, x) to $(x, 1)$ contains exactly one point $(x, g_i(x))$ such that $f_i(x, g_i(x)) = 1/2$. In particular, $g_1(0) = t_1 < t_2 = g_2(0)$. Because $f_2(t_1, 1) > f_2(t_2, 1)$, it follows that $g_2(t_1) < 1 = g_1(t_1)$. Let $A = \{x : 0 < x < t_1$ and $g_1(x) < g_2(x)\}$ and set $\alpha = \sup(A)$. Because f_1 and f_2 are continuous, $0 < \alpha < t_1$. If $g_1(\alpha) < g_2(\alpha)$, then by the continuity of f_1 and f_2, there is a $t > \alpha$ with $g_1(t) < g_2(t)$. This is impossible. If $g_1(\alpha) > g_2(\alpha)$, then there is an $\epsilon > 0$ such that $g_1(x) > g_2(x)$ for

$\alpha - \epsilon < x < \alpha$, contradicting the fact that $\alpha = \sup(A)$. Thus it must be the case that $g_1(\alpha) = g_2(\alpha)$, so $(x_0, y_0) = (\alpha, g_1(\alpha))$ is a point with $f_1(x_0, y_0) = f_2(x_0, y_0) = \frac{1}{2}$.

Also solved by Roy Barbara (Lebanon), Michel Bataille (France), John Christopher, Daniele Donini (Italy), Knut Dale (Norway), Robert L. Doucette, Brian D. Ginsberg, Tom Jager, Elias Lampakis (Greece), James M. Meehan, Stephen Noltie, Rolf Richberg (Germany), Ralph Rush, Jawad Sadek, Achilleas Sinefakopoulos, John W. Spellmann and Sam H. Creswell, Nora Thornber, Dave Trautman, Daniel G. Treat, Paula Grafton Young, Li Zhou, and the proposer.

Disconnected Magic **April 2002**

1645. *Proposed by Philip Straffin, Stephen Goodloe, and Tamas Varga, Beloit College, Beloit, WI.*

A graph is called *magic* if it has $n \geq 1$ edges and its edges can be labeled by the integers $1, 2, \ldots, n$ with each integer used once, and so that the sum of the labels of the edges at any vertex is the same. Are there any magic graphs which are not connected?

Solution by Daniele Donini, Bertinoro, Italy.

We show that for any integer $m \geq 1$, there exists a magic graph with m components. One such graph is the graph

$$G_m = K_{4,4} \cup \cdots \cup K_{4,4},$$

consisting of m copies of the complete bipartite graph $K_{4,4}$. We use magic squares to generate a magic numbering scheme for the edges of G_m.

Write the numbers from 1 to $16m$ in the cells of a $4 \times 4m$ grid, as illustrated in Table 1. Starting from the left, partition the grid into m 4×4 grids. Replace any diagonal element x in one of these grids by its "complement" $16m - x + 1$. See Table 2. This process performs interchanges between selected pairs of numbers in positions that are symmetric about the center point of the grid, so the numbers in the resulting table are still the numbers 1 through $16m$. In addition, each 4×4 subgrid is now a magic square in which the elements in each row and each column sum to $32m + 2$.

TABLE 1:

1	2	3	4	5	6	7	8	$\cdots$	$4m-3$	$4m-2$	$4m-1$	$4m$
$4m+1$	$4m+2$	$4m+3$	$4m+4$	$4m+5$	$4m+6$	$4m+7$	$4m+8$	$\cdots$	$8m-3$	$8m-2$	$8m-1$	$8m$
$8m+1$	$8m+2$	$8m+3$	$8m+4$	$8m+5$	$8m+6$	$8m+7$	$8m+8$	$\cdots$	$12m-3$	$12m-2$	$12m-1$	$12m$
$12m+1$	$12m+2$	$12m+3$	$12m+4$	$12m+5$	$12m+6$	$12m+7$	$12m+8$	$\cdots$	$16m-3$	$16m-2$	$16m-1$	$16m$

TABLE 2:

$16m$	2	3	$16m-3$	$16m-4$	6	7	$16m-7$	$\cdots$	$12m+4$	$4m-2$	$4m-1$	$12m+1$
$4m+1$	$12m-1$	$12m-2$	$4m+4$	$4m+5$	$12m-5$	$12m-6$	$4m+8$	$\cdots$	$8m-3$	$8m+3$	$8m+2$	$8m$
$8m+1$	$8m-1$	$8m-2$	$8m+4$	$8m+5$	$8m-5$	$8m-6$	$8m+8$	$\cdots$	$12m-3$	$4m+3$	$4m+2$	$12m$
$4m$	$12m+2$	$12m+3$	$4m-3$	$4m-4$	$12m+6$	$12m+7$	$4m-7$	$\cdots$	4	$16m-2$	$16m-1$	1

For $1 \leq k \leq m$, the edge labels for the k-th component of G_m are determined by the k-th 4×4 grid,

$$T_k = \begin{array}{|c|c|c|c|} \hline 16m-4k+4 & 4k-2 & 4k-1 & 16m-4k+1 \\ \hline 4m+4k-3 & 12m-4k+3 & 12m-4k+2 & 4m+4k \\ \hline 8m+4k-3 & 8m-4k+3 & 8m-4k+2 & 8m+4k \\ \hline 4m-4k+4 & 12m+4k-2 & 12m+4k-1 & 4m-4k+1 \\ \hline \end{array}$$

The corresponding component $K_{4,4}$ has vertices $r_{k1}, r_{k2}, r_{k3}, r_{k4}, c_{1k}, c_{2k}, c_{3k}, c_{4k}$ corresponding to the rows and columns of T_k, and edge set $\{r_{ki}c_{jk} : 1 \leq i, j \leq 4\}$. If we assign to edge $r_{ki}c_{jk}$ the entry in row i and column j of T_k, then the sum of the edges at each vertex corresponds to a row or column sum from T_k. It follows that the resulting graph G_m is magic, with the edge labels at each vertex summing to $32m + 2$.

Note. Many readers submitted simple examples of nonconnected magic multigraphs, that is, graphs in which at least one pair of vertices is joined by more than one edge.

Also solved by The Carroll College Problem Solving Group, Eddie Cheng, Marty Getz and Dixon Jones, Khudija S. Jamil, and the proposers.

Power Sums in a Sequence **April 2002**

1646. *Proposed by Erwin Just (Emeritus), Bronx Community College, Bronx, NY.*

Let $a > 0$, b, $k > 0$, and $m > 0$ be integers, and assume that the arithmetic progression $\{an + b\}_{n=0}^{\infty}$ contains the kth power of an integer. Prove that there are an infinite number of values of n for which $an + b$ is the sum of m k-th powers of nonzero integers.

Solution by Kenneth Bernstein, Belmont, MA.

Let $an_0 + b = c^k$, and let $c_1, c_2, \ldots, c_{m-1}$ be any nonnegative integers. Then for $n = n_0 + a^{k-1}(c_1^k + c_2^k + \cdots + c_{m-1}^k)$, $an + b$ is the sum of m k-th powers of nonzero integers.

Also solved by Roy Barbara (Lebanon), Jany C. Binz (Switzerland), Knut Dale (Norway), Daniele Donini (Italy), Ovidiu Furdui, Brian D. Ginsberg, Tom Jager, Lenny Jones, Elias Lampakis (Greece), Peter W. Lindstrom, Rolf Richberg (Germany), Li Zhou, and the proposer.

Answers

Solutions to the Quickies from page 152.

A929. Assume that Q lies on a sphere of radius 1 and has sides a_i and angles α_i, $i = 1, 2, 3, 4$. The spherical supplement of Q, denoted by Q^*, has sides a_i^* and angles α_i^* satisfying $a_i + \alpha_i^* = a_i^* + \alpha_i = \pi$, $1 \leq i \leq 4$. (See, for example, George Polya's Mathematics and Plausible Reasoning, Volume 1: Induction and Analogy in Mathematics, Princeton University Press, 1954, p. 57.) Because $a_i^* = \pi - \alpha_i$, it follows that if the opposite angles of Q are equal, then the opposite sides of Q^* are equal. Next consider the two triangles formed from Q^* by drawing one of the diagonals of Q^*. By the SSS congruence theorem for spherical triangles, the two triangles are congruent. It follows that opposite angles in Q^* are equal, and then that the opposite sides of Q are equal. (Additional solution on p. 106.)

A930. Observe that $(1 - x) + (1 - y) + (1 - z) = 1$. By the weighted arithmetic/geometric mean inequality,

$$\begin{aligned} x^{1-x}y^{1-y}z^{1-z} &\leq x(1-x) + y(1-y) + z(1-z) \\ x^{1-y}y^{1-z}z^{1-x} &\leq x(1-y) + y(1-z) + z(1-x) \\ x^{1-z}y^{1-x}z^{1-y} &\leq x(1-z) + y(1-x) + z(1-y). \end{aligned}$$

Adding these inequalities gives the desired result.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles, books, and other materials 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.

Arney, Chris, and Donald Small (eds.), *Changing Core Mathematics*, MAA, 2002; xi + 183 pp, \$28.95 (P) (\$22.95 to members). ISBN 0–88385–172–5.

This volume was sent to mathematics departments throughout the country to stimulate change in the first two years of college mathematics for mathematics, science, and engineering students. For years, that core has been dominated by calculus. The authors suggest that the development and application of information technology engender the need for a "revolutionary" change, to focus on problem-solving skills and "learning how to learn" through structuring core mathematics around "modeling and inquiry"—in other words, around process rather than content. The first part of the volume surveys the history of core mathematics, makes the case for inquiry and modeling, and suggests an integrated curriculum for the first year. The first semester would begin with graph-theoretic models, proceed through probabilistic models and matrices (including eigenvalues and eigenvectors), concentrate on discrete dynamical systems, and conclude with an introduction to continuous change. The second semester would introduce differential equations as antiderivative problems, consider Euler's method and numerical integration, and concentrate on differential equations. The bulk of the volume consists of diverse essays by workshop participants; and an appendix contains three of COMAP's Interdisciplinary Lively Applications Projects (ILAPs), which are problem-solving projects for teams of students. The perspective of this book and its proposed reorientation of mathematics for science-oriented students has much to recommend it. However, not all students who study calculus are science-inclined or even science-interested; and faculty who like to teach mathematics from a theorem-proof viewpoint or as an intellectual endeavor in historical context would feel left out in the proposed curriculum. Unfortunately, the theme of interdisciplinarity might hit a roadblock with both students and mathematics faculty who are interested in mathematics for its own sake: They may have no interest in (nor any experience with) models in statics, electrical circuits, and fluid dynamics. In fact, older faculty members may remember that many of the applications in one edition of a calculus book by Finney and Thomas were cut out of the next because of exactly that kind of resistance (plus—crucially—lack of knowledge about applications on the part of teaching assistants!). Finally, some of the pedagogical recommendations, such as "assign new material for students to prepare before it is discussed in class," would likely meet strong resistance from students in all areas.

Sossinsky, Alexei, *Knots: Mathematics with a Twist*, Harvard University Press, 2002; xix + 126 pp, \$24.95. ISBN 0–674–00944–4.

This gem of a book organizes an exposition of knot theory into chapters that each start with a simple original idea and explore its implications. The order is largely chronological, and technical details are minimized, though this book demands more of the reader than the Wilson book on the four-color problem (see below). Unlike that book, the endnotes here are not keyed to the references and—my pet peeve!—there is no index.

Wilson, Robin, *Four Colors Suffice: How the Map Problem Was Solved*, Princeton University Press, 2002; xiv + 262 pp, $24.95. ISBN 0–691–11533–8.

"The publication of this book coincides with the 150th anniversary of the four-colour problem, and the 25th anniversary of the publication of its proof." Author Wilson provides a delightful and well-organized history of the four-color problem and its solution by Haken and Appel, requiring no mathematical background and clothed in a beautifully laid-out book. The philosophical objections to a computer-assisted proof may have died down in the past 25 years. However, Wilson relates that investigators who tried in the 1990s to check the Haken-Appel proof gave up and instead created a new and simpler proof of their own, along the lines of the original proof. (Curiously, the only color in this book is on the map on the dust cover; all the book's illustrations are in halftone.)

Krantz, Steven G., *Mathematical Apocrypha: Stories and Anecdotes of Mathematicians and the Mathematical*, MAA, 2002; xiii + 214 pp (P), $32.95 ($25.95 for members). ISBN 0-88385–539–9.

You will vastly enjoy dipping into this collection of stories about mathematicians, even though some may not be verifiable (or even true). Despite author Krantz's assertion that he has avoided stories that are "mean-spirited or critical or that depict people in a bad light," an entire section is devoted to "mathematical foolishness" and in some cases names are omitted to protect the individuals involved from embarrassment. Oh, well, enjoy yourself nevertheless—just be sure to keep it all in the family.

dePillis, John, *777 Mathematical Conversation Starters*, MAA, 2002; xvi + 344 pp, $37.95 (P) ($29.95 to members). ISBN 0–88385–540–2.

Beginning a conversation by mentioning that you are a mathematician is usually a non-starter, you are probably too honest to lie and say instead that you are a tennis coach, and the anecdotes of *Mathematical Apocrypha* (see above) are mostly "in" jokes meaningful only to mathematicians. So, to enhance your social life beyond a small circle of mathematician friends, you need "conversation starters" that are appropriate for interaction with the vast majority of people, who do not recognize the name of any mathematician since Euclid. Here may be the answer to your prayers: a cartoon-illustrated collection of thought-provoking quotations and brief discussions, arranged alphabetically by topic. Many are familiar, but some will be new to you. Sample, by Stanley Osher, UCLA: "I write the algorithms that make the computer sing. I'm the Barry Manilow of mathematics."

Hayes, Brian, Science on the far side, *American Scientist* (November-December 2002) 499–502; Science on the farther shore, `http://www.americanscientist.org/Issues/Comsci02/02-11Hayes.html`.

The priority for invention of the method of least-squares goes to Gauss, who used it to help re-find the asteroid Ceres in 1801. His diary records his earlier discovery of the method, which he said he had used since 1795. He published only in 1809, four years after Legendre had published it as an orbit-finding technique (without justification or mention of probability). Legendre: "There is no discovery that one cannot claim for oneself by saying that one had found the same thing some years previously." At about the same time as Gauss, "a citizen of a developing country" far away, the Irish-born American Robert Adrain (1775–1843), published an account of least-squares and of the normal distribution, in connection with a surveying problem. Adrain published in an American journal that died shortly thereafter—did any copies ever reach England, much less the Continent? Adrain acquired Legendre's 1805 book, but we do not know whether before or after writing his paper. Author Hayes speculates on the chance of a movie about Adrain (theme: backwoods bumpkin beats out whining wig-wearers), notes the difference that today's technology makes for scientists far from the centers of science, and makes the important observation that "It takes more than a village to raise a scientist. It takes a village full of scientists." [Adrain's paper was reprinted in Stephen M. Stigler (ed.), *American Contributions to Mathematical Statistics in the Nineteenth Century*; New York: Arno Press, 1980.]

NEWS AND LETTERS

31st United States of America Mathematical Olympiad

May 3 and May 4, 2002

edited by Titu Andreescu and Zuming Feng

Problems

1. Let S be a set with 2002 elements, and let N be an integer with $0 \le N \le 2^{2002}$. Prove that it is possible to color every subset of S either blue or red so that the following conditions hold:
   (a) the union of any two red subsets is red;
   (b) the union of any two blue subsets is blue;
   (c) there are exactly N red subsets.
2. Let ABC be a triangle such that
$$\left(\cot\frac{A}{2}\right)^2 + \left(2\cot\frac{B}{2}\right)^2 + \left(3\cot\frac{C}{2}\right)^2 = \left(\frac{6s}{7r}\right)^2,$$
where s and r denote its semiperimeter and its inradius, respectively. Prove that triangle ABC is similar to a triangle T whose side lengths are all positive integers with no common divisor and determine these integers.
3. Prove that any monic polynomial (a polynomial with leading coefficient 1) of degree n with real coefficients is the average of two monic polynomials of degree n with n real roots.
4. Let $\mathbb{R}$ be the set of real numbers. Determine all functions $f : \mathbb{R} \to \mathbb{R}$ such that
$$f(x^2 - y^2) = xf(x) - yf(y)$$
for all pairs of real numbers x and y.
5. Let a, b be integers greater than 2. Prove that there exists a positive integer k and a finite sequence $n_1, n_2, \ldots, n_k$ of positive integers such that $n_1 = a$, $n_k = b$, and $n_i n_{i+1}$ is divisible by $n_i + n_{i+1}$ for each i $(1 \le i < k)$.
6. I have an $n \times n$ sheet of stamps, from which I've been asked to tear out blocks of three adjacent stamps in a single row or column. (I can only tear along the perforations separating adjacent stamps, and each block must come out of a sheet in one piece.) Let $b(n)$ be the smallest number of blocks I can tear out and make it impossible to tear out any more blocks. Prove that there are constants c and d such that
$$\frac{1}{7}n^2 - cn \le b(n) \le \frac{1}{5}n^2 + dn$$
for all $n > 0$.

Solutions

Note: For interested readers, the editors recommend the *USA and International Mathematical Olympiads 2002*. There many Olympiad problems are presented, together with a collection of remarkable solutions developed by the examination committees, contestants, and experts, during or after the contests.

1. If $N = 0$, we color every subset blue; if $N = 2^{2002}$, we color every subset red. Now suppose neither of these holds. We may assume that $S = \{0, 1, 2, \ldots, 2001\}$. Write N in binary representation: $N = 2^{a_1} + 2^{a_2} + \cdots + 2^{a_k}$, where the a_i are all distinct; then each a_i is an element of S. Color each a_i red, and color all the other elements of S blue. Now declare each nonempty subset of S to be the color of its largest element, and color the empty subset blue. If T, U are any two nonempty subsets of S, then the largest element of $T \cup U$ equals the largest element of T or the largest element of U, and if T is empty, then $T \cup U = U$. It readily follows that (a) and (b) are satisfied. To verify (c), notice that, for each i, there are 2^{a_i} subsets of S whose largest element is a_i (obtained by taking a_i in combination with any of the elements $0, 1, \ldots, a_i - 1$). If we sum over all i, each red subset is counted exactly once, and we get $2^{a_1} + 2^{a_2} + \cdots + 2^{a_k} = N$ red subsets.
2. Define $a = BC$, $b = CA$, $c = AB$, and $u = \cot A/2$, $v = \cot B/2$, $w = \cot C/2$. Denote by I the incenter of triangle ABC, and let D, E, F be the points of tangency of the incircle with sides BC, CA, AB, respectively. Then $EI = r$, and by standard results about incircles, $AE = s - a$. So $u = \cot A/2 = AE/EI = s - a/r$, and similarly $v = s - b/r$, $w = s - c/r$. Because
$$\frac{s}{r} = \frac{(s-a) + (s-b) + (s-c)}{r} = u + v + w,$$
we can rewrite the given relation as $49[u^2 + 4v^2 + 9w^2] = 36(u + v + w)^2$, which is the equality case of the Cauchy-Schwarz Inequality
$$(6^2 + 3^2 + 2^2)\left[u^2 + (2v)^2 + (3w)^2\right] \geq (6 \cdot u + 3 \cdot 2v + 2 \cdot 3w)^2.$$
After multiplying by r, we see that
$$\begin{aligned}\frac{s-a}{36} &= \frac{s-b}{9} = \frac{s-c}{4}\\ &= \frac{2s-b-c}{9+4} = \frac{2s-c-a}{4+36} = \frac{2s-a-b}{36+9} = \frac{a}{13} = \frac{b}{40} = \frac{c}{45},\end{aligned}$$
that is, triangle ABC is similar to a triangle with side lengths 13, 40, 45.
3. Let $F(x)$ be the monic real polynomial of degree n. If $n = 1$, then $F(x) = x + a$ for some real number a. Then $F(x)$ is the average of $x + 2a$ and x, each of which has 1 real root. Now we assume that $n > 1$. Define the polynomial $G(x) = (x-2)(x-4)\cdots(x - 2(n-1))$. The degree of $G(x)$ is $n - 1$. Consider the polynomials $P(x) = x^n - kG(x)$ and $Q(x) = 2F(x) - P(x) = 2F(x) - x^n + kG(x)$. We will show that for large enough k these two polynomials have n real roots. Since they are monic and their average is clearly $F(x)$, this will solve the problem.

   Consider the values of polynomial $G(x)$ at n points $x = 1, 3, 5, \ldots, 2n - 1$. These values alternate in sign and have magnitude at least 1 (since at most two of the factors have magnitude 1 and the others have magnitude at least 2). On the other hand, there is a constant $c > 0$ such that for $0 \leq x \leq n$, we have $|x^n| < c$ and $|2F(x) - x^n| < c$. Take $k > c$. Then we see that $P(x)$ and $Q(x)$ evaluated at n points $x = 1, 3, 5, \ldots, 2n-1$ alternate in sign. Thus, polynomials $P(x)$ and $Q(x)$ each have at least $n-1$ real roots—one in each interval $(1, 3), \ldots, (2n-3, 2n-1)$. However, since they are polynomials of degree n, they must then each have n real roots (as in the previous solution), as desired.

4. Setting $x = y = 0$ in the given condition yields $f(0) = 0$. Because $-xf(-x) - yf(y) = f([-x]^2 - y^2) = f(x^2 - y^2) = xf(x) - yf(y)$, we have $f(-x) = -f(x)$ for $x \neq 0$. Hence $f(x)$ is odd. From now on, we assume $x, y \geq 0$.

   Setting $y = 0$ in the given condition yields $f(x^2) = xf(x)$. Hence $f(x^2 - y^2) = f(x^2) - f(y^2)$, or, $f(x^2) = f(x^2 - y^2) + f(y^2)$. Since for $x \geq 0$ there is a unique $t \geq 0$ such that $t^2 = x$, it follows that

   $$f(x) = f(x-y) + f(y) \tag{1}$$

   Setting $x = 2t$ and $y = t$ in (1) gives

   $$f(2t) = 2f(t). \tag{2}$$

   Setting $x = t+1$ and $y = t$ in the given condition yields

   $$f(2t+1) = (t+1)f(t+1) - tf(t). \tag{3}$$

   By (2) and by setting $x = 2t+1$ and $y = 1$ in (1), the left-hand side of (3) becomes

   $$f(2t+1) = f(2t) + f(1) = 2f(t) + f(1). \tag{4}$$

   On the other hand, by setting $x = t+1$ and $y = 1$ in (1), the right-hand side of (3) reads $(t+1)f(t+1) - tf(t) = (t+1)[f(t) + f(1)] - tf(t)$, or,

   $$(t+1)f(t+1) - tf(t) = f(t) + (t+1)f(1). \tag{5}$$

   Putting (3), (4), and (5) together leads to $2f(t) + f(1) = f(t) + (t+1)f(1)$, or, $f(t) = tf(1)$ for $t \geq 0$. Recall that $f(x)$ is odd; we conclude that $f(-t) = -f(t) = -tf(1)$ for $t \geq 0$. Hence $f(x) = kx$ for all x, where $k = f(1)$ is a constant. It is not difficult to see that all such functions indeed satisfy the conditions of the problem.

5. We may say two positive integers a and b are *connected*, denoted by $a \leftrightarrow b$, if there exists a positive integer k and a finite sequence $n_1, n_2, \ldots, n_k$ of positive integers such that $n_1 = a$, $n_k = b$, and $n_i n_{i+1}$ is divisible by $n_i + n_{i+1}$ for each i $(1 \leq i < k)$. The problem asks to prove that $a \leftrightarrow b$ for all $a, b > 2$. Note that for positive integer n with $n \geq 3$, $n \leftrightarrow 2n$, as the sequence $n \leftrightarrow n(n-1) \leftrightarrow n(n-1)(n-2) \leftrightarrow n(n-2) \leftrightarrow 2n$ satisfies the conditions of the problem. For positive integer $n \geq 4$, $n' = (n-1)(n-2) \geq 3$, hence $n' \leftrightarrow 2n'$ by the above argument. It follows that $n \leftrightarrow n-1$ for $n \geq 4$ by $n' \leftrightarrow 2n'$ and by the sequences $n \leftrightarrow n(n-1) \leftrightarrow n(n-1)(n-2) \leftrightarrow n(n-1)(n-2)(n-3) \leftrightarrow 2(n-1)(n-2) \leftrightarrow (n-1)(n-2) \leftrightarrow n-1$. Iterating this, we connect all integers larger than 2.

6. The upper bound requires an example of a set of $n^2/5 + dn$ blocks whose removal makes it impossible to remove any further blocks. We note first that we can tile the plane with tiles that contain one block for every five stamps, so that no more blocks can be chosen. Two such tilings are shown below with one tile outlined in heavy lines. Assume that there are x unit squares in each tile. Then there are $x/5$ blocks in each tile. Choose a constant m such that the basic tile fits inside an $(m+1) \times (m+1)$ square. Given an $n \times n$ section of the tiling, take all tiles lying entirely within that section and add as many additional tiles, which lie partially in and partially out of that section, as possible. Let k denote the total number of

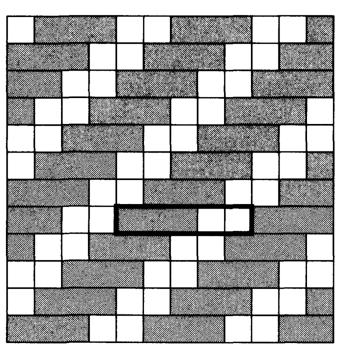
chosen tiles. Hence there are $kx/5$ blocks contained in the k chosen tiles. The $n \times n$ section is covered by all the chosen tiles, and these are all contained in a concentric $(n + 2m) \times (n + 2m)$ square. Then $kx \leq (n + 2m)^2$, and so there are at most

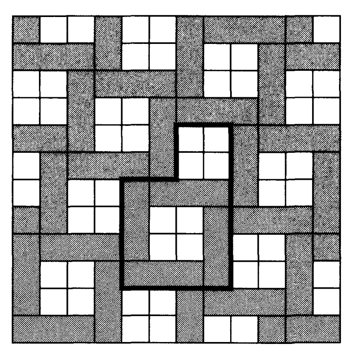
$$\frac{1}{5}kx \leq \frac{1}{5}(n + 2m)^2 \leq \frac{1}{5}n^2 + \frac{4m^2 + 4m}{5}n$$

blocks total. We can classify all the above blocks into three categories (i) blocks lying completely in the $n \times n$ section; (ii) blocks lying partially in the section; (iii) blocks lying completely outside of the section. Suppose there are x_1, x_2, x_3 blocks in categories (i), (ii), (iii), respectively. We do not have to worry about blocks in category (iii), and we take all the blocks in category (i). We need to deal with blocks in category (ii) with more care. By the conditions of the problem, we can not take out those blocks from the $n \times n$ section. All the blocks in category (ii) are on the border of the section. Hence there are at most $4n$ blocks in category (ii), and so these blocks contain at most $8n$ stamps in the $n \times n$ square. We might need additional blocks to deal with these stamps. Each of the additional blocks must contain one of these stamps. Thus there are at most $8n$ additional blocks. Thus there are at most

$$x_1 + 8n \leq x_1 + x_2 + x_3 + 8n \leq \frac{1}{5}n^2 + \frac{4m^2 + 4m + 40}{5}n$$

blocks needed.





The lower bound requires an argument. In fact, we'll prove the lower bound

$$b(n) \geq \frac{1}{6}(n^2 - 2n).$$

Each block can be classified as "horizontal" or "vertical" in the obvious fashion. Given an arrangement of blocks, let H and V be the numbers of horizontal and vertical blocks. Without loss of generality, we may assume $V \leq H$.

We associate each unused stamp that is not in one of the two leftmost columns to the first block one encounters proceeding leftward from the stamp. Note that one never has to proceed leftward more than two stamps; otherwise, there would be another block to remove. Each block is associated to at most two stamps in each row that it occupies. In particular, each horizontal block is associated to at most two stamps. Moreover, a vertical block cannot have an unused stamp on its immediate right in each of the three rows it covers; otherwise, those three stamps would form

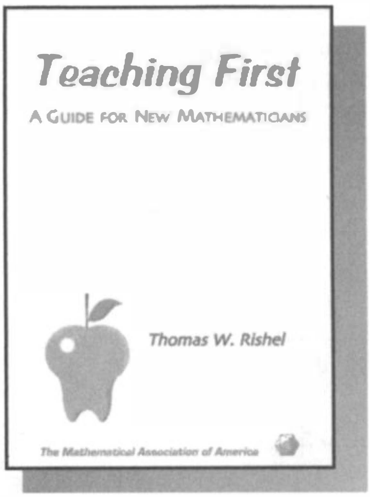
(continued on page 150)

The Mathematical Association of America

Teaching First: A Guide for New Mathematicians

Thomas W. Rishel

Series: MAA Notes



In this volume Thomas Rishel draws on his nearly forty years of teaching experience to address the "nuts and bolts" issues of teaching college mathematics. This book is written for the mathematics TA or young faculty member who may be wondering just where and how to start. Rishel opens the readers' eyes to pitfalls they may never have considered, and offers advice for balancing an obligation "to the student" with an obligation "to mathematics." Throughout, he provides answers to seemingly daunting questions shared by most new TAs, such as how to keep a classroom active and lively; how to prepare writing assignments, tests, and quizzes; how exactly to write a letter of recommendation; and how to pace, minute by minute, the "mathematical talks" one will be called upon to give.

This book is Rishel's answer to those who may suggest that good teaching is innate and cannot be taught. This he emphatically denies, and he insists that solid teaching starts with often overlooked "seeming trivialites" that one needs to master before exploring theories of learning. Along the way he also covers the general issues that teachers of all subjects eventually experience: fairness in grading, professionalism among students and colleagues, identifying and understanding student "types", technology in the classroom. All of the subjects in this book are considered within the context of Rishel's experience as a mathematics teacher. All are illustrated with anecdotes and suggestions specific to the teaching of mathematics.

Teaching First is a comprehensive guide for a mathematics TA, from the first semester preparations through the unforseen challenges of accepting a faculty position. Its aim is to prepare the new TA with clear suggestions for rapidly improving their teaching abilities.

Catalog Code: NTE-54/JR 150 pp., Paperbound, 2000 ISBN 088385-165-2 List: $19.00 MAA Member: $15.00

Name______________________ Credit Card No.______________________

Address______________________ Signature______________________ Exp. Date____/____

City______________________ Qty________ Price $__________ Amount $__________

State__________ Zip __________ Shipping and Handling $__________

Phone______________________ Catalog Code: NTE-54/JR Total $__________

Shipping and Handling: USA orders (shipped via UPS): $3.00 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 2-3 weeks of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas Orders:** $4.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $10.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit card orders are accepted for all customers. All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Order Via:

Phone: 1 (800) 331.1622

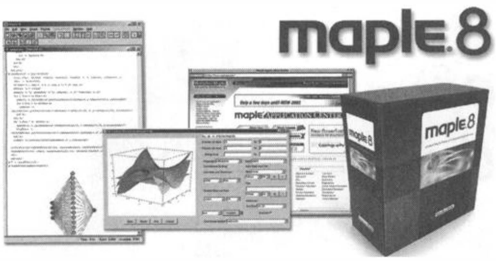
Fax: (301) 206.9789

Mail: Mathematical Association of America
PO Box 91112
Washington, DC 20090-1112

Web: www.maa.org

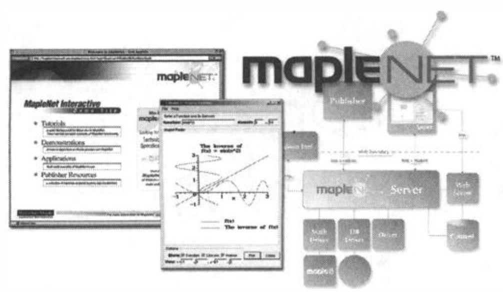
Distance • Asynchronous • Online • On Campus

Education is no longer just about classrooms and labs. With the growing diversity and complexity of educational programs, you need a software system that lets you efficiently deliver effective learning tools to literally, the world. Maple® now offers you a choice to address the reality of today's mathematics education.



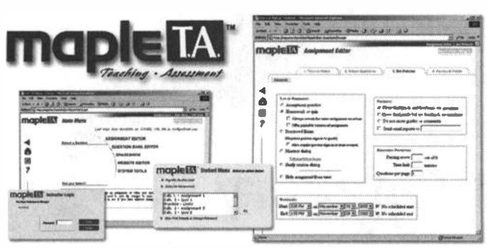
Maple 8 – the standard

Perfect for students in mathematics, sciences, and engineering. Maple 8 offers all the power, flexibility, and resources your technical students need to manage even the most complex mathematical concepts.



MapleNET – online education

A complete standards-based solution for authoring, delivering, and managing interactive learning modules through browsers. Derived from the legendary Maple engine, MapleNet is the only comprehensive solution for Web based education in mathematics.



Maple T.A. – automated assessment

Online math skills assessment without resorting to multiple choice or true or false. MapleT.A. supports natural entry and intelligent testing of answers. A complete question-set management infrastructure lets you manage even the largest class sizes.

Give your institution and your students the competitive edge.

For a **FREE 30-day Maple 8 Trial CD**, or to register for a **FREE MapleNet or Maple T.A. Online Seminar** call ***1.800.267.6583*** or e-mail ***info@maplesoft.com***.

Maplesoft
command the brilliance

WWW.MAPLESOFT.COM | INFO@MAPLESOFT.COM | WWW.MAPLEAPPS.COM | US & CANADA 800.267.6583

© Maplesoft, a division of Waterloo Maple Inc. 2003. Maplesoft, Maple, MapleNet and Maple T.A. are trademark of Waterloo Maple Inc. All other trademarks are property of their respective owners.

Use the
MAA's
Catalog of
Commercial Products
to Find

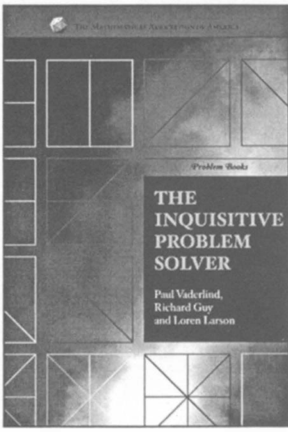
Thousands Of Books For You To Consider For Your Mathematics Courses.

Browse By Subject Category at:

www.MathDL.org/lcp.html

Part of the MAA's Mathematical Sciences Digital Library	*The Mathematical Association of America*
www.MathDL.org	**www.MAA.org**

NEW! from the Mathematical Association of America

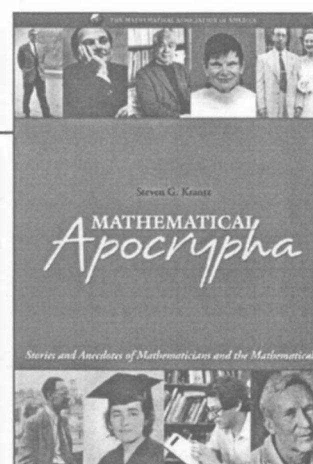


The Inquisitive Problem Solver

Vaderlind, Guy, & Larson

The Inquisitive Problem Solver is a collection of 256 mathematical miniatures composed to stimulate and entertain. However, on a deeper level, these little puzzles, accessible to a general audience, provide a setting rich in mathematical themes. One of the larger purposes of the book is to show how everyday situations can lead an inquisitive problem solver to profound and far-reaching mathematical principles. Discussions accompanying the problems reinforce important techniques in discrete mathematics, and the solutions—which require verbal arguments—show that proofs and careful reasoning are at the core of doing mathematics. In addition, anyone reading this book will learn that asking good questions is just as important to the progress of mathematics as answering questions.

Catalog Code: IPS/JR • 344 pp., Paperbound, 2002 • ISBN 0-88385-806-1
List: $34.95 • MAA Member: $24.95



Mathematical Apocrypha

Krantz

Mathematical Apocrypha is a book of stories about mathematicians and the mathematical. It differs from other books of its kind in that it includes many stories about contemporary mathematicians. Many of these stories are derived from the author's direct or second-hand experience, and have never before appeared in print. The stories are told in a brisk and engaging style, and are enhanced by numerous photographs and illustrations. The theme of the book is strictly mathematical. Some of the stories, however, are about people who adhere to mathematics but cannot strictly be called mathematicians. Included are stories about Bertrand Russell, Alfred North Whitehead, and Albert Einstein, along with stories about mathematicians Erdös, Doob, Besicovitch, Atiyah, Wiener, Mary Ellen and Walter Rudin, Pólya, Halmos, Littlewood and many, many more legendary mathematicians.

Catalog Code: APC/JR • 280 pp., Paperbound, 2002 • ISBN 0-88385-539-9
List: $28.95 • MAA Member: $22.95

Call 1-800-331-1622 to order!

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

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