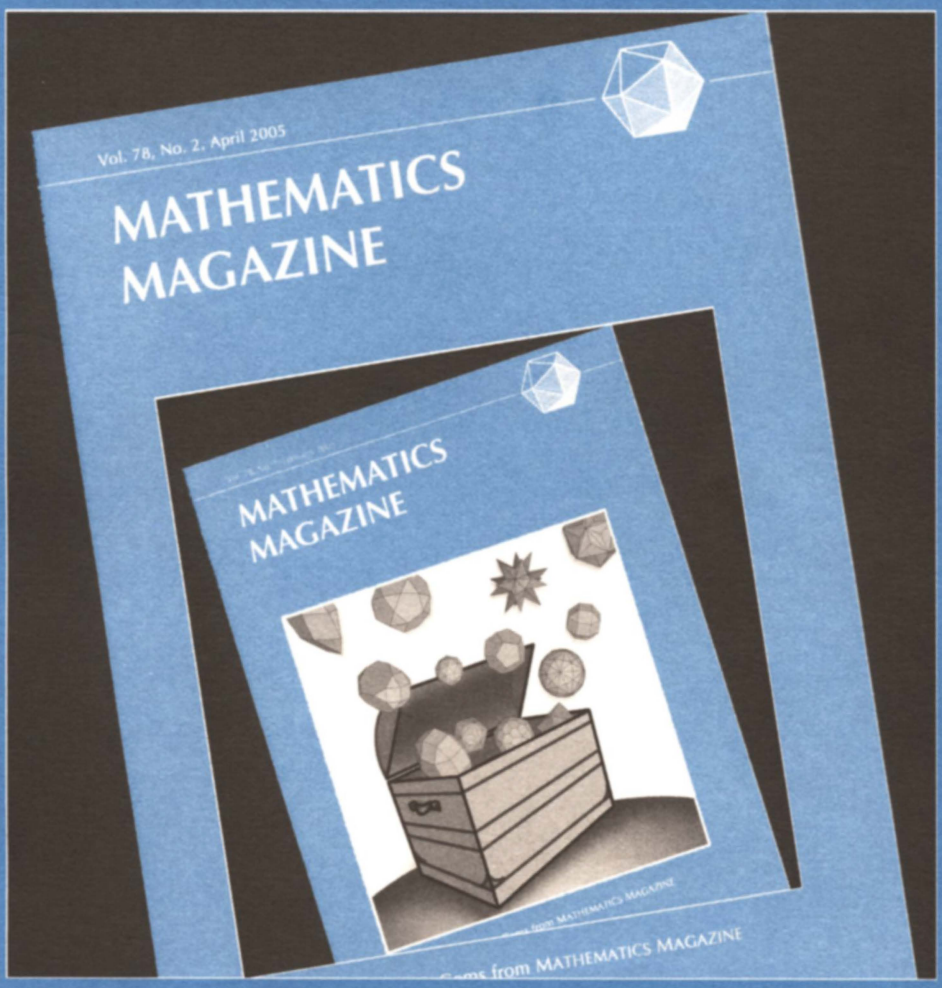
Vol. 78, No. 2, April 2005

MATHEMATICS MAGAZINE



Gems from MATHEMATICS MAGAZINE

- Groups of Arithmetical Functions
- Outwitting the Lying Oracle
- Twentieth-Century Gems from MATHEMATICS MAGAZINE

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 Allen Schwenk, Editor-Elect, Mathematics Magazine, Department of Mathematics, Western Michigan University, Kalamazoo, MI, 49008. 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: *Gems from Gems from Mathematics Magazine from Mathematics Magazine*, by Jason Challas. Jason teaches at West Valley College in Saratoga, CA where he instructs Jason to instruct Jason to instruct art.

AUTHORS

Jim Delany graduated from San Diego State College in 1961, then studied under Tom Head at Iowa State University, where he wrote a dissertation in lattice theory. Since 1970, he has been at Cal Poly, teaching part-time since retiring in 2001. He enjoys number theory, algebra, and trying to solve problems published monthly in MAA journals. For years he was an avid long-distance cyclist, and has fond memories of riding the Paris-Brest-Paris *randonné* in 1983.

Robb T. Koether received his B.S. from the University of Richmond in 1973 and his M.A. and Ph.D. in commutative algebra from the University of Oklahoma in 1974 and 1978, respectively. Since 1981 he has been teaching mathematics and computer science at Hampden-Sydney College in Virginia, where he is currently the chair of the department. His main mathematical interest is number theory, although he enjoys teaching in all areas, and in recent years he has spent much of his time working in computer science. Outside of the office he enjoys backpacking on the Appalachian Trail.

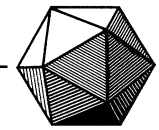
John K. Osoinach, Jr. received his B.S. from Vanderbilt University in 1990, his M.A. from Rice University in 1993, and his Ph.D. in topology from the University of Texas at Austin in 1998. He has taught at Eureka College and at Hampden-Sydney College, where he is an assistant professor. He remains interested in the geometry and topology of 3-manifolds, and he has found great satisfaction in directing undergraduate research projects in topology at Hampden-Sydney. His interest in game theory arose from teaching the mathematics of strategy in his introductory mathematics courses. His current work involves helping his 6-month-old son, Allen, determine which shapes are most easily thrown across the room.

G. L. (Jerry) Alexanderson and Peter Ross were both surprised to read in Steven Krantz's MAA book *Mathematical Apocrapha* that the general Stone-Weierstrass Theorem was published first in MATHEMATICS MAGAZINE. This was the stimulus for their article on gems in the MAGAZINE.

Jerry is in the Department of Mathematics and Computer Science at Santa Clara University, where he is the Valeriote Professor of Science. His most recent book was a biography of his advisor at Stanford, George Pólya. Having served as Editor of MATHEMATICS MAGAZINE, First Vice President, Secretary, and President of the MAA, he was a member of the MAA's Board of Governors for a total of 25 years. He received his reprieve from the Board in January 2005. Currently he edits the Spectrum book series for the MAA.

Peter is in the same department, which he joined in 1982. He completed both his M.A. and Ph.D. at the University of California, Berkeley, with an interlude teaching secondary school in India as a Peace Corps Volunteer from 1963 to 1965. Peter's hobbies include singing choral music, riding an exciting bicycle commute in Silicon Valley, and writing Media Highlights for the *College Mathematics Journal*.

Vol. 78, No. 2, April 2005



MATHEMATICS MAGAZINE

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

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

Advertising correspondence should be addressed to Frank Peterson (*FPetersonj@aol.com*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

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

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

Printed in the United States of America

Groups of Arithmetical Functions

JAMES E. DELANY, *Emeritus*
California Polytechnic State University
San Luis Obispo, CA 93407
jdelany@calpoly.edu

An *arithmetical function* is a mapping from the positive integers to the complex numbers. The more interesting ones involve some number-theoretic property, such as

$\tau(n) =$ the number of positive divisors of n,

$\sigma(n) =$ the sum of the positive divisors of n, and

$\phi(n) =$ the number of positive integers $k \leq n$ such that $\gcd(k, n) = 1$.

A typical introductory number theory book includes a chapter on these functions, showing that they form a commutative ring with unity under pointwise addition

$$(f + g)(n) = f(n) + g(n)$$

and *Dirichlet multiplication*

$$(f * g)(n) = \sum_{d|n} f(d)g(n/d).$$

Here the sum is taken over all positive integer divisors d of n. This somewhat surprising choice of a product is quite fruitful, allowing one to obtain interesting number-theoretic formulas from simple computations in the ring. In particular, the useful functions mentioned above can all be expressed in terms of two simple elements of this ring.

In this MAGAZINE, Berberian [**2**] discussed (among other things) the group of units of this ring. He showed that τ, σ, and ϕ can be expressed in terms of two very simple functions and proved that those two functions are linearly independent. In this article we extend his pair to an uncountably infinite set. In the process, we present answers to other questions posed in his article, including a description of the structure of the group of units.

In the interest of accessibility, most of the discussion is confined to real-valued arithmetical functions. Except for a bit of abelian group theory, the algebraic ideas come from introductory linear algebra and abstract algebra. For many readers the only novel concept will be Bell series, a powerful tool developed by E. T. Bell in the early twentieth century.

NOTATION. The symbols $\mathbb{P}$, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ will denote the positive integers, nonnegative integers, integers, rational numbers, real numbers, and complex numbers, respectively.

Background

We develop some basic principles of the ring of arithmetical functions. Our presentation is self-contained, but the reader desiring more information may consult various introductory number theory books, such as Niven and Zuckerman [**6**, Chapter 4] or Rosen [**7**, Chapter 7]. Apostol [**1**, Chapter 2] is particularly helpful.

First, the Dirichlet product can also be expressed as

$$(f * g)(n) = \sum_{d_1 d_2 = n} f(d_1) g(d_2) \tag{1}$$

where the sum extends over all ordered pairs of positive divisors of n whose product is n. Extending this notation, the associative law states that

$$\big(f * (g * h)\big)(n) = \big((f * g) * h\big)(n) = \sum_{d_1 d_2 d_3 = n} f(d_1) g(d_2) h(d_3).$$

The Dirichlet product is particularly easy to evaluate at a prime power, p^k:

$$(f * g)(p^k) = \sum_{i=0}^{k} f(p^i) g(p^{k-i}).$$

The multiplicative identity of the ring is

$$I(n) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

To determine the group of units, we ask which arithmetical functions are invertible, in the sense of the Dirichlet product. As long as $f(1) \neq 0$, we can obtain f^{-1} inductively: $f^{-1}(1) = 1/f(1)$ and, when $n > 1$,

$$f^{-1}(n) = -\frac{1}{f(1)} \sum_{\substack{d \mid n \\ d < n}} f^{-1}(d) f(n/d).$$

Again, the formula simplifies for a prime power, p^k, $k > 0$:

$$f^{-1}(p^k) = -\frac{1}{f(1)} \sum_{i=0}^{k-1} f^{-1}(p^i) f(p^{k-i}). \tag{2}$$

Scalar multiplication is defined as usual: $(cf)(n) = cf(n)$. Equation (1) makes it clear that $(cf) * g = f * (cg) = c(f * g)$.

Of particular interest are the functions that are *multiplicative*, those having the properties that $f(1) = 1$ and $f(mn) = f(m) f(n)$ whenever $\gcd(m, n) = 1$. The functions I, τ, σ, and ϕ are all multiplicative. The multiplicative functions form a subgroup of the group of units [**1**, Section 2.10]. A multiplicative function is uniquely determined by its values on the prime powers: if $p_1, \ldots, p_r$ are distinct primes, then

$$f(p_1^{k_1} \cdots p_r^{k_r}) = \prod_{i=1}^{r} f(p_i^{k_i}).$$

EXAMPLE. For each real α, the function ε_α defined by $\varepsilon_\alpha(n) = n^\alpha$ is multiplicative. In fact, it is *totally multiplicative* or *completely mutiplicative* in that $\varepsilon_\alpha(mn) = \varepsilon_\alpha(m)\varepsilon_\alpha(n)$ for all $m, n \in \mathbb{P}$. Then ε_α^{-1} must also be multiplicative, so $\varepsilon_\alpha^{-1}(1) = 1$ and it suffices to compute $\varepsilon_\alpha^{-1}(p^k)$, where p is a prime and k is a positive integer. From (2) we have $\varepsilon_\alpha^{-1}(p) = -\varepsilon_\alpha^{-1}(1)\varepsilon_\alpha(p) = -p^\alpha$. A routine induction, again using (2), shows that $\varepsilon_\alpha^{-1}(p^k) = 0$ when $k \geq 2$. Now suppose that $n > 1$ has prime factorization $n = p_1^{k_1} \cdots p_r^{k_r}$. Then $\varepsilon_\alpha^{-1}(n) = \prod_{i=1}^{r} \varepsilon_\alpha^{-1}(p_i^{k_i})$. This is zero if any one of the

k_i exceeds one. If each $k_i = 1$ we have $n = \prod_{i=1}^{r} p_i$ and $\varepsilon_\alpha^{-1}(n) = \prod_{i=1}^{r} \varepsilon_\alpha^{-1}(p_i) = \prod_{i=1}^{r}(-p_i^\alpha) = (-1)^r n^\alpha$. In summary, we have proved

$$\varepsilon_\alpha^{-1}(n) = \begin{cases} 1 & \text{if } n = 1 \\ (-1)^r n^\alpha & \text{if } n \text{ is the product of } r \text{ distinct primes} \\ 0 & \text{otherwise.} \end{cases}$$

The functions ε, ε_1, and μ Note that $\varepsilon_0(n) = 1$ and $\varepsilon_1(n) = n$ for all n. These functions play a key role; in fact, ε_0 and ε_1 are the two functions featured by Berberian in his article [**2**]. The function ε_0 occurs so often that we let $\varepsilon = \varepsilon_0$. Despite their importance, there is little agreement on notation, as seen in Table 1.

TABLE 1: Notation for I, ε, and ε_1

Author(s)	I	ε	ε_1
Apostol [**1**]	I	u	N
Berberian [**2**]	u	γ	ε
McCarthy [**5**]	δ	ζ	ζ_1
Niven and Zuckerman [**6**]	I	U	E
Rosen [**7**]	ι	ν	–

We can express τ, σ, and ϕ in terms of ε and ε_1 using the following idea: Suppose f is an arithmetical function and let

$$F(n) = \sum_{d|n} f(d) = \sum_{d|n} f(d)\varepsilon(n/d) = (f * \varepsilon)(n),$$

showing that $F = f * \varepsilon$. Since $\tau(n) = \sum_{d|n} 1$ and $\sigma(n) = \sum_{d|n} d$ we have

$$\tau = \varepsilon * \varepsilon$$

$$\sigma = \varepsilon_1 * \varepsilon.$$

There is a pretty equation involving Euler's ϕ function, the third on our initial list of examples:

$$\sum_{d|n} \phi(d) = n. \tag{3}$$

To see this, note that $\phi(d)$ equals the number of reduced fractions having denominator d in the interval (0, 1]. If we partition the set $\{1/n, 2/n, \ldots, n/n\}$ according to the denominators of the fractions in reduced form, the sum adds the cardinalities of these equivalence classes, and this total must be n. In terms of the Dirichlet product, (3) says $\phi * \varepsilon = \varepsilon_1$, or

$$\phi = \varepsilon_1 * \varepsilon^{-1}.$$

Thus τ, σ, and ϕ are each expressible in terms of ε and ε_1.

The Möbius function $\mu = \varepsilon^{-1}$ appears often, as in $\phi = \varepsilon_1 * \mu$. Taking $\alpha = 0$ in the formula for ε_α^{-1} yields

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1 \\ (-1)^r & \text{if } n \text{ is the product of } r \text{ distinct primes} \\ 0 & \text{otherwise.} \end{cases}$$

The *Möbius inversion formula* [**1**, p. 32] states that if $F(n) = \sum_{d|n} f(d)$ then $f(n) = \sum_{d|n} F(d)\mu(n/d)$. In our setting this reduces to the assertion that if $F = f * \varepsilon$ then $f = F * \varepsilon^{-1}$.

The group of units

Our initial goal is to obtain an algebraic description of the group of units. We begin by showing it is the direct sum of three subgroups: the scalars, the multiplicative functions, and a group to be defined momentarily.

First we split off the scalar functions.

Let $U = \{f \mid f(1) \neq 0\}$, $U_1 = \{f \mid f(1) = 1\}$, and $C = \{cI \mid c \in \mathbb{R}, c \neq 0\}$. Then C and U_1 are subgroups of U and $C \cap U_1 = \{I\}$. If $f \in U$ and $c = f(1)$ then $f = (cI) * (\frac{1}{c} f)$ with $cI \in C$ and $\frac{1}{c} f \in U_1$. Thus $U = C \oplus U_1$.

There are many important functions for which $f(1) \neq 1$.

EXAMPLE. (SUMS OF SQUARES) Hardy and Wright [**4**, p. 314] used $r_k(n)$ to denote the number of k-tuples $(a_1, a_2, \ldots, a_k)$ of integers for which $a_1^2 + a_2^2 + \cdots + a_k^2 = n$. The two most familiar cases are r_2 and r_4. It turns out that each is a scalar times a multiplicative function. In the first case, $r_2(n) = 4 \sum_{d|n} \chi(d)$, where χ is the completely multiplicative function given by

$$\chi(n) = \begin{cases} 0 & \text{if } n \text{ is even} \\ (-1)^{(n-1)/2} & \text{if } n \text{ is odd} \end{cases}$$

[**4**, p. 241]. Thus $r_2 = 4I * \chi * \varepsilon$, and $\chi * \varepsilon$ is multiplicative.

Lagrange showed that every positive integer is expressible as the sum of four squares, so r_4 is always positive. One formula is

$$r_4(n) = 8 \sum_{d|n, 4\nmid d} d$$

[**4**, p. 314]. Defining the multiplicative function f by

$$f(n) = \begin{cases} 0 & \text{if } 4 \mid n \\ n & \text{otherwise,} \end{cases}$$

we have $r_4(n) = 8 \sum_{d|n} f(d)$, and $r_4 = 8I * f * \varepsilon$ with $f * \varepsilon$ multiplicative.

The first few values of r_2 and r_4 are shown in Table 2. One reason these functions get special attention is that they can be related to factorization in the Gaussian integers and integer quaternions, respectively [**4**, Chapter 20].

TABLE 2: Number of ways to express n as the sum of two or four squares

n	1	2	3	4	5	6	7	8	9	10	11	12
χ	1	0	−1	0	1	0	−1	0	1	0	−1	0
$r_2 = 4I * \chi * \varepsilon$	4	4	0	4	8	0	0	4	4	8	0	0
f	1	2	3	0	5	6	7	0	9	10	11	0
$r_4 = 8I * f * \varepsilon$	8	24	32	24	48	96	64	24	104	144	96	96

Antimultiplicative functions Let U_M denote the subgroup of multiplicative functions in U. The desired complement of U_M in U_1 consists of functions we will call *antimultiplicative*, meaning $f(1) = 1$ and $f(p^k) = 0$ whenever p^k is a prime power with $k > 0$. Let U_A be the set of such functions.

To begin with, U_A is a subgroup of U_1. It is nonempty since $I \in U_A$. If $f, g \in U_A$ and $k > 0$ then $(f * g)(p^k) = \sum_{i=0}^{k} f(p^i)g(p^{k-i}) = \sum_{i=0}^{k} 0 = 0$ so $f * g \in U_A$. When $k > 0$ we also have $f^{-1}(p^k) = -\sum_{i=0}^{k-1} f^{-1}(p^i)f(p^{k-i}) = -\sum_{i=0}^{k-1} 0 = 0$ so $f^{-1} \in U_A$. Thus U_A is a group.

It is clear that $U_M \cap U_A = \{I\}$. We would like to be able to separate an arithmetical function into multiplicative and antimultiplicative pieces. Given $f \in U_1$, define g by

$$g(p_1^{k_1} \cdots p_r^{k_r}) = \prod_{i=1}^{r} f(p_i^{k_i})$$

and let $h = g^{-1} * f$. Then $g \in U_M$ and we claim $h \in U_A$. For $k > 0$, we compute $h(p^k) = (g^{-1} * f)(p^k) = \sum_{i=0}^{k} g^{-1}(p^i)f(p^{k-i}) = \sum_{i=0}^{k} g^{-1}(p^i)g(p^{k-i}) = (g^{-1} * g)(p^k) = I(p^k) = 0$. Thus $f = g * h$ with $g \in U_M$ and $h \in U_A$, so $U_M \oplus U_A = U_1$.

EXAMPLE. Von Mangoldt's Λ function [**1**, p. 32] is given by

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k,\ p \text{ prime},\ k > 0 \\ 0 & \text{otherwise.} \end{cases}$$

It is useful in studying the distribution of primes. Although Λ is not a unit, e^{Λ} is in U_1. Let

$$f(n) = e^{\Lambda(n)} = \begin{cases} p & \text{if } n = p^k,\ p \text{ prime},\ k > 0 \\ 1 & \text{otherwise.} \end{cases}$$

Now let us compute the multiplicative component of f. It turns out to involve the *core* of an integer, which is the product of its distinct prime divisors: $\gamma(p_1^{k_1} \cdots p_r^{k_r}) = p_1 \cdots p_r$. From the definition, the multiplicative part of f is the function g given by

$$g(p_1^{k_1} \cdots p_r^{k_r}) = \prod_{i=1}^{r} f(p_i^{k_i}) = p_1 \cdots p_r = \gamma(p_1^{k_1} \cdots p_r^{k_r}).$$

Table 3 shows the first few values of e^{Λ}, its multiplicative component γ, and its antimultiplicative component $\gamma^{-1} * e^{\Lambda}$. It is instructive to compute a few examples to verify that $f = g * h$.

TABLE 3: Multiplicative and antimultiplicative components of e^{Λ}

n	1	2	3	4	5	6	7	8	9	10	11	12
$e^{\Lambda} = f$	1	2	3	2	5	1	7	2	3	1	11	1
$\gamma = g$	1	2	3	2	5	6	7	2	3	10	11	6
$\gamma^{-1} * e^{\Lambda} = h$	1	0	0	0	0	−5	0	0	0	−9	0	5

This completes the first step in the description of the group of units: $U = C \oplus U_M \oplus U_A$.

Powers It is actually possible to regard U_1 as a rational vector space in which U_M and U_A are complementary subspaces. This vector space structure is somewhat surprising, as it does not involve addition, but instead entails raising units to rational powers with respect to the Dirichlet product. For $f \in U$ let $f^{(0)} = I$, $f^{(1)} = f$, $f^{(2)} = f * f$, etc. When m is a negative integer, let $f^{(m)} = (f^{-1})^{(-m)}$. Needless to say, these powers obey the usual laws of exponents. The associative law extends to m factors as

$$(f_1 * \cdots * f_m)(n) = \sum_{d_1 \cdots d_m = n} f_1(d_1) \cdots f_m(d_m).$$

When $f_1 = \cdots = f_m = f$ this reduces to

$$f^{(m)}(n) = \sum_{d_1 \cdots d_m = n} f(d_1) \cdots f(d_m).$$

As an example, consider powers of ε and $\mu = \varepsilon^{-1}$. If m is a positive integer then

$$\varepsilon^{(m)}(n) = \sum_{d_1 \cdots d_m = n} \varepsilon(d_1) \cdots \varepsilon(d_m) = \sum_{d_1 \cdots d_m = n} 1.$$

In other words, this counts the number of ways to express n as the product of m positive divisors, taking the order of the factors into account. Since ε is multiplicative, $\varepsilon^{(m)}$ is also multiplicative and it suffices to determine it on prime powers. Here

$$\varepsilon^{(m)}(p^k) = \sum_{p^{k_1} \cdots p^{k_m} = p^k} 1 = \sum_{k_1 + \cdots + k_m = k} 1,$$

where each $k_i \in \mathbb{N}$. The number of ways to express k as the sum of m nonnegative integers is $\binom{m+k-1}{m-1}$. To see this, form a row of $m + k - 1$ 1s, choose $m - 1$ of them to be replaced by $+$ signs; regrouping gives an expression for k as desired. For instance, if $k = 5$ and $m = 3$, the row 1 1 1 1 1 1 1 could become 1 1 + +1 1 1, leading to $5 = 2 + 0 + 3$. Thus,

$$\varepsilon^{(m)}(p^k) = \binom{m+k-1}{m-1} = \binom{m+k-1}{k}.$$

We claim that $\mu^{(m)}(p^k) = (-1)^k \binom{m}{k}$ for $m \geq 0$. To begin with, $I(p^k) = (-1)^k \binom{0}{k}$. The induction step follows from $\mu^{(m+1)} = \mu * \mu^{(m)}$ and the identity

$$\binom{m}{k} + \binom{m}{k-1} = \binom{m+1}{k}.$$

When $m < 0$,

$$\mu^{(m)}(p^k) = \varepsilon^{(-m)}(p^k) = \binom{-m+k-1}{k} = (-1)^k \binom{m}{k}.$$

In other words, the formula for positive m also works for negative m. Similarly, if $m < 0$, then

$$\varepsilon^{(m)}(p^k) = \mu^{(-m)}(p^k) = (-1)^k \binom{-m}{k} = \binom{m+k-1}{k},$$

just as in the case $m \geq 0$. In summary,

PROPOSITION 1. *If m is an integer and $n = p_1^{k_1} \cdots p_r^{k_r}$ then*

$$\varepsilon^{(m)}(n) = \mu^{(-m)}(n) = \prod_{i=1}^{r} \binom{m + k_i - 1}{k_i}$$

and

$$\mu^{(m)}(n) = \varepsilon^{(-m)}(n) = \prod_{i=1}^{r} (-1)^{k_i} \binom{m}{k_i}.$$

Furthermore, when $m > 0$, $\varepsilon^{(m)}(n)$ equals the number of ways to express n as the product of m positive integers, taking the order of the factors into account.

For example, $\tau(n) = \varepsilon^{(2)}(n) = \prod_{i=1}^{r} \binom{1+k_i}{k_i} = \prod_{i=1}^{r} (1 + k_i)$.

EXERCISE. For $m \geq 0$ show that $\varepsilon_\alpha^{(m)} = \varepsilon_\alpha \varepsilon^{(m)}$. More generally, if f is any arithmetical function then $(\varepsilon_\alpha f)^{(m)} = \varepsilon_\alpha f^{(m)}$. Here $\varepsilon_\alpha f$ denotes ordinary multiplication.

Elements of finite order Does U_1 have any elements of finite order? A group is said to be *torsion-free* if the only element of finite order is the identity. The group U_1 is torsion-free.

To see this, assume $f^{(m)} = I$ with $m > 0$. Suppose $n > 1$ and $f(d) = 0$ for all $1 < d < n$. Then

$$0 = f^{(m)}(n) = \sum_{d_1 \cdots d_m = n} f(d_1) \cdots f(d_m).$$

The only summands that might not be zero are those in which one of the factors d_i equals n and the rest equal 1. Then $0 = f^{(m)}(n) = mf(n)$ and $f(n) = 0$.

The next step in defining rational powers is to show the existence of roots. In a torsion-free abelian group, roots are unique when they exist. In our situation the reasoning is that if $f^{(m)} = g^{(m)}$ then $(f * g^{-1})^{(m)} = I$, and since U_1 is torsion-free, $f * g^{-1} = I$ and $f = g$.

Roots What happens when we try to construct roots? Suppose $g \in U_1$ and m is a positive integer. We are looking for $f \in U_1$ such that $f^{(m)} = g$. To begin with, $f(1) = 1$. For $n > 1$,

$$g(n) = f^{(m)}(n) = \sum_{d_1 \cdots d_m = n} f(d_1) \cdots f(d_m).$$

Separating out the summands involving $f(n)$ gives

$$g(n) = mf(n) + \sum_{\substack{d_1 \cdots d_m = n \\ d_1, \ldots, d_m < n}} f(d_1) \cdots f(d_m)$$

and solving for $f(n)$ we get

$$f(n) = \frac{1}{m} \left(g(n) - \sum_{\substack{d_1 \cdots d_m = n \\ d_1, \ldots, d_m < n}} f(d_1) \cdots f(d_m) \right).$$

Thus $f(n)$ can be determined inductively, and it is unique.

When $f^{(m)} = g$ we write $g^{(1/m)} = f$.

EXAMPLE. Let us find $\varepsilon^{(1/2)}$ and $\varepsilon^{(1/3)}$. Let f be the function such that $f^{(2)} = \varepsilon$. The preceding formula tells us

$$f(p) = (1/2)(1 - 0) = 1/2$$
$$f(p^2) = (1/2)(1 - f(p)^2) = 3/8$$
$$f(p^3) = (1/2)(1 - 2f(p)f(p^2)) = 5/16$$

These values of $\varepsilon^{(1/2)}(p^k)$ are displayed in Table 4. The mth root of a multiplicative function is also multiplicative, for reasons to be explained soon. This allows us to compute the remaining displayed values of $\varepsilon^{(1/2)}$. Similar calculations yield the indicated values of $\varepsilon^{(1/3)}$.

TABLE 4: Dirichlet roots of ε

n	1	2	3	4	5	6	7	8	9	10	11	12
$\varepsilon^{(1/2)}$	1	1/2	1/2	3/8	1/2	1/4	1/2	5/16	3/8	1/4	1/2	3/16
$\varepsilon^{(1/3)}$	1	1/3	1/3	2/9	1/3	1/9	1/3	14/81	2/9	1/9	1/3	2/27

The vector space $(U_1, *)$ Before describing the structure of U_1, we review some properties of abelian groups, bearing in mind that they are normally discussed using additive notation. An abelian group $(G, +)$ is said to be *divisible* if for each $g \in G$ and each $n \in \mathbb{P}$ there is an $x \in G$ such that $nx = g$. If G is torsion-free, x will be unique. This allows one to view G as a vector space over $\mathbb{Q}$ by letting $(m/n)g$ be the unique solution to $nx = mg$. We have shown that $(U_1, *)$ is a divisible torsion-free group, so we can view it as a vector space over the rationals.

For a thorough discussion of divisible groups, see Fuchs [**3**, Chapter IV]. Divisible groups have some extremely nice properties. A subgroup of a divisible group is a direct summand if and only if it is divisible. Since U_M and U_A are complementary summands of the group U_1, they are divisible subgroups and therefore complementary subspaces when U_1 is regarded as a vector space. In particular, this implies that the nth root of a multiplicative function is multiplicative (as asserted in constructing Table 4) and that the nth root of an antimultiplicative function is antimultiplicative. In summary,

THEOREM 1. *For $f \in U_1$ and $m/n \in \mathbb{Q}$ let $f^{(m/n)}$ denote the unique $g \in U_1$ such that $g^{(n)} = f^{(m)}$. Defining scalar multiplication $\mathbb{Q} \times U_1 \to U_1$ by $(q, f) \to f^{(q)}$ makes the group $(U_1, *)$ a vector space over $\mathbb{Q}$. Furthermore, U_M and U_A are complementary subspaces: $U_1 = U_M \oplus U_A$.*

This result answers two questions posed by Berberian [**2**]. It describes the structure of the group U_M of multiplicative functions and, since $U = C \oplus U_M \oplus U_A$, it also describes the quotient group U/U_M.

He also showed that the two functions ε and ε_1 are linearly independent, and posed the problem of finding a third. In order to study independence, we introduce a new tool.

Bell series

With each arithmetical function we associate a family of formal power series called *generating functions* that distills much of the information about the function. If f is an arithmetical function and p is a prime, then the *Bell series of f with respect to p* is

the formal power series

$$f_p(X) = \sum_{k=0}^{\infty} f(p^k)X^k.$$

Discussions of the basic ideas of these series can be found in Apostol [**1**, p. 43] and McCarthy [**5**, p. 60]. When speaking of Bell series we often omit the phrase "for each prime p." The statement $I_p(X) = 1$ is intended to mean that this is true for each prime p.

Formulas involving Maclaurin series carry over to Bell series. The geometric series is especially useful. For example,

$$\varepsilon_p(X) = 1 + X + X^2 + \cdots = 1/(1 - X),$$

meaning that $1 + X + X^2 + \cdots = (1 - X)^{-1}$ in $\mathbb{R}[[X]]$, the ring of formal power series. More generally,

$$(\varepsilon_\alpha)_p(X) = 1 + p^\alpha X + p^{2\alpha}X^2 + \cdots = 1/(1 - p^\alpha X).$$

Another well-known function is the Liouville λ [**1**, p. 37] given by

$$\lambda(p_1^{k_1} \cdots p_r^{k_r}) = (-1)^{k_1 + \cdots k_r}.$$

Here $\lambda(p^k) = (-1)^k$ and

$$\lambda_p(X) = 1 - X + X^2 - X^3 + \cdots = 1/(1 + X).$$

The calculation of the series for the core function γ is only slightly more complicated:

$$\gamma_p(X) = 1 + pX + pX^2 + pX^3 + \cdots = 1 + \frac{pX}{(1 - X)} = \frac{1 - (1 - p)X}{1 - X}.$$

The feature of Bell series that makes them so valuable is that $(f * g)_p(X) = f_p(X)g_p(X)$. This follows from the rule for multiplying power series:

$$\begin{aligned} f_p(X)g_p(X) &= \left(\sum_{i=0}^{\infty} f(p^i)X^i\right)\left(\sum_{j=0}^{\infty} g(p^j)X^j\right) \\ &= \sum_{k=0}^{\infty}\left(\sum_{i=0}^{k} f(p^i)g(p^{k-i})\right)X^k \\ &= \sum_{k=0}^{\infty}(f * g)(p^k)X^k = (f * g)_p(X). \end{aligned}$$

Bell series are most useful in studying multiplicative functions. If $f, g \in U_M$ then, since a multiplicative function is determined by its values on prime powers, $f = g$ if and only if $f_p(X) = g_p(X)$ for each prime p. On the other hand, they are useless when it comes to antimultiplicative functions: $f \in U_1$ is antimultiplicative if and only if $f_p(X) = 1$ for each prime p. If $f \in U_1$, then the Bell series of f equals the Bell series of its multiplicative component.

The multiplicative property implies that if $m \in \mathbb{N}$ then $(f^{(m)})_p(X) = f_p(X)^m$. When $f \in U$ we have $(f^{-1})_p(X)f_p(X) = I_p(X) = 1$ so $(f^{-1})_p(X) = f_p(X)^{-1}$. Then $f_p^{(m)}(X) = f_p(X)^m$ when m is a negative integer as well.

Table 5 lists a few Bell series and illustrates some basic properties: μ and ε are inverses, and their series are reciprocals; σ and ϕ illustrate the product rule; τ is an example of a power.

TABLE 5: Selected Bell series

f	$f_p(X)$
I	1
μ	$1 - X$
ε	$1/(1 - X)$
ε_1	$1/(1 - pX)$
$\tau = \varepsilon * \varepsilon$	$1/(1 - X)^2$
$\sigma = \varepsilon * \varepsilon_1$	$1/\big((1 - X)(1 - pX)\big)$
$\phi = \mu * \varepsilon_1$	$(1 - X)/(1 - pX)$
γ	$\big(1 - (1 - p)X\big)/(1 - X)$
λ	$1/(1 + X)$

What about Bell series of rational powers? Suppose $f \in U_1$ and $f_p(X) = F(X) = 1 + \sum_{k=1}^{\infty} a_k X^k$. If m is a positive integer, there is a unique $g \in U_1$ for which $g^{(m)} = f$, in which case $g_p(X)^m = f_p(X)$. On the other hand, there is a unique series $G(x) = 1 + \sum_{k=1}^{\infty} b_k X^k$ such that $G(X)^n = F(X)$: Its coefficients can be calculated inductively and are uniquely determined. Then $g_p(X)$ must equal $G(X)$, so there is no ambiguity in writing $(f^{(1/n)})_p(X) = f_p(X)^{1/n}$ or, for that matter, $(f^{(m/n)})_p(X) = f_p(X)^{m/n}$.

Completely multiplicative functions We illustrate the value of Bell series by using them to determine rational powers of completely multiplicative functions. As noted earlier, each ε_α has this property. Two more examples are Liouville's λ and the χ used in computing r_2. We can generalize the former to λ_β, $\beta \neq 0$, by letting

$$\lambda_\beta(p_1^{k_1} \cdots p_r^{k_r}) = \beta^{k_1 + \cdots k_r}.$$

Each λ_β is completely multiplicative, as are all the products $\varepsilon_\alpha \lambda_\beta$.

If f is completely multiplicative, then $f(p^k) = f(p)^k$ and f is determined by its values on the primes. In that case the Bell series are

$$f_p(X) = \sum_{k=0}^{\infty} f(p^k) X^k = \sum_{k=0}^{\infty} f(p)^k X^k = 1/(1 - f(p)X).$$

For example, $(\varepsilon_\alpha \lambda_\beta)_p(X) = \sum_{k=0}^{\infty} p^{k\alpha} \beta^k X^k = 1/(1 - p^\alpha \beta X)$.

Calculating rational powers of completely multiplicative functions involves binomial series. Many calculus students know that, for $s \in \mathbb{R}$,

$$(1 + x)^s = 1 + \sum_{k=1}^{\infty} \binom{s}{k} x^k$$

where $\binom{s}{k} = s(s - 1) \cdots (s - k + 1)/k!$, even when s is not an integer [**8**, p. 809]. This series converges for $|x| < 1$. In our setting, namely the power series ring $\mathbb{R}[[X]]$, defining $(1 + X)^s$ to mean the series $1 + \sum_{k=1}^{\infty} \binom{s}{k} X^k$ extends the binomial theorem for integer exponents to real exponents in a manner consistent with the ring operations.

Replacing s by $-s$ and X by $-CX$ yields the formula

$$(1 - CX)^{-s} = 1 + \sum_{k=1}^{\infty} \binom{-s}{k} (-1)^k (CX)^k.$$

But

$$\binom{-s}{k}(-1)^k = \binom{s+k-1}{k},$$

so

$$(1 - CX)^{-s} = 1 + \sum_{k=1}^{\infty} \binom{s+k-1}{k} C^k X^k.$$

For example, if q is rational then

$$(\varepsilon^{(q)})_p(X) = (1 - X)^{-q} = \sum_{k=0}^{\infty} \binom{q+k-1}{k} X^k$$

and

$$\varepsilon^{(q)}(p^k) = \binom{q+k-1}{k}.$$

Thus Proposition 1 extends to rational powers:

$$\varepsilon^{(q)}(p_1^{k_1} \cdots p_r^{k_r}) = \prod_{i=1}^{r} \binom{q+k_i-1}{k_i}. \tag{4}$$

EXAMPLE. $\varepsilon^{(1/2)}(p^k) = \binom{k-1/2}{k}$, where

$$\begin{aligned}\binom{k-\frac{1}{2}}{k} &= \frac{(k-\frac{1}{2})(k-\frac{3}{2})\cdots(\frac{1}{2})}{k!}\left(\frac{2^k}{2^k}\right) \\ &= \frac{(1)(3)\cdots(2k-1)}{2^k k!},\end{aligned}$$

so $\varepsilon^{(1/2)}(p) = 1/2$, $\varepsilon^{(1/2)}(p^2) = 3/8$, and $\varepsilon^{(1/2)}(p^3) = 5/16$, just as in Table 4. We could also write

$$\begin{aligned}\varepsilon^{(1/2)}(p^k) &= \left(\frac{(1)(3)\cdots(2k-1)}{2^k k!}\right)\left(\frac{(2)(4)(6)\cdots(2k)}{2^k k!}\right) \\ &= \frac{1}{4^k}\binom{2k}{k}.\end{aligned}$$

Similarly

$$\varepsilon^{(1/3)}(p^k) = \binom{k-\frac{2}{3}}{k} = \frac{(1)(4)(7)\cdots(3k-2)}{3^k k!}.$$

As in Table 4, $\varepsilon^{(1/3)}(p) = 1/3$, $\varepsilon^{(1/3)}(p^2) = 2/9$, and $\varepsilon^{(1/3)}(p^3) = 14/81$.

When f is completely multiplicative,

$$\begin{aligned}(f^{(q)})_p(X) &= f_p(X)^q = (1 - f(p)X)^{-q} \\ &= \sum_{k=0}^{\infty} \binom{q+k-1}{k} f(p)^k X^k \\ &= \sum_{k=0}^{\infty} \varepsilon^{(q)}(p^k) f(p^k) X^k \\ &= (\varepsilon^{(q)} f)_p(X).\end{aligned}$$

In other words, $f^{(q)} = \varepsilon^{(q)} f$. Then, from the formula (4) for $\varepsilon^{(q)}$, we obtain

PROPOSITION 2. *If f is completely multiplicative and $q \in \mathbb{Q}$ then $f^{(q)} = \varepsilon^{(q)} f$. When $n = p_1^{k_1} \cdots p_r^{k_r}$,*

$$f^{(q)}(n) = f(n) \prod_{i=1}^{r} \binom{q+k_i-1}{k_i}.$$

Linear independence With Bell series at our disposal, we can obtain some sweeping results about linear independence in U_M. In the sense of this vector space, a set $\mathcal{F} \subseteq U_1$ is linearly dependent if and only if there exist distinct functions $f_1, \ldots, f_r \in \mathcal{F}$ and rationals $q_1, \ldots, q_r$, not all zero, such that $f_1^{(q_1)} * \cdots * f_r^{(q_r)} = I$. In that case, let N be a positive integer such that all the $m_i = q_i N$ are integers. Then $f_1^{(q_1 N)} * \cdots * f_r^{(q_r N)} = I^{(N)}$, or $f_1^{(m_1)} * \cdots * f_r^{(m_r)} = I$. In other words, it suffices to consider integer exponents. In U_M we have $f_1^{(m_1)} * \cdots * f_r^{(m_r)} = I$ if and only if $\prod_{i=1}^{r} (f_i)_p(X)^{m_i} = 1$ for each prime p. The following is useful in dealing with such products.

LEMMA 1. *Suppose $\prod_{i=1}^{r} P_i(X)^{m_i} = 1$, where the P_i are nonconstant polynomials, not necessarily distinct, and the m_i are integers. If some P_k is relatively prime to all the others, then $m_k = 0$.*

In particular, when $\prod_{i=1}^{r} (1 - C_i X)^{-m_i} = 1$, where the C_i are nonzero constants, if some C_k is different from all the others, then $m_k = 0$.

Proof. Rewrite the equation as

$$\prod_{m_i < 0} P_i(X)^{-m_i} = \prod_{m_j \geq 0} P_j(X)^{m_j}$$

to obtain polynomials on both sides of the equation. If P_k is relatively prime to the others and $m_k \neq 0$, then P_k divides one side but not the other, a contradiction. ∎

Before offering our principal result on independence, we illustrate the ideas involved with a special case.

PROPOSITION 3. *The functions $\{\varepsilon_\alpha \lambda_\beta \mid \alpha, \beta \in \mathbb{R}, \beta \neq 0\}$ are linearly independent.*

Proof. Suppose $f_1^{(m_1)} * \cdots * f_r^{(m_r)} = I$, where the m_i are integers, $f_i = \varepsilon_{\alpha_i} \lambda_{\beta_i}$, and the f_i are all distinct. We must show that each m_i is zero. In terms of Bell series we have $\prod_{i=1}^{r} (f_i)_p(X)^{m_i} = 1$ for every prime p, or

$$\prod_{i=1}^{r} (1 - p^{\alpha_i} \beta_i X)^{-m_i} = 1.$$

Let $C_i(p) = p^{\alpha_i}\beta_i$. Consider all the equations $C_i(p) = C_j(p)$ with $i \neq j$. Each equation is satisfied by at most one prime p: If a solution to $p^{\alpha_i}\beta_i = p^{\alpha_j}\beta_j$ does exist it can only be $p = (\beta_j/\beta_i)^{1/(\alpha_i-\alpha_j)}$. Thus there are at most a finite number of solutions altogether, so there must be a prime that makes the $C_i(p)$ all different. Lemma 1 then implies that each m_i is zero. ∎

EXERCISE. (POWERS OF γ) For $\alpha \in \mathbb{R}$ show that

$$(\gamma^\alpha)_p(X) = \frac{1-(1-p^\alpha)X}{1-X}$$

and that $\{\gamma^\alpha \mid \alpha \in \mathbb{R}\}$ is a linearly independent set.

Extension to $\mathbb{C}$ All of the foregoing extends to complex-valued arithmetical functions. In this setting $\varepsilon_\alpha(n) = n^\alpha = e^{\alpha \ln n}$. The function λ_β needs no special consideration since the exponents involved are all integers. The only delicate point arises in the proof of the preceding proposition, where we used the fact that each equation $C_i(p) = C_j(p)$, $i \neq j$, had at most one solution. This is not the case in the complex numbers, since it is possible to have $p^\alpha = q^\alpha$ for distinct primes p, q. For instance, if $\alpha = 2\pi i/\ln(3/2)$, then $(3/2)^\alpha = e^{\alpha \ln(3/2)} = e^{2\pi i} = 1$ and $3^\alpha = 2^\alpha$.

On the other hand, if $\alpha, \beta \in \mathbb{C}$ and $\alpha \neq 0$, there are at most two primes p such that $p^\alpha = \beta$. To see this, suppose that p, q, r are distinct primes and $p^\alpha = q^\alpha = r^\alpha = \beta$. Then $(p/q)^\alpha = (p/r)^\alpha = 1$. Taking logarithms, $\alpha \ln(p/q) = 2k\pi i$ and $\alpha \ln(p/r) = 2l\pi i$ with k and l integers. The equation $\alpha l \ln(p/q) = \alpha k \ln(p/r)$ quickly leads to $p^l r^k = p^k q^l$ and the fundamental theorem of arithmetic implies that $k = l = 0$. But in that case, $\alpha \ln(p/q) = \alpha \ln(p/r) = 0$, an impossibility, since $\alpha \neq 0$ and p, q, r are distinct.

We are now in a position to state our main result on independence. Berberian [2] proved that ε and ε_1 are independent. But $\varepsilon = \varepsilon_0\lambda_1$ and $\varepsilon_1 = \varepsilon_1\lambda_1$ are just two members of the following uncountable independent set.

THEOREM 2. *The functions $\{\varepsilon_\alpha\lambda_\beta \mid \alpha, \beta \in \mathbb{C}, \beta \neq 0\}$ are linearly independent.*

Proof. Proceed as in the proof of Proposition 3. The equation $C_i(p) = C_j(p)$, $i \neq j$, reduces to $p^{\alpha_i-\alpha_j} = \beta_j/\beta_i$. When $\alpha_i = \alpha_j$ there are no solutions, since we can't also have $\beta_i = \beta_j$. When $\alpha_i \neq \alpha_j$ there are at most two solutions, as we have just seen. Once again, we only need to avoid a finite number of primes to find one that makes all the $C_i(p)$ distinct, so Lemma 1 again implies that all the exponents are zero. ∎

The functions gcd(m, ·) As a final application we consider some functions involving the greatest common divisor. For $m, n \in \mathbb{P}$ let $G_m(n) = \gcd(m, n)$. Each G_m is multiplicative. These functions are not independent: $G_2 * G_3 = G_1 * G_6$, for example. Bell series allow us to uncover such relations.

PROPOSITION 4. *Let (a, b) and $[a, b]$ denote the gcd and lcm of a and b.*

1. *$G_a * G_b = G_{(a,b)} * G_{[a,b]}$ for all $a, b \in \mathbb{P}$.*
2. *If $(a, b) = 1$ then $G_{ab} = \mu * G_a * G_b$.*
3. *If $m = m_1 m_2 \cdots m_r$ and the m_i are pairwise relatively prime then*

$$G_m = \mu^{(r-1)} * G_{m_1} * \cdots * G_{m_r}.$$

4. *If the prime factorization of m is $p_1^{k_1} \cdots p_r^{k_r}$ then*

$$G_m = \mu^{(r-1)} * G_{p_1^{k_1}} * \cdots * G_{p_r^{k_r}}.$$

Proof. First note that if h is the largest integer such that $p^h \mid m$, then

$$(G_m)_p(X) = 1 + pX + p^2X^2 + \cdots + p^{h-1}X^{h-1} + p^hX^h + p^hX^{h+1} + p^hX^{h+2} + \cdots$$
$$= \frac{1 - p^hX^h}{1 - pX} + \frac{p^hX^h}{1 - X} = \frac{1 - X - (p^{h+1} - p^h)X^{h+1}}{(1 - pX)(1 - X)}.$$

Call this $g(p, h, X)$. Now suppose $a, b \in \mathbb{P}$. For a given prime p let p^h be the highest power of p dividing a and let p^k be the highest dividing b. Then $(G_a * G_b)_p(X) = g(p, h, X)g(p, k, X)$. The highest power dividing (a, b) is p^m, where $m = \min(h, k)$, and the highest dividing $[a, b]$ is p^M, with $M = \max(h, k)$. Here $(G_{(a,b)} * G_{[a,b]})_p(X) = g(p, m, X)g(p, M, X)$. But either $h \le k$, $m = h$, and $M = k$, or $k \le h$, $m = k$, and $M = h$. In either case $g(p, h, X)g(p, k, X) = g(p, m, X)g(p, M, X)$. Thus $(G_a * G_b)_p(X) = (G_{(a,b)} * G_{[a,b]})_p(X)$ for all p, and $G_a * G_b = G_{(a,b)} * G_{[a,b]}$.

Noting that $G_1 = \varepsilon$, we have $G_a * G_b = \varepsilon * G_{ab}$ when $(a, b) = 1$. This implies the second equation. The third equation comes from repeated application of the second, and the fourth is a special case of the third. ∎

As an example, $G_{12} * G_{18} = G_6 * G_{36} = \mu^{(2)} * G_2 * G_3 * G_4 * G_9$.

This naturally raises the question of independence. In fact it can be shown that the functions $\{G_{q^k} \mid q \text{ prime}, k \in \mathbb{P}\}$ are linearly independent. Noting that $(G_{q^k})_p(X) = 1/(1 - X)$ if $p \neq q$, and

$$(G_{p^k})_p(X) = \frac{1 + (p - 1)X + (p^2 - p)X^2 + \cdots + (p^k - p^{k-1})X^k}{1 - X},$$

we could eventually establish the assertion by applying Lemma 1 to the numerators of the latter expressions, as in the proof of Proposition 3. The details of the argument, though interesting, are lengthy enough to divert us from the focus of this article so we will not pursue this point.

Further investigation Interesting areas of research lie in many directions.

- A systematic survey of multiplicative functions would be in order, in which the Bell series of families of functions are used to study their dependence relations in the rational vector space U_M. An excellent source of such families is McCarthy [**5**].
- What about the group of functions we have termed antimultiplicative? We haven't said anything about these, leaving the topic for readers to pursue.
- Is U_1 actually a vector space over $\mathbb{R}$ or even $\mathbb{C}$? For completely multiplicative f the right side of the formula in Proposition 2 for $f^{(q)}$ can be evaluated if q is real or even complex. What problems arise when one tries to extend exponentiation to real or complex exponents?

The study of algebraic properties of the ring of arithmetical functions offers research opportunities at many levels. A *Mathematica* notebook that facilitates experimentation with these functions is available at the MAGAZINE website, `www.maa.org/pubs/mathmag.html`.

REFERENCES

1. Tom M. Apostol, *Introduction to Analytic Number Theory,* Springer-Verlag, New York, 1976.
2. S. K. Berberian, *Number-Theoretic Functions via Convolution Rings,* this MAGAZINE, **65**, 1992, 75–90.
3. László Fuchs, *Infinite Abelian Groups,* Academic Press, New York, 1970.

4. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers,* 5th ed., Clarendon Press, Oxford, 1979.
5. Paul J. McCarthy, *Introduction to Arithmetical Functions,* Springer-Verlag, New York, 1986.
6. Ivan Niven and Herbert S. Zuckerman, *An Introduction to the Theory of Numbers,* 4th ed., John Wiley & Sons, New York, 1980.
7. Kenneth H. Rosen, *Elementary Number Theory and Its Applications,* 4th ed., Addison Wesley Longman, Reading, Massachusetts, 2000.
8. James Stewart, *Calculus,* 5th ed., Brooks/Cole, Belmont, California, 2003.

Letter to the Editor: Sury on Binet

In "A Parent of Binet's Formula?," October 2004, B. Sury asks if there is "a more natural motivation explaining the polynomial identity"

$$\sum_{i\geq 0}(-1)^i\binom{n-i}{i}(XY)^i(X+Y)^{n-2i} = X^n + X^{n-1}Y + \cdots + XY^{n-1} + Y^n.$$

Here is a simple combinatorial explanation of this identity.

The $\binom{n-i}{i}$ term counts the ways to tile a strip of length n with i dominoes of length two and $n-2i$ squares of length one (since such a strip has $n-i$ tiles altogether, from which we choose i of them to be dominoes). Now imagine that we are tiling a strip of length n with squares and dominoes but our squares can be colored in $X+Y$ ways, say X of the colors are *light* and Y of colors are *dark*. Also we will allow both halves of our dominoes to be colored, but the left half is always given a light color and the right half is given a dark color. Thus each domino can be colored in XY ways. Hence the number of tilings with exactly i dominoes (and thus $n-2i$ squares) would be

$$\binom{n-i}{i}(XY)^i(X+Y)^{n-2i}.$$

The total number of tilings is the sum of the above expression over all values of i (to be nonzero, we must have $0 \leq i \leq n/2$).

The left side of the polynomial identity is the number of colored tilings with an even number of dominoes minus the number of colored tilings with an odd number of dominoes. I claim that this difference is "almost zero" since there is an easy way to change the parity of the number of dominoes in "practically every" colored tiling. Specifically, for any tiling, look for the first occurrence of either A) a colored domino or B) a light square followed by a dark square.

If the first such occurrence is a colored domino then chop that domino in half to produce a light square followed by a dark square, producing a tiling of type B. If the first such occurrence is of type B, then join the colored squares together to form a domino, thus creating a tiling of type A. Notice that when we go from A to B or from B to A, we change the parity of the number of dominoes. Thus practically every tiling of type A holds hands with a tiling of type B and vice versa.

What are the exceptions? Simply those tilings that have no dominoes and never have a light square followed by a dark square. Such tilings consist of i dark squares followed by $n-i$ light squares for some $0 \leq i \leq n$, which can be done Y^iX^{n-i} ways. In total, the number of tilings with no light-dark pattern is $X^n + X^{n-1}Y + \cdots + XY^{n-1} + Y^n$, as desired.

——ARTHUR T. BENJAMIN
HARVEY MUDD COLLEGE
CLAREMONT, CA 91711

Outwitting the Lying Oracle

ROBB T. KOETHER
JOHN K. OSOINACH, JR.
Hampden-Sydney College
Hampden-Sydney, VA 23943
rkoether@hsc.edu
josoinach@hsc.edu

Lamentations 2:14 "Your prophets have seen for you false and foolish visions; ... they have seen for you false and misleading oracles."
(*New American Standard Bible*)

At the Delphi Casino an oracle operates a table where gamblers place bets on coin flips. The gamblers win or lose the amounts they bet, depending on whether they correctly predict the outcomes of the coin flips. As you approach the table, the oracle says to you, "I know how the coin will land each time and I am willing to tell you, but I must warn you, I will try to win your bet by occasionally lying to you." This does not strike you as a very promising game, but after some negotiation, the oracle agrees to lie no more than once during the next three coin flips, provided that before each flip you first tell the oracle the amount of your wager.

The question is: How should you place your bets on the three coin tosses so that you win the greatest amount of money in the end, no matter what the oracle does and no matter what the coin tosses are? We assume that the oracle is always agreeable to any amount that you wish to wager, but you cannot wager more than you currently possess.

We first encountered this problem in an article in *Scientific American* [**3**]. A very similar, but more general, problem appeared as Problem 10801 in the *American Mathematical Monthly* [**2**], along with its solution [**1**]. We gave this problem as a "Problem of the Fortnight" at Hampden-Sydney College, where we assumed you began with \$100 and the coin was flipped three times. One student solved the problem in the following manner (slightly paraphrased):

"The greatest amount of money that you can be guaranteed to receive, regardless of what the oracle does and regardless of what the coin flips are, is \$200: You should bet \$50 on the first flip and agree with the oracle's prediction. If the oracle lies, then you will still have \$50 left, but will correctly guess the remaining two flips for \$200; if the oracle is truthful, then you will have \$150. On the next flip again bet \$50 and agree with the oracle's prediction. If the oracle lies, then you have \$100 with one flip remaining, which you will guess correctly for \$200; if the oracle is truthful, then you will still have \$200 and will bet \$0 on the final flip."

While this answer is not entirely rigorous, the key ideas are present: No matter what the oracle does and what the results of the tosses are, if you always agree with the oracle, you have a strategy that guarantees that you double your initial amount.

This answer raised some interesting questions: How does the solution change if we increase the number of flips and allow the oracle to lie more than once? Can you outwit the oracle by disagreeing with the oracle's prediction? Or, stated differently, is there a strategy by which you could expect to win more than the maximum *guaranteed* outcome? Furthermore, is it possible to use the size of the bet to influence the oracle either to lie or tell the truth?

In this article, we first analyze the original game, but with any number of flips, followed by a simple generalization where the oracle may lie more than once. We then investigate the problem of trying to outwit the oracle, that is, finding strategies that give you the best chance for a better expected outcome, if possible, as well as strategies that the oracle should employ to minimize your chances for a better outcome. The mathematics involves relatively straightforward applications of game theory and probability, leading to some interesting results.

Believing the oracle

Our initial strategy will be always to agree with the oracle's prediction and make our bets on the basis of that strategy. We will start by solving the basic problem, where the oracle may lie at most once, and then allow the oracle to lie multiple times.

Multiple flips, one lie We begin with a restatement of the basic problem.

The Lying Oracle Problem: The oracle agrees to flip the coin a specified number of times and to predict the outcome accurately, except for possibly one lie. Before each prediction, you may bet any amount up to your current holdings. The oracle will then announce the outcome, after which you must state the outcome on which you wish to bet. How should you place your bets for the coin tosses so that you win the greatest amount of money in the end, no matter what the oracle does and no matter what the coin tosses are?

Solution: Using the terminology of game theory, we will henceforth refer to you, the bettor, as "the player."

Let w_n represent the proportion of the player's current holdings that the player should wager when there are n flips remaining in order to optimize the final outcome, and let A_n be the ratio of the player's final winnings to the current holdings, when the player wagers the optimal amounts on the remaining n flips.

If the oracle tells the truth on the first of the remaining n flips, then the player has the proportion $1 + w_n$ of the player's current holdings. The player must continue to place bets cautiously since the oracle may still lie. Thus, the final proportion of the player's winnings would be $(1 + w_n)A_{n-1}$.

On the other hand, if the oracle lies, then the player has the proportion $1 - w_n$, but now the player is free to bet the maximum amount on all $n - 1$ remaining flips, thereby doubling the player's money each time. In that case, the final proportion would be $(1 - w_n)2^{n-1}$.

We wish to find the value of w_n that will make these two expressions equal, nullifying the effect of the oracle's lie. That is, we wish to find the value of w_n such that

$$(1 - w_n)2^{n-1} = (1 + w_n)A_{n-1}. \tag{1}$$

This common value will be the value of A_n. In particular,

$$A_n = (1 - w_n)2^{n-1}. \tag{2}$$

Now by applying the above reasoning again, we see that $A_{n-1} = (1 - w_{n-1})2^{n-2}$. Substituting this into (1) produces

$$(1 - w_n)2^{n-1} = (1 + w_n)(1 - w_{n-1})2^{n-2}.$$

Solving for w_n yields the recurrence relation

$$w_1 = 0,$$
$$w_n = \frac{1 + w_{n-1}}{3 - w_{n-1}}, \quad n \geq 2.$$

(The condition $w_1 = 0$ follows from the observation that with one flip and one lie, the player cannot guarantee a correct prediction; hence, the player should wager nothing.) An easy induction shows that the solution to this relation is

$$w_n = \frac{n-1}{n+1}, \quad n \geq 1. \tag{3}$$

A formula for A_n is obtained by substituting the expression for w_n given in equation (3) into equation (2). We summarize everything we have learned in the following theorem.

THEOREM 1. *If the oracle has not yet lied and there are n coin tosses remaining, then the player should bet* $(n-1)/(n+1)$ *of the player's current amount of money. However, if the oracle has lied and therefore cannot lie again, the player should bet everything. In either case, the player's final amount will be exactly* $2^n/(n+1)$ *times the player's current holdings.*

Multiple flips, multiple lies A natural generalization of this problem is to allow the oracle to lie more than once, so suppose that the oracle may lie up to k times during the coin flips. If the player's strategy is to continue to agree with the oracle's prediction, how should the player place the bets now so that the player again gets the greatest amount of money in the end?

In this case, we are dealing with a family of games, $G_{n,k}$, where $G_{n,k}$ represents the game of n coin flips and at most k lies. Thus, if the oracle tells the truth, then play proceeds to game $G_{n-1,k}$, and if the oracle lies, then play proceeds to game $G_{n-1,k-1}$. FIGURE 1 shows how the games proceed, beginning with 4 flips and 2 possible lies by the oracle. Note that game $G_{i,i}$ is always followed by game $G_{i-1,i-1}$, for all $i \geq 1$.

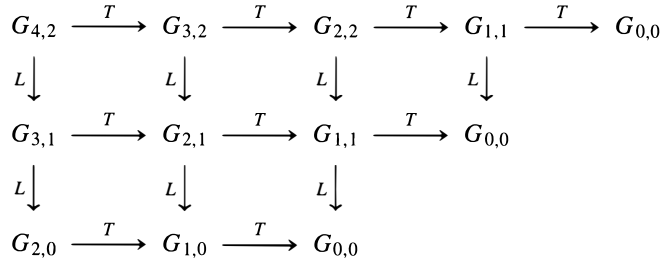


Figure 1 The game tree starting with 4 flips and 2 lies

Let $w_{n,k}$ represent the proportion of the player's current holdings that the player should wager in game $G_{n,k}$ in order to optimize the final winnings, and let $A_{n,k}$ represent the ratio of the final winnings to the player's current holdings, provided the player wagers the optimal amounts in game $G_{n,k}$ and all succeeding games. Let us call $w_{n,k}$ the *critical wager*.

If the player continues to believe the oracle's predictions, then whenever k equals n, the player should bet \$0 from that stage on, as the oracle could lie every time, giving

the player no opportunity to recover from a loss. Notice also that the games $G_{n,0}$ simply double the player's money with each coin toss, and that the games $G_{n,1}$ were analyzed above.

THEOREM 2. *For all $n \geq 1$ and for all k, $0 \leq k \leq n$, in the game $G_{n,k}$,*

$$A_{n,k} = \frac{2^n}{\sum_{i=0}^{k} \binom{n}{i}} \tag{4}$$

and

$$w_{n,k} = \frac{\binom{n-1}{k}}{\sum_{i=0}^{k} \binom{n}{i}}. \tag{5}$$

Proof. We will first establish a recurrence relation among the numbers $A_{n,k}$. Consider the first of n flips. If the oracle has told the truth, then the player would win $w_{n,k}$ of the player's current holdings on that flip. If the oracle has lied, then the player would lose the proportion $w_{n,k}$ on that flip. In the first case, the player's final winnings would be $(1 + w_{n,k})A_{n-1,k}$ and in the second case it would be $(1 - w_{n,k})A_{n-1,k-1}$.

In order to maximize the player's guaranteed winnings, these two amounts should be equal. Setting them equal and solving for $w_{n,k}$ yields

$$w_{n,k} = \frac{A_{n-1,k-1} - A_{n-1,k}}{A_{n-1,k-1} + A_{n-1,k}}. \tag{6}$$

It follows that

$$\begin{aligned} A_{n,k} = (1 + w_{n,k})A_{n-1,k} &= \left(1 + \left(\frac{A_{n-1,k-1} - A_{n-1,k}}{A_{n-1,k-1} + A_{n-1,k}}\right)\right) A_{n-1,k} \\ &= \frac{2A_{n-1,k-1}A_{n-1,k}}{A_{n-1,k-1} + A_{n-1,k}}. \end{aligned}$$

We see from this equation that $A_{n,k}$ is the harmonic mean of $A_{n-1,k-1}$ and $A_{n-1,k}$, that is,

$$\frac{1}{A_{n,k}} = \frac{1}{2}\left(\frac{1}{A_{n-1,k-1}} + \frac{1}{A_{n-1,k}}\right). \tag{7}$$

We will use equation (7) to establish (4) by induction.

First, it is clear that $w_{n,0} = 1$, since the player will bet the full amount if the player knows that the oracle will not lie, and that $w_{n,n} = 0$, since the player will bet nothing if the oracle cannot be counted on to tell the truth at least once. It follows that $A_{n,0} = 2^n$ and $A_{n,n} = 1$, for all $n \geq 1$.

Thus, equation (4) holds for all n when $k = 0$ or $k = n$. In particular, it holds for all k, $0 \leq k \leq n$, when $n = 0$ or $n = 1$. We proceed by induction on n. Let us assume that equation (4) is correct for all k, $0 \leq k \leq n$, for some $n \geq 1$, and consider $A_{n+1,k}$, for some k where $0 < k < n + 1$. We complete the induction by computing

$$\begin{aligned} \frac{1}{A_{n+1,k}} = \frac{1}{2}\left(\frac{1}{A_{n,k-1}} + \frac{1}{A_{n,k}}\right) &= \frac{1}{2}\left(\frac{\sum_{i=0}^{k-1}\binom{n}{i}}{2^n} + \frac{\sum_{i=0}^{k}\binom{n}{i}}{2^n}\right) \\ &= \frac{\sum_{i=0}^{k-1}\binom{n}{i} + \sum_{i=0}^{k}\binom{n}{i}}{2^{n+1}} = \frac{1 + \sum_{i=1}^{k}\left(\binom{n}{i-1} + \binom{n}{i}\right)}{2^{n+1}} \\ &= \frac{\sum_{i=0}^{k}\binom{n+1}{i}}{2^{n+1}}. \end{aligned}$$

Now the reader can easily use (6) to verify that

$$w_{n,k} = \frac{\binom{n-1}{k}}{\sum_{i=0}^{k}\binom{n}{i}}.$$

■

TABLES 1 and 2 give the values of $A_{n,k}$ and $w_{n,k}$ for $1 \le n \le 7$ and $0 \le k \le 6$. It is interesting to note that the same solution was obtained by Pudaite [**2**], where the assumption was equivalent to the oracle's lying *exactly* k times in n coin flips.

TABLE 1: Table of final winnings $A_{n,k}$

n \ k	0	1	2	3	4	5	6
1	2	1	1	1	1	1	1
2	4	$\frac{4}{3}$	1	1	1	1	1
3	8	$\frac{8}{4}$	$\frac{8}{7}$	1	1	1	1
4	16	$\frac{16}{5}$	$\frac{16}{11}$	$\frac{16}{15}$	1	1	1
5	32	$\frac{32}{6}$	$\frac{32}{16}$	$\frac{32}{26}$	$\frac{32}{31}$	1	1
6	64	$\frac{64}{7}$	$\frac{64}{22}$	$\frac{64}{42}$	$\frac{64}{57}$	$\frac{64}{63}$	1
7	128	$\frac{128}{8}$	$\frac{128}{29}$	$\frac{128}{64}$	$\frac{128}{99}$	$\frac{128}{120}$	$\frac{128}{127}$

TABLE 2: Table of critical wagers $w_{n,k}$

n \ k	0	1	2	3	4	5	6
1	1	0	0	0	0	0	0
2	1	$\frac{1}{3}$	0	0	0	0	0
3	1	$\frac{2}{4}$	$\frac{1}{7}$	0	0	0	0
4	1	$\frac{3}{5}$	$\frac{3}{11}$	$\frac{1}{15}$	0	0	0
5	1	$\frac{4}{6}$	$\frac{6}{16}$	$\frac{4}{26}$	$\frac{1}{31}$	0	0
6	1	$\frac{5}{7}$	$\frac{10}{22}$	$\frac{10}{42}$	$\frac{5}{57}$	$\frac{1}{63}$	0
7	1	$\frac{6}{8}$	$\frac{15}{29}$	$\frac{20}{64}$	$\frac{15}{99}$	$\frac{6}{120}$	$\frac{1}{127}$

Note that if the player does not bet the critical wager $w_{n,k}$ at each stage, then the oracle can follow a pure strategy that guarantees the player's final outcome to be less than if the player had bet the critical wager. If the player bets more than the critical wager, then the oracle will lie, reducing the amount the player has to start with for the next game by more than the critical wager. Likewise, if the player bets less than the critical wager, then the oracle will tell the truth, which will increase the amount the player has to start with for the next game by less than the critical wager. Thus, by betting an amount different from the critical wager, the player can induce the oracle to lie or be truthful, but always at a disadvantage to the player, provided the player continues to believe the oracle.

Outwitting the oracle

The above analysis makes two crucial assumptions: the player will always agree with the oracle, and the oracle knows that the player will always agree. These assumptions ensure that the player will never receive less than the guaranteed amount, no matter what the oracle does or how the coin is flipped, but they also guarantee that the player will never receive more than that amount. But what if the player suspects that the oracle is lying? Can the player expect to increase the final winnings by *not* agreeing with the oracle? Indeed, can the player induce the oracle to lie by betting a large amount, and then win that amount by disagreeing with the oracle? As we investigate this possibility, we will also assume that the oracle now suspects that the player may disagree.

A single flip Let's begin with a simple example. Suppose we have exactly one flip and the oracle has one lie. If the oracle knows that the player will always agree with the oracle's prediction, then the oracle will lie if the player bets any amount at all. However, if the player is unpredictable—the player may choose to disagree—is it to the player's advantage to bet some amount? Is there a strategy for betting a certain wager so that the player's *expected* payoff is more than the amount guaranteed by the previous analysis? After all, in this game the player following the previous strategy would bet nothing.

This game may be modeled by a simple two-by-two matrix, where the entries represent the payoffs for the player. The rows indicate the player's two strategies (Agree or Disagree), while the columns represent the oracle's two strategies (tell the Truth or tell a Lie). Hence, in this example, we have the following payoff matrix for the player:

$$\begin{array}{c} \\ \text{Agree} \\ \text{Disagree} \end{array} \begin{array}{c} \begin{array}{cc} \text{Truth} & \text{Lie} \end{array} \\ \begin{pmatrix} 1+w & 1-w \\ 1-w & 1+w \end{pmatrix} \end{array},$$

where w is the proportion wagered (whether optimal or not). Let p_T and p_L be the probabilities that the oracle will tell the truth or lie, respectively. Similarly, let p_A and p_D be the probabilities that the player will agree or disagree with the oracle, respectively. In order to decide which strategy to pursue, the player computes the expected payoff of each row of the payoff matrix; the player then chooses the strategy (row) whose expected payoff is the greater of the two. The player's expected payoff of agreeing with the oracle is $p_T(1+w) + p_L(1-w)$; likewise, the expected payoff of disagreeing with the oracle is $p_T(1-w) + p_L(1+w)$.

On the other hand, the oracle's optimal strategy occurs when these two expected payoffs are equal. Setting the expected payoffs from the two rows equal yields

$$p_T(1+w) + p_L(1-w) = p_T(1-w) + p_L(1+w).$$

Solving for p_T and p_L gives $p_T = p_L$ (assuming that $w > 0$). Since $p_T + p_L = 1$, we have that $p_T = p_L = 1/2$. Substituting these values into the player's expected payoff from the first row gives us an expected payoff of 1. A similar calculation yields $p_A = p_D = 1/2$. Thus, the player cannot expect to do any better in this case than in the original scenario.

Multiple flips A similar analysis works in general. Let $E_{n,k}$ denote the expected payoff of the game $G_{n,k}$ when the player and the oracle employ their optimal strategies. In this game, the payoff matrix is

$$\begin{array}{c} \\ \text{Agree} \\ \text{Disagree} \end{array} \begin{array}{c} \begin{array}{cc} \text{Truth} & \text{Lie} \end{array} \\ \begin{pmatrix} (1+w)E_{n-1,k} & (1-w)E_{n-1,k-1} \\ (1-w)E_{n-1,k} & (1+w)E_{n-1,k-1} \end{pmatrix} \end{array},$$

where w is the proportion of the wager. The strategies adopted by the players depend, of course, on the values of w, $E_{n-1,k}$, and $E_{n-1,k-1}$.

LEMMA 1. *In the game $G_{n,k}$, for all $n \geq 1$ and for all k, $1 \leq k \leq n$,*

$$E_{n,0} = 2^n, \tag{8}$$

$$E_{n,n} = 1, \tag{9}$$

$$E_{n,k} < E_{n,k-1}, \tag{10}$$

$$E_{n,k} = \frac{2E_{n-1,k-1}E_{n-1,k}}{E_{n-1,k-1} + E_{n-1,k}}. \tag{11}$$

Proof. We will establish (8) and (9) first. The game $G_{1,0}$ is trivial. The oracle must tell the truth and the player will agree. Therefore, $E_{1,0} = 2$. Notice also that we have already analyzed the game $G_{1,1}$ and found that $E_{1,1} = 1$.

In the game $G_{n,0}$, the oracle must always tell the truth, which gives the player a pure strategy of agreeing with the oracle each time. Thus, the player's optimal strategy is to wager the entire amount and will therefore double the amount wagered each time. Hence we have $E_{n,0} = 2^n$.

The game $G_{n,n}$ has payoff matrix

$$\begin{array}{c} \\ \text{Agree} \\ \text{Disagree} \end{array} \begin{array}{c} \begin{array}{cc} \text{Truth} & \text{Lie} \end{array} \\ \begin{pmatrix} (1+w)E_{n-1,n-1} & (1-w)E_{n-1,n-1} \\ (1-w)E_{n-1,n-1} & (1+w)E_{n-1,n-1} \end{pmatrix} \end{array}.$$

(Recall that whether or not the oracle lies, the next game is $G_{n-1,n-1}$.) Again, a straightforward analysis shows that $E_{n,n} = E_{n-1,n-1}$ and it follows that $E_{n,n} = 1$ for all $n \geq 1$.

We now establish parts (10) and (11) of the lemma. First, note that we have already shown that $E_{1,1} < E_{1,0}$. Also, if we define $E_{0,1} = E_{0,0} = 1$, then we see that

$$E_{1,1} = \frac{2E_{0,0}E_{0,1}}{E_{0,0} + E_{0,1}} = 1.$$

We will proceed by induction on n. Suppose that (10) and (11) hold for some $n \geq 1$ and for all k, $1 \leq k \leq n$. Consider, for some such k, the payoff matrix of game $G_{n+1,k}$:

$$\begin{array}{c} \\ \text{Agree} \\ \text{Disagree} \end{array} \begin{array}{c} \begin{array}{cc} \text{Truth} & \text{Lie} \end{array} \\ \begin{pmatrix} (1+w)E_{n,k} & (1-w)E_{n,k-1} \\ (1-w)E_{n,k} & (1+w)E_{n,k-1} \end{pmatrix} \end{array}.$$

We say that a row is a *dominated row* if its entries are never greater than the corresponding entries of the other row in the payoff matrix. On the other hand, we say that a column is a *dominated column* if its entries are never less than the corresponding entries of the other column in the payoff matrix. This difference reflects the fact that the oracle's goal is to reduce the amount that the player wins; hence, the oracle always seeks the smallest possible payoff for the player.

Clearly, neither row is dominated by the other (assuming that $w > 0$). It is also clear, from the assumption that $E_{n,k} < E_{n,k-1}$, that column 1 cannot be dominated by column 2. However, column 2 will be dominated by column 1 if

$$(1 + w)E_{n,k} \leq (1 - w)E_{n,k-1}.$$

This occurs when

$$w \leq \frac{E_{n,k-1} - E_{n,k}}{E_{n,k-1} + E_{n,k}}. \tag{12}$$

In this case, the oracle has a pure strategy: always tell the truth, in which case the player also has a pure strategy: always agree. This produces a payoff of $(1 + w)E_{n,k}$. Subject to the inequality (12), this expression reaches a maximum value of

$$\frac{2E_{n,k-1}E_{n,k}}{E_{n,k-1} + E_{n,k}}$$

when

$$w = \frac{E_{n,k-1} - E_{n,k}}{E_{n,k-1} + E_{n,k}}.$$

On the other hand, when $w > (E_{n,k-1} - E_{n,k})/(E_{n,k-1} + E_{n,k})$, neither column is dominated by the other, in which case the oracle has a mixed strategy. The oracle's optimal strategy (p_T, p_L) will make the expected payoff of row 1 equal to the expected payoff of row 2. That is,

$$p_T(1 + w)E_{n,k} + p_L(1 - w)E_{n,k-1} = p_T(1 - w)E_{n,k} + p_L(1 + w)E_{n,k-1}. \tag{13}$$

This simplifies, since $w > 0$, to

$$p_T E_{n,k} = p_L E_{n,k-1}.$$

Using the fact that $p_L = 1 - p_T$, we may solve for p_T and p_L:

$$p_T = \frac{E_{n,k-1}}{E_{n,k-1} + E_{n,k}}, \tag{14a}$$

$$p_L = \frac{E_{n,k}}{E_{n,k-1} + E_{n,k}}. \tag{14b}$$

Now, by substituting these expressions into either side of (13), we compute the expected payoff to be

$$\frac{2E_{n,k-1}E_{n,k}}{E_{n,k-1} + E_{n,k}}.$$

This establishes that the optimal payoff occurs when

$$w \geq \frac{E_{n,k-1} - E_{n,k}}{E_{n,k-1} + E_{n,k}},$$

in which case the oracle will utilize the optimal strategy given by (14a) and (14b). Thus

$$E_{n+1,k} = \frac{2E_{n,k-1}E_{n,k}}{E_{n,k-1} + E_{n,k}}, \tag{15}$$

which establishes (11).

As was remarked in (7), equation (15) implies that $E_{n+1,k}$ is the *harmonic mean* of $E_{n,k-1}$ and $E_{n,k}$; that is,

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

Therefore,

$$E_{n,k} < E_{n+1,k} < E_{n,k-1}.$$

Since these inequalities hold for all k, $1 \leq k \leq n$, it follows that

$$1 = E_{n,n} < E_{n+1,n} < E_{n,n-1} < \cdots < E_{n,1} < E_{n+1,1} < E_{n,0} = 2^n,$$

establishing that

$$E_{n+1,k} < E_{n+1,k-1}$$

for all k, $2 \leq k \leq n$. As special cases, we have already shown that $E_{n+1,n+1} = 1$ and $E_{n+1,0} = 2^{n+1}$, so we may conclude that inequality (10) of the lemma holds in general. ∎

COROLLARY 1. *If the player and the oracle follow their optimal strategies in the game $G_{n,k}$, then an optimal wager is any amount w such that*

$$\frac{E_{n-1,k-1} - E_{n-1,k}}{E_{n-1,k-1} + E_{n-1,k}} \leq w \leq 1.$$

As before, we will call the value

$$w_{n,k} = \frac{E_{n-1,k-1} - E_{n-1,k}}{E_{n-1,k-1} + E_{n-1,k}} \tag{16}$$

the *critical wager.*

COROLLARY 2. *Let w be the amount of the wager in the game $G_{n,k}$. Then the oracle's optimal strategy is given by*

$$p_T = \begin{cases} 1 & \text{if } 0 \leq w \leq w_{n,k} \\ \frac{1}{2} + \frac{1}{2}w_{n,k} & \text{if } w_{n,k} < w \leq 1 \end{cases} \tag{17}$$

and the player's optimal strategy is given by

$$p_A = \begin{cases} 1 & \text{if } 0 \leq w \leq w_{n,k} \\ \frac{1}{2} + \frac{1}{2}\left(\frac{w_{n,k}}{w}\right) & \text{if } w_{n,k} < w \leq 1. \end{cases}$$

Proof. By using formulae (14a), (14b), and (16), we see that

$$p_T - p_L = w_{n,k}$$

when the oracle has a mixed strategy, from which the formula for the oracle's strategy follows. We will now compute the player's optimal strategy (p_A, p_D) for the game $G_{n,k}$. This strategy occurs when the expected values of the two columns of the payoff matrix are equal, giving the equation

$$p_A(1+w)E_{n-1,k} + p_D(1-w)E_{n-1,k} = p_A(1-w)E_{n-1,k-1} + p_D(1+w)E_{n-1,k-1}.$$

This simplifies to

$$w(p_A - p_D) = w_{n,k},$$

from which the formula for the player's strategy follows. ∎

If the player bets any amount up to the critical wager $w_{n,k}$, then Corollary 2 prescribes a pure strategy: always agree. On the other hand, it is now rational for the player to bet more than the critical wager. Indeed, it is rational for the player to bet even the full amount ($w = 1$), provided the player is willing to disagree with the oracle occasionally. We will pursue this possibility further in the next section.

It is interesting to note that, if the player bets more than the critical wager $w_{n,k}$, then the oracle's mixed strategy is *not* dependent on the size of the wager, even though the player's mixed strategy is.

THEOREM 3. *For all $n \geq 1$ and for all k, $0 \leq k \leq n$, in the game $G_{n,k}$,*

$$E_{n,k} = \frac{2^n}{\sum_{i=0}^{k} \binom{n}{i}}$$

and

$$w_{n,k} = \frac{\binom{n-1}{k}}{\sum_{i=0}^{k} \binom{n}{i}}.$$

Proof. Lemma 1 establishes the same recurrence relation for $E_{n,k}$ that was earlier established for $A_{n,k}$. Thus, the solution for $E_{n,k}$ is the same as the solution for $A_{n,k}$. Furthermore, equation (16) is of the same form as equation (6), so $w_{n,k}$ will be the same as before. ∎

We see that the expected payoff in this case is the same as the guaranteed payoff in the earlier case where the player always agreed with the oracle. Therefore, *you can't outwit the oracle (in the long run) by disagreeing with the oracle!* You might as well agree with the oracle, even though you know the oracle might lie.

Probability of a given sequence

We have now seen that by betting a sufficiently large amount and occasionally disagreeing with the oracle, we can induce the oracle to follow a predictable mixed strategy; that is, the oracle will tell the truth with known probability p_T. That makes it possible to calculate the probability of any particular sequence of truths and lies.

Note the significance of the denominator in Theorems 2 and 3. The term $\binom{n}{i}$ represents the number of ways in which the oracle can lie *exactly* i times with n flips

remaining. Hence, the denominator $\sum_{i=0}^{k} \binom{n}{i}$ represents the total number of ways in which the oracle can lie with n flips and up to k lies. It turns out, as shown in the following theorem, that these different sequences of truths and lies are all equally likely. This seems reasonable, since this gives the player the least amount of information on which to choose whether to agree or disagree.

THEOREM 4. *Beginning with game $G_{n,k}$, the probability of any given sequence of truths and lies for the n coin tosses is* $1/\sum_{i=0}^{k} \binom{n}{i}$.

Proof. Let T and L represent truths and lies, respectively, in a sequence of flips. We proceed by induction. In the game $G_{1,1}$ there are only two possible sequences: T or L. As we have already seen, the probability of each is $1/2$. Now suppose that for some $n \geq 1$, the likelihood of any particular sequence of truths and lies beginning with the game $G_{n,k}$ is $1/\sum_{i=0}^{k} \binom{n}{i}$, for all $k, 0 \leq k \leq n$. Consider a sequence beginning with the game $G_{n+1,k}$ for some $k, 0 \leq k \leq n+1$. The first term of the sequence is either T or L. By substituting the expressions in Theorem 3 into formula (17) and simplifying, we find that the probability that the first term is T is

$$p_T = \frac{\sum_{i=0}^{k} \binom{n}{i}}{\sum_{i=0}^{k} \binom{n+1}{i}},$$

and the probability that the first term is L is

$$p_L = \frac{\sum_{i=0}^{k-1} \binom{n}{i}}{\sum_{i=0}^{k} \binom{n+1}{i}}.$$

By hypothesis, the probability of the remaining n terms of the sequence is either $1/\sum_{i=0}^{k} \binom{n}{i}$ or $1/\sum_{i=0}^{k-1} \binom{n}{i}$, depending on the number of lies remaining. Therefore, if the first term is T, then the probability of the full sequence is

$$p_T\left(\frac{1}{\sum_{i=0}^{k} \binom{n}{i}}\right) = \frac{1}{\sum_{i=0}^{k} \binom{n+1}{i}}$$

and if the first term is L, then the probability of the full sequence is

$$p_L\left(\frac{1}{\sum_{i=0}^{k-1} \binom{n}{i}}\right) = \frac{1}{\sum_{i=0}^{k} \binom{n+1}{i}}.$$

Thus, regardless of the first term, the probability of every sequence beginning with game $G_{n+1,k}$ is $1/\sum_{i=0}^{k} \binom{n+1}{i}$. This completes the induction. ∎

The significance of the equation

$$E_{n,k} = \frac{2^n}{\sum_{i=0}^{k} \binom{n}{i}}$$

now becomes more apparent. The expected payoff of game $G_{n,k}$ does not depend on the size of the wager w, provided $w \geq w_{n,k}$. Therefore, consider the simple case where $w = 1$ in every game. If the oracle tells so much as a single lie, then the player loses everything, ending up with \$0. However, if the player succeeds (by chance) in outwitting the oracle every time, then the player ends up with $\$2^n$. This happens with

probability $1/\sum_{i=0}^{k}\binom{n}{i}$. Therefore, the expected payoff is

$$E_{n,k} = \frac{2^n}{\sum_{i=0}^{k}\binom{n}{i}},$$

just as we calculated earlier.

Final thoughts

Playing the game with the lying oracle is best suited for those who are averse to risk. After all, if you try to outwit the oracle, you can't expect to do any better than if you simply believe the oracle each time. Furthermore, by betting large sums of money you cannot tempt the oracle into trying to outwit you, provided the oracle suspects that you may disagree. Indeed, the oracle simply mixes the predictions with lies and truths in a fixed fashion, aloof to the amount you bet, unless you are too cautious with your bet.

Listed below are some variants of the game which may make for some interesting further investigation. In each case, what is the player's optimal strategy and expected payoff?

- In the above analysis, the oracle need not lie at all during the course of the coin flips. Suppose there is a minimum number of lies that the oracle must tell.
- The oracle might also require you to place *all* your wagers before the first coin flip, expressed as a proportion of the amount you'd have before each coin flip.
- Similarly, the oracle might require you to place all your wagers before the first coin flip, but expressed as *absolute* amounts. If your holdings ever drop below your next wager, then you lose everything [**3**].
- Suppose the oracle improves the payoff for guessing the coin flip correctly (say, a correct guess pays 3:1). (See the editorial comment in [**1**].)
- What if the probability distribution of the possible outcomes isn't uniform (say, the coin is weighted)?
- What if, instead of a coin, the oracle uses a die (or any other object where the number of possible outcomes is greater than 2)?
- Consider a k-gullible oracle, that is, an oracle that continues to believe that the player will agree until the player has disagreed k times. From that point on, the oracle suspects that the player may disagree.

Acknowledgment. We would like to thank the referees for their excellent suggestions, which led to many improvements of this paper. We would also like to thank Bill Gasarch for drawing our attention to Problem 10801 in the *Monthly*.

REFERENCES

1. Owen Byer and Deirdre Smeltzer, A gambler urns his money, *Amer. Math. Monthly* **109** (2002), 394–395.
2. Paul R. Pudaite, Problem 10801, *Amer. Math. Monthly* **107** (2000), 368.
3. Dennis E. Shasha, The Delphi flip, *Scientific American* **285**:2 (2001), 94.

Twentieth-Century Gems from MATHEMATICS MAGAZINE

GERALD L. ALEXANDERSON
PETER ROSS
Santa Clara University
Santa Clara, CA 95053-0290
galexand@math.scu.edu
pross@scu.edu

Would we turn to MATHEMATICS MAGAZINE today for a seminal research paper in analysis by one of the current giants of mathematics? Perhaps not. It would have been a good idea in 1948, though, when Marshall H. Stone published in the MAGAZINE [**10**] what came to be known as the general Stone-Weierstrass theorem.

Though this may be the most striking example of a new and important result to appear first in MATHEMATICS MAGAZINE, there are many other examples of interesting articles written for the MAGAZINE by eminent mathematicians. The MAGAZINE has also been a rich resource of gems by less well-known mathematicians, and its backlog can be mined for all sorts of purposes. In the following pages we'll sample some of the items that we found attractive and interesting. They are highly personal choices, of course. The emphasis will be on the earlier years of the MAGAZINE, in part because many readers will be familiar with more recent material. In addition, the first author was Editor of the MAGAZINE from 1986 to 1990 and finds it too difficult to choose among his children!

We begin with a brief history of the MAGAZINE, whose past is more varied than that of most journals. Continuing the historical theme, we note some of the gems that established the history of mathematics as one strength of the MAGAZINE'S offerings. We then show what can be learned from these pages about 20th-century mathematical culture, including the culture of mathematical awards. The longest part of the article consists of collections of gems by mathematical theme: algebra, calculus & analysis, combinatorics, games & puzzles, geometry, number theory, probability, and teaching & pedagogy. We close with a brief account of the Problems section, with its own distinctive cast of characters.

Several Proofs Without Words and book reviews are discussed in appropriate sections. Roger Nelsen has published two collections of selected "Proofs Without Words" [**7, 8**], most of which are taken from the MAGAZINE. For the reader left wanting more, the backlog of the MAGAZINE itself is the place to seek more gems. Readers may wish to consult the *Mathematics Magazine/College Mathematics Journal* database at `http://www.math.hmc.edu/journalsearch/`. This free, easy-to-search database contains the title, author, and either a summary or the first paragraph of almost every article, note, and proof without words that has appeared in the MAGAZINE (and the *College Mathematics Journal*).

The history of the MAGAZINE

The earliest history of the MAGAZINE (first called the *Mathematics News Letter*, then the *National Mathematics Magazine*) is rather obscure. It started as a series of pamphlets published by the Louisiana-Mississippi Section of the MAA in 1926 and continued there under the editorship of Samuel T. Sanders of Louisiana State University

until 1945. (The first subscription rate was 50 cents a year, for ten issues, "the price of two good movie entertainments.") Sanders himself was the most prolific contributor in the early years, although some of his contributions were pleas for more articles on mathematics.

In an amusing early article, "The angle trisection chimera once more," **6**:3 (1931/32), 1–6, Sanders destroys an angle-trisection "proof" given by The Very Reverend J. J. Callahan, President of Duquesne University. (Since almost all references are to articles in the MAGAZINE, we shall use this format to give precise references to these papers instead of having an extensive list of references at the end.) Fr. Callahan's error was not one of the more subtle ones; he merely showed how with straightedge and compass one could construct an angle that is three times as large as a given angle!

Sanders includes an explanation, given by Tobias Dantzig in his classic *Number, the Language of Science*, of why the Greeks chose such limiting tools for their constructions: "The impossibility of the classic problem was imposed by a restriction which was so old as to be considered natural, so natural, indeed, that it was rarely mentioned. When the Greek spoke of a geometrical construction, he meant a construction by straight-edge and compasses. These were the instruments of the gods; all others were banned as unworthy of the speculation of the philosopher. For Greek philosophy, we must remember, was essentially aristocratic. The methods of the artisan, ingenious and elegant though they may have been, were regarded as vulgar and banal, and general contempt attached to all those who used their knowledge for gainful ends . . . ."

The very earliest years of the MAGAZINE were lean on mathematical content, with articles focusing instead on high school algebra, commercial arithmetic, pedagogy, and so on. A pleasant exception is a note by Zena Garrett of Mississippi Delta State Teachers College on "Perfect Numbers," **3**:6 (1928/29), 17–19. She describes her introduction to perfect numbers—integers that equal the sum of their proper divisors—in a class where students were given the first two, 6 and 28, and were told to find the next one by themselves, without using the library. She gives a charming description of her investigation and eventual success, using finite geometric series, in finding 496 and 8128, and then the general formula $2^{n-1}(2^n - 1)$ with $2^n - 1$ a prime.

The editor ran an experiment by distributing copies of the 1959 issues of the MAGAZINE through newsstands. He reported that in this post-Sputnik era they sold 40% of the copies distributed.

The management of MATHEMATICS MAGAZINE was taken over by the MAA in 1961 after a new editor had been appointed in 1960, Robert E. Horton. At that time the MAA's Board of Governors decided that the volume year of the MAGAZINE should coincide with the calendar year, not the academic year, a change evident in the references to articles below.

Rufus Isaac's one-page article, "Two Mathematical Papers without Words," **48** (1975), 198, consisted solely of the two images shown in FIGURE 1.

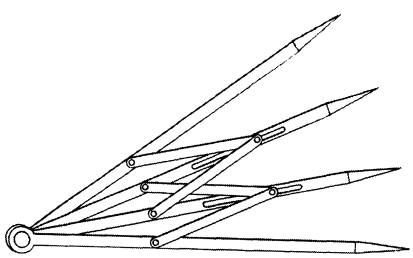
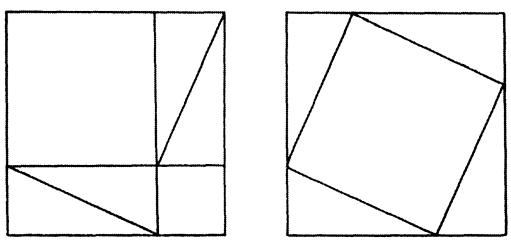


Figure 1 On trisecting an angle



A proof of the Pythagorean theorem

A reader was quick to point out that the second "proof by words" appeared as "A 'look-see' proof" in Martin Gardner's Mathematical Games column of *Scientific American* in October 1964, and was in fact millennia old. In January of 1976, MAGAZINE editors Seebach and Steen encouraged further contributions of these newly-named "proofs without words," and they have since become popular staples of the MAGAZINE as well as other publications like the *College Mathematics Journal*.

The MAGAZINE on the history of mathematics

The history of mathematics has always played a prominent role in the MAGAZINE. Earlier articles on it were broader and aimed more at the general reader, while in the last half-century articles on history have tended to be more specialized and technical.

• E. T. Bell's *Men of Mathematics*, published in 1937 and still an enormously popular book, has no doubt inspired many to study higher mathematics. In recent years many have criticized the book for its historical inaccuracies. That has not, however, discouraged its fans. The MAGAZINE review, by G. Waldo Dunnington, **11** (1936/37), 406–7, should have alerted later critics to possible historical shortcomings: "Dr. Bell is a seasoned, skillful writer with a fluent style; he writes with a realistic, curt, potent wit ... [the] careful reader will find a considerable number of examples of ... 'exaggeration.' ... [Bell] frequently recounts the charming legends, interesting traditions, and melodramatic fictions concerning the various mathematicians, but fortunately he usually discounts them properly by elucidating that they may be apocryphal anecdotes, rather than fact ... Dr. Bell allows his imagination to play; conjecture, personal opinion and speculation are abundant."

• A historical gem is E. T. Bell's two-part article "Gauss and the early development of algebraic numbers," **18** (1943/44), 188–204, 219–33. Bell surveys Gauss's contributions to many areas such as proofs of the fundamental theorems of arithmetic and algebra, elliptic functions (before Abel and Jacobi), reciprocity theorems, Gauss's "epochal achievement in introducing algebraic integers into arithmetic," and his ambivalent attitude towards Fermat's last theorem. Regarding the latter, Bell notes that in 1808 "Gauss had stopped, baffled, before his special case $n = 7$ of Fermat's last theorem." Eight years later Gauss asserted in a letter, "the Fermat theorem as an isolated proposition has little interest for me, since a multitude of such propositions, which one can neither prove nor refute, can be easily promulgated." However, later in the same letter Gauss says that if a "lucky star" prevails in allowing him a "*great* extension of the higher arithmetic", then "also the Fermat theorem will thereby appear as only one of the least interesting corollaries." Needless to say, the lucky star never appeared. Bell, as in his popular books, writes both pithily and with insight. For example, in contrasting operations with complex numbers (a field) with operations with algebraic integers (only a ring), Bell observes, "Generally the distinction between algebra and arithmetic has been roughly summarized in the dictum that division is only exceptionally impossible in algebra and only exceptionally possible in arithmetic." Bell's comments are occasionally provocative, for example, "Considered objectively, biography is the meanest form of gossip."

• Readers interested in the history of mathematics may want to look at the unusual article "Mathematics and mathematicians from Abel to Zermelo," **26** (1952/53), 127–46, by Einar Hille, the distinguished analyst at Yale, who became President of the American Mathematical Society in 1947–48. Viewing mathematics as a function of several variables, two of which are time and field of research, Hille presented two cross-sections in the article. The first, "Who was who in mathematics in 1852?",

focuses on mathematicians in France, Germany, and Great Britain (including Ireland), since at that time "outside of these countries research mathematicians were few and far between." The second cross-section is on "The development of analysis, particularly complex function theory, until the time of the first world war." Here the contrast between the approach of Weierstrass (1815–97) and that of Riemann (1826–66) is striking: "Weierstrass had the local point of view, Riemann the global one." Hille observes that the character of their work was also fundamentally different: "Weierstrass finished what he started," while mathematicians have been working on Riemann's ideas and trying to prove his conjectures since the mid-19th century.

• In volume **35** (1962), 153–54, Underwood Dudley, building on some data collected by Augustus De Morgan and others, tabulated 45 values of π calculated to five decimal places by various mid-19th century calculators. He then examined π_t, defined to be the ratio of the circumference of a circle to its diameter at time t, where t is taken between 1832 and 1879. He constructed the least squares linear function $\pi_t = .0000056060t + 3.14281$ where t is measured in years C.E., and, extrapolating, found that $\pi_{1962} = 3.15381$. He also noted that the Biblical value of $\pi_t = 3$ was excellent for its time and that calculations will be easier in June of 10,201 when π_t will be 3.20000. Further, π_t equaled 3.1415926535 . . . on November 10, 219 B.C.E., at around 10:54 in the evening. Dudley goes on to calculate the date of creation, 560,615 B.C.E., which agrees "neither with astronomical theory nor with Archbishop Ussher's chronology," and concludes, "Clearly more research is needed."

• The influential mathematician and historian B. L. van der Waerden became a MAGAZINE author with his "Hamilton's discovery of quaternions," **49** (1976), 227–34. Quoting extensively from Hamilton's mathematical papers and letters, van der Waerden traces Hamilton's train of thought, first through his attempts to multiply triplets (a, b, c) and (x, y, z) so that the "law of the moduli holds,"

$$(a^2 + b^2 + c^2)(x^2 + y^2 + z^2) = u^2 + v^2 + w^2,$$

and finally to Hamilton's "lightning stroke" of continuing to four dimensions. This last metaphor is apt; in a letter to his son describing the famous walk on which he carved the quaternion formulas on Brougham Bridge in Dublin, Hamilton wrote, "An electric current seemed to close; and a spark flashed forth . . . ." Van der Waerden observes that Hamilton would have quickly given up his search to multiply triplets if he had read Legendre's great work *Théorie des nombres*, since in it Legendre remarks that 3 and 21 are sums of three squares, $3 = 1 + 1 + 1$ and $21 = 1 + 4 + 16$, but their product 63 is not since it is of the form $8n + 7$. (Square numbers modulo 8 are either 0, 1, or 4, so it is easily checked that $u^2 + v^2 + w^2$ is not congruent to 7 modulo 8.)

• In 1986, Israel Kleiner published an extraordinary survey of the history of groups, "The evolution of group theory," **59** (1986), 195–213. It is nicely illustrated and gives, in addition to an illuminating text, a schematic showing the various stages of group theory from its sources and its specialized theories of permutation groups, abelian groups, and transformation groups, to the "divergence" in the 1920s and beyond: finite group theory, combinatorial group theory, topological group theory, and so on. It's one of a series of insightful historical articles by Kleiner in the 80s and 90s.

Mathematical culture

We present some glimpses of the culture of mathematics, which collectively illustrate that portraying mathematical culture has often been a significant feature of the MAGAZINE.

• Anyone who has suffered from having an obviously worthy manuscript rejected by an editor should take comfort in an article by Arnold Emch of the University of Illinois **11** (1936/37), 186–89, entitled "Rejected papers of three famous mathematicians." After stating that merit should be the only criterion for deciding whether to publish, adding that "there is no other science in which this assertion should appear more evident than in mathematics," Emch describes the sorry vicissitudes of three papers by Ludwig Schläfli, Bernhard Riemann, and Ernest de Jonquières. Each paper was published more than a quarter-century after it was submitted.

• George Bergman, now a distinguished professor at Berkeley but then a 12-year old student at Junior High School 246 in Brooklyn, published in the MAGAZINE, **31** (1957/58), 98–110, a 13-page piece on a number system with the irrational base φ, the golden mean. This resulted in George's being interviewed by Mike Wallace (of *60 Minutes* fame) for the *New York Post*. This interview was reprinted in the MAGAZINE, **31** (1957/58), 282. At the end of the interview Wallace asked: "George, when you wake up in the morning, what's the first thing you think about? Mathematics?" The student replied: "Oh, don't be silly. I think about breakfast."

Cecil Rousseau, in "The Phi Number System Revisited," **68** (1995), 283–84, observed that since Bergman's 1957 article, "the *phi number system* has become part of the folklore of elementary mathematics and has, for example, appeared as an exercise in Knuth's *The Art of Computer Programming* [**3**]. One of the basic results involving the phi number system is that every positive integer has a finite expansion."

Awards and the MAGAZINE The Chauvenet Prizes are the most prestigious and venerable of the awards given by the MAA for expository writing. Awarded since 1925, they had usually been awarded for papers in the *Monthly* or the *Bulletin of the American Mathematical Society*, but in 1944 the Chauvenet Prize was given to R. H. Cameron for a MAGAZINE article on Fourier transforms. Not until 1991 was the Chauvenet Prize again awarded for a MATHEMATICS MAGAZINE article. This was the extraordinary paper by W. B. R. Lickorish of Cambridge University and Kenneth C. Millett of the University of California, Santa Barbara, that explained the new knot polynomials of Vaughn Jones ("The new polynomial invariants of knots and links," **61** (1988), 2–23.)

The Allendoerfer Awards were set up in 1976 to recognize outstanding expository articles published in MATHEMATICS MAGAZINE. Bart Braden won this award in 1986 for his article "Design of an oscillating sprinkler," **58** (1985), 29–38, in which he analyzes the engineering behind the design of a lawn sprinkler that spreads water uniformly on a level lawn. An accompanying illustration (our FIGURE 2) provides an alternative design.

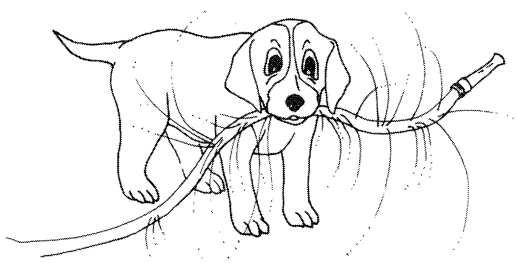


Figure 2 A sprinkler design from 1985

Some other award-winning MAGAZINE gems are included in the section on mathematical themes below.

Mathematical themes

Algebra The MAGAZINE contains relatively few algebra gems, in part because algebra played a minor role in the MAGAZINE's early years.

• In "Remarks on the functional equation $f(x + y) = f(x) + f(y)$," **42** (1969), 121–23, Edwin Hewitt and Herbert S. Zuckerman note that since Cauchy's time it was known that the only continuous, additive homomorphisms of the real numbers were the linear maps $f(x) = kx$. In 1905, Georg Hamel constructed (many!) discontinuous homomorphisms of $\mathbb{R}$ and proved that, surprisingly, the graph of each such function is dense in the plane $\mathbb{R}^2$. Hewitt and Zuckerman give an elementary proof of this result, without using the axiom of choice as Hamel did. They also show that all solutions to the very different looking functional equation $f(x + y) = g(x) + h(y)$ come from, in a simple manner, homomorphisms of $\mathbb{R}$.

• In volume **47** (1974), 226–27, Andy Magid of the University of Oklahoma used commutative ring theory to prove the following theorem: Every trigonometric identity is a consequence of $\sin^2 x + \cos^2 x = 1$. Here "trigonometric identity" means an identity in x that has been first simplified to a polynomial identity of the form $f(\sin x, \cos x) = 0$. In a letter to the editor, **48** (1975), 4, Harry W. Hickey noted that it is not necessary to use commutative ring theory to prove this theorem; more elementary means will do.

Calculus and analysis Let us return to the article by Stone. The classical Weierstrass approximation theorem [**13**] asserts that any continuous function on a closed interval $[a, b]$ can be uniformly approximated there by a polynomial function. This theorem can be reformulated in terms of the algebra $C([a, b])$ of all continuous functions on $[a, b]$, which contains as a subalgebra the family P of all polynomials in a single variable x. $C([a, b])$ is a complete metric space under the so-called supremum norm, where the distance between two continuous functions f and g is

$$d(f, g) = \max_{x \in [a,b]} |f(x) - g(x)|.$$

Weierstrass's approximation theorem then asserts that the uniform closure of P is $C([a, b])$ itself or, equivalently, that P is dense in $C([a, b])$.

In 1937 Stone, then at Harvard, gave a generalization of the Weierstrass theorem near the end of a very long paper [**9**] that focused principally on Boolean rings and Boolean spaces. The stimulus for the generalization came from a conversation Stone had had with John von Neumann who, according to Stone [**11**], asked the "right" question. In the following decade Stone improved the original proof, modified and extended his theorem, and found, along with others, "many interesting applications to classical problems of analysis." [**10**] At the end of the decade, Stone, by then ushering in the famous "Stone Age" as department chair at the University of Chicago, chose MATHEMATICS MAGAZINE for "collecting the relevant material in an expository article where everything could be presented in the light of our most recent knowledge." [**10**] Steven G. Krantz in [**4**] explained that Stone sent this paper to the MAGAZINE "because he had promised them a paper to help them get off to a good start," referring to the late 1940s revival of the MAGAZINE by Editor Glenn James.

Stone's MATHEMATICS MAGAZINE article on "the generalized Weierstrass approximation theorem," which we now know as the Stone-Weierstrass theorem, was published in two parts [**10**]. In the first Stone extended the classical Weierstrass theorem by (a) replacing the interval $[a, b]$ by an arbitrary compact topological space X,

and (b) replacing the subalgebra P of polynomials by any subalgebra U of $C(X)$ (the continuous functions on X) that "separates points in X," meaning that for any two points x and y in X there is a function f in U with $f(x) \neq f(y)$. (Stone used slightly older notation and terminology.) He then concluded that the uniform closure of U is either $C(X)$ itself, or the ideal of all functions in $C(X)$ that vanish at a particular point x_0.

In the second part, Stone extended his generalized Weierstrass approximation theorem to complex-valued functions and to locally compact topological spaces (like $\mathbb{R}^n$, where every point has a compact neighborhood). He gave several applications of his theorems to approximate continuous functions by "trigonometric polynomials" (such as finite Fourier series), by Laguerre functions, and by Hermite functions. The final application of this landmark paper was a proof of the celebrated Peter-Weyl theorem on approximating continuous functions on a compact topological group. All in MATHEMATICS MAGAZINE.

Figure 3 Marshall H. Stone

• Richard Bellman, an applied mathematician who was then at the RAND Corporation, began his article "Inequalities," **28** (1954/55), 21–26, by asserting that mathematics "is fundamentally the study of inequalities rather than equalities." Among other things he sketched a "most ingenious" proof, using two types of induction, of the arithmetic mean-geometric mean inequality:

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \ldots a_n} \qquad (\mathrm{I}_n)$$

for any positive real numbers $a_1, a_2, \ldots, a_n$. The first part of the proof used forward, or ordinary, induction to prove I_n for powers of 2, $n = 2, 4, 8, \ldots$. The second part, which fills in the gaps, featured a *backward* induction showing that I_n implies I_{n-1} by a clever use of the hypothesis I_n. This proof goes back to Cauchy [**1**]. Bellman observed that this "is perhaps the only application" of this form of mathematical induction, but in recent years some computer scientists have used it. For example, Udi Manber uses what he calls reversed induction to prove that certain very dense graphs have a Hamiltonian cycle. [**6**]

• R. C. Buck, an analyst, discusses how topology can illuminate analysis in "Topology and analysis," **40** (1967), 71–74. By analysis here he means elementary calculus, intermediate calculus, and advanced analysis. For elementary calculus he cites the Intermediate Value Theorem, which is just a consequence of the topological theorem

that continuous maps send connected sets to connected sets. For intermediate calculus Buck observes that pointwise convergence and uniform convergence give two competing topologies for the function space $\mathcal{F}$ consisting of all real-valued functions on $(-\infty, \infty)$. The important question—Is the subset consisting of all continuous functions a *closed* set in $\mathcal{F}$?—then has opposite answers for the two topologies.

• The set $\{\sin 1, \sin 2, \sin 3, \ldots\}$ seems to be dense in $[-1, 1]$, yet one might expect the proof to be difficult. On the contrary, C. Stanley Ogilvy in "The sequence $\{\sin n\}$," **42** (1969), 94, gives a two-paragraph proof using only the irrationality of π.

• Problems in the MAGAZINE sometimes generate later articles. Problem 711 in 1968 was to show that for any positive numbers $a_1, a_2, \ldots, a_n$,

$$\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} \frac{1}{a_i}\right) \geq n^2.$$

None of the five solutions presented in "A product of sums," **42** (1969), 161–62, applied the arithmetic mean-geometric mean inequality to each sum:

$$\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} \frac{1}{a_i}\right) \geq \left(n\sqrt[n]{a_1 a_2 \ldots a_n}\right)\left(n\sqrt[n]{\frac{1}{a_1}\frac{1}{a_2}\ldots\frac{1}{a_n}}\right).$$

The most elegant of the five solutions presented applied the Cauchy-Schwarz inequality to

$$\left(\sum_{i=1}^{n} (\sqrt{a_i})^2\right)\left(\sum_{i=1}^{n} \left(\frac{1}{\sqrt{a_i}}\right)^2\right),$$

for a one line proof.

• Calculus instructors who assign exercises of the form $\lim_{x\to 0^+} f(x)^{g(x)}$, such as

$$\lim_{x\to 0^+} (\sin x)^{\tan x} \quad \text{and} \quad \lim_{x\to 0^+} (e^{x+1} - e)^x,$$

may wonder at the ubiquity of 1 as the answer. The article "The indeterminate form 0^0," **50** (1977), 41–42, by Louis M. Rotando and Henry Korn, explains why; the limit is 1 whenever f and g are nonzero *analytic* functions at $x = 0$ (and, of course, $f(x)$ is nonnegative for all positive x sufficiently close to 0), that is, representable by a Taylor series there. The article mentions the counterexample $\lim_{x\to 0^+} x^{a/\ln x}$, which is generalized in an earlier *Monthly* article by G. C. Watson [**12**], while the article itself gives a counterexample involving e^{-1/x^2} to show that the assumption of "analytic at 0" cannot be weakened to "infinitely differentiable at 0."

Combinatorics

• An interesting feature of Louis W. Shapiro's "Finite groups acting on sets with applications," **46** (1973), 136–47, is that it is in the form of a tutorial, with a few key definitions and comments interspersed among numerous exercises. Only an elementary knowledge of group theory is needed for most exercises, and the article begins with the observation that "The concept of a group acting on a set is a small generalization of the idea of a permutation group." The twenty-fifth exercise is the Pólya-Burnside theorem, which Burnside proved in 1897 and Pólya applied in 1937 to get the Pólya enumeration formula. Many of the applications mentioned in the title are to group theory itself, where a group G acts on various sets of subsets of G (such as the orbits under

right multiplication by elements of G). The three Sylow theorems of group theory, for example, are included in the fifty-ninth through sixty-third exercises.

Pólya's enumeration formula itself is the focus of Alan Tucker's "Pólya's enumeration formula by example," **47** (1974), 248–56. Tucker was a teaching assistant for Pólya in a course at Stanford University on combinatorial mathematics in which Pólya omitted a proof of his formula and gave only examples. Tucker proceeds similarly, showing how, with well-chosen examples, one might discover Pólya's formula. In his concluding paragraph he observes, "The importance of our approach is that theory and precise mathematical statements have been avoided ... in favor of the underlying ideas that motivated Pólya."

Games and puzzles Recreational mathematics has always had a modest role in the MAGAZINE. We highlight below several gems that investigate parts of recreational mathematics with tools of modern mathematics such as group theory.

- Flexagons are a part of recreational mathematics yet their inventors were serious research mathematicians or physicists: Arthur Stone, John Tukey, Bryant Tuckerman, and (sometimes serious) Richard Feynman. A trihexaflexagon is a figure made by folding a strip of paper into ten equilateral triangles and then pasting the two end triangles together (yielding, topologically, a Möbius band). Finding the symmetry group of the trihexaflexagon would be a challenging exercise for a modern algebra class. The answer, found in "Symmetries of the trihexaflexagon," by Michael Gilpin, **49** (1976), 189–92, is that the symmetry group is D_9, the group of symmetries of a regular nonagon. A variation of the exercise, for a trihexaflexagon with a design on it that enlarges the symmetry group to D_{18}, is given in "The Faces of the Tri-Hexaflexagon," by Peter Hilton, Jean Pedersen, and Hans Walser, **70** (1997), 243–51. Ethan Berkove and Jeffrey Dumont have continued this thread in "It's Okay to Be Square If You're a Flexagon," **77** (2004), 335–48.

- The entire January 1978 issue of the MAGAZINE was devoted to recreational mathematics and games, such as Jerome L. Paul's "Tic-Tac-Toe in n-dimensions," **51** (1978), 45–49, which is played on a hypercube. The lead article by John Horton Conway on "A gamut of game theories," **51** (1978), 5–12, introduces operations on games, such as their (disjunctive) sum and (conjunctive) join, and value functions on compound games that lead to Conway's surreal numbers. The issue includes Doris Schattschneider's Allendoerfer Award-winning article "Tiling the plane with congruent pentagons," **51** (1978), 29–44. Her survey includes the charming story of how several amateurs—including a homemaker, Marjorie Rice, with barely a high school knowledge of mathematics—independently discovered new pentagonal tilings after reading a column by Martin Gardner in the July 1975 *Scientific American*.

Geometry Geometry has had a prominent role in the MAGAZINE since its inception. But the type of problems and theorems discussed has somewhat broadened from those of the classical geometries of two or three dimensions.

- The title of David Singmaster's article "On round pegs in square holes and square pegs in round holes," **37** (1964), 335–37, almost reveals the question: Which fits better, a round peg in a square hole or a square peg in a round hole? Using the ratio of areas as the yardstick, simple geometry shows that a round peg fits better in a square hole than a square peg in a round hole (as $\pi/4 > 2/\pi$). The generalization to n dimensions provides a surprise: the n-ball fits better in the n-cube than the n-cube fits in the n-ball if and only if $n \leq 8$. The proof of this generalization uses the formula for the volume V_n

of an n-ball with radius r,

$$V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} r^n,$$

where $\Gamma(x)$ is the gamma function. The derivation of this formula, by an argument "accessible to a multivariable calculus class," is given in Jeffrey Nunemacher's "The largest unit ball in any Euclidean space," **59** (1986), 170–71.

• Carl Allendoerfer's purpose in his "Generalizations of theorems about triangles," **38** (1965), 253–59, is to describe appropriate generalizations of theorems about triangles to theorems about tetrahedra that are "generally unknown" (of the many mathematicians he asked, only Pólya knew the generalization described below), even though several of them go back to Descartes and Grassmann. The angle-sum theorem for a triangle—that the sum of the angles is π—is a nice example, since any obvious generalization is doomed: the sum of the solid angles of a tetrahedron is *not* a constant. The key to a successful generalization involves two steps:

1. Reformulating the angle-sum theorem in the equivalent form that the sum of the external angles of a triangle is 2π.
2. Defining the external angle of a triangle at a vertex not as the angle between the two directed sides, but as the angle between the outer normals to these sides. These two steps lead to an elegant proof of the generalization that the sum of the external solid angles of any tetrahedron is 4π. And, just as "triangle" in the previous step may be replaced by "convex polygon," in the preceding generalization "tetrahedron" may be replaced by "convex polyhedron."

• Standard calculus texts seem not to include Steiner's problem, even though it involves minimizing a function of two variables and the answer is appealing. Using geometry, Steiner proved that for any acute-angled triangle ABC the point P which minimizes $PA + PB + PC$ is such that each side of the triangle subtends an angle of 120° at P. "A note on Steiner's problem," by P. N. Bajaj, **40** (1967), 273, gives a short, elegant proof of Steiner's result that uses only simple trigonometry to find the critical point of the function $f(x, y) = PA + PB + PC$ when $P = (x, y)$.

• Before 1976 authors in the MAGAZINE were eligible for the Lester R. Ford Award (now restricted to *Monthly* authors). The last Ford winner in the MAGAZINE was "Geometric extremum problems," by G. D. Chakerian and L. H. Lange, **44** (1971), 57–69. Among other things the authors provide complete solutions to the problems of finding a rectangle of maximum area inscribed in a given triangle or a given ellipse. Special cases of these problems are standard exercises in calculus texts: a side of the rectangle is assumed to be along the base of the triangle or parallel to an axis of the ellipse. Chakerian and Lange use geometry and affine transformations of the plane to show that the solutions for the special cases are actually solutions to the more general optimization problems.

• The interesting question of extending Venn diagrams in a nice way to more than three sets has stimulated much research; volume **15** (1984) of the *College Mathematics Journal* alone has two articles on the topic by the renowned geometer Branko Grünbaum, who also has two such articles in the MAGAZINE, "Venn diagrams and independent families of sets," **48** (1975), 12–23, and the cleverly-named "Diagrams Venn and how," coauthored with J. Chris Fisher and E. L. Koh, **61** (1988), 36–40.

The first of these discusses various extensions of the classic three-circle Venn diagram to more sets, using general curves, convex curves, polygons, and so on. Venn himself in 1880 asserted that with circles the maximal number of sets that could be represented is 3 while with ellipses the maximal number is 4. This last assertion wasn't

proved but was repeated by many authors, even as recently as 1967. In fact, it's false; the maximum number of sets using ellipses is 5 and Grünbaum's first article shows such a Venn diagram:

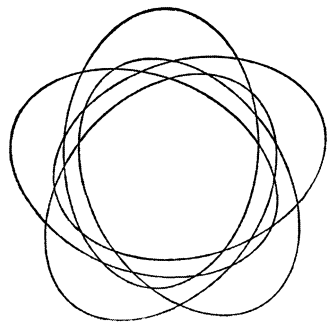


Figure 4 Grünbaum's Venn diagram with 5 sets

• In 1899, George Pick discovered a nice formula for the area A of any simple polygon P whose vertices are lattice points in the plane:

$$A = I + \frac{B}{2} - 1,$$

where I and B are, respectively, the number of lattice points in the interior and on the boundary of P. Here "simple" means that P does not intersect itself. A major theme of "Triangulations and Pick's theorem," by R. W. Gaskell, M. S. Klamkin, and P. Watson, **49** (1976), 35–37, is that "Pick's theorem is not really about area, but a combinatorial result which essentially belongs to topology." The article justifies this assertion by presenting a two-step proof of Pick's theorem involving

1. counting the number of primitive triangles (those containing no lattice points other than its vertices) in any primitive triangulation of P, and
2. showing that the area of each primitive triangle is $1/2$.

The main step, (1), is accomplished by cleverly gluing two rubber copies of P together and inflating this "balloon", then applying Euler's famous topological formula $V - E + F = 2$ to the resulting spherical polyhedron. The upshot is that the number of primitive triangles in the triangulation is found to be $2I + B - 2$, and this with step (2) yields Pick's formula.

Number theory

• The article "Some interesting algebraic identities," by William S. McCulley, **34** (1960/61), 203–6, presents the evoluton of some identities that are important in number theory. It starts with the simple identity

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) = (x_1y_1 + x_2y_2)^2 + (x_1y_2 - x_2y_1)^2, \qquad (*)$$

which expresses a product of two sums of two squares as a sum of two squares. In 1748, Euler generalized $(*)$ to sums of four squares, getting

$$(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) = z_1^2 + z_2^2 + z_3^2 + z_4^2$$

for suitable z_i. The forms in this identity led to Hamilton's quaternions, while Cayley's generalization in 1845 to eight squares

$$\left(\sum_{i=1}^{8} x_i^2\right)\left(\sum_{i=1}^{8} y_i^2\right) = \sum_{i=1}^{8} z_i^2,$$

led to the octonions, or Cayley numbers. This line of generalization stops here, as in 1898 Hurwitz proved that such "reproducing forms" exist only for $n = 2$, 4, and 8. In another direction, Cauchy in 1821 generalized $(*)$ to products of n squares:

$$\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) = \left(\sum_{i=1}^{n} x_i y_i\right)^2 + \sum_{1\le i<j\le n} \left(x_i y_j - x_j y_i\right)^2.$$

Note that omitting the second sum on the right side, which is nonnegative, yields the Cauchy-Schwarz Inequality.

• A well-known result from probabilistic number theory is that the probability that two integers are relatively prime is $6/\pi^2 = 1/\zeta(2)$, where $\zeta(2) = \sum_{n=1}^{\infty} 1/n^2$ is the value of the Riemann Zeta Function at 2. Alan H. Stein generalizes this result in "On almost relatively prime integers," **48** (1975), 169–70, by asking what is the probability that the greatest common divisor of two integers is in X, for any subset X of positive integers? His answer is the proportion of $\zeta(2)$ contributed by integers in X. More precisely, the probability is $(\sum_{n\in X} 1/n^2)/\zeta(2)$. For example, the probability that two integers have greatest common divisor 2 is one-fourth of the probability that they are relatively prime, or about 0.15.

• Of Paul Erdős's 1514 mathematical papers, four appeared in the MAGAZINE, beginning in 1975. In "Some unconventional problems in number theory," **52** (1979), 67–70, Erdős provides "a mélange of simply posed conjectures with frustratingly elusive solutions." An example: "Forty years ago I asked: does $x^x y^y = z^z$ have any nontrivial solutions in integers? Chao Ko [in 1940] found infinitely many solutions; perhaps he found them all." Another example: "It is extremely difficult to obtain results on the difference of consecutive primes. A well-known conjecture of Cramer states that

$$\limsup_{n\to\infty} \frac{p_{n+1} - p_n}{(\log n)^2} = 1.$$

This conjecture is completely unattackable by present day methods and I expect that it will stay in this class for a very long time." Erdős's prediction may well be correct, but Helmut Maier and Carl Pomerance [**5**] have evidence that Cramer's conjecture may not be the right one.

Probability The sample below is representative of many other MAGAZINE articles that answer interesting questions on probability by elementary methods, often with surprising conclusions.

• A mathematics major who is beginning a first course in either probability theory or measure theory would do well to read Truman Botts's tour de force, "Probability theory and the Lebesgue integral," **42** (1969), 105–11. Botts starts from scratch—a single coin flip or die roll for discrete probability, and a single spin on a rotary dial for continuous probability—and gently works his way to probability spaces and Lebesgue measure. Along the way he offers valuable insights on the advantages of the Lebesgue integral over the Riemann integral (it integrates more functions), and the reasons why the Lebesgue integral is a natural tool in probability theory while the Riemann integral is "inadequate."

• With a large class that has seen the ordinary birthday problem one might try the variation discussed in "Another generalization of the birthday problem," by J. E. Nymann, **48** (1975), 46–47. For a class of n students, with $n \geq 23$, suppose that students were asked in succession to call out their birthdays; how many such students would be needed before the probability of a birthday match with someone *in the whole class* would be greater than or equal to $1/2$? The answers are surprisingly small: only 7 for $n = 40$ and 3 for $n = 86$, for example. The derivation of the relevant probability formula is easy and is quite similar to that of the probability formula for the ordinary birthday problem.

A seemingly more difficult derivation, for expected value instead of probability, is given in "A direct attack on a birthday problem," by Samuel Goldberg, **49** (1976), 130–31. The author finds that the expected number of different birthdays among n people is simply

$$365 - \frac{364^n}{365^{n-1}} = 365 - 365\left(\frac{364}{365}\right)^n.$$

Teaching and pedagogy It is probably safe to say that MAGAZINE interest in primarily educational issues has somewhat declined over the years, perhaps since there are now many other venues for such articles. Of course, any contributions after one by George Pólya—the first gem below—would seem to indicate a decline!

• Ten years before the publication of George Pólya's *How To Solve It*, a much truncated version of his advice on problem solving, **9** (1934/35), 172–74, appeared in the MAGAZINE. Pólya was then still a professor in Zürich, but for MAGAZINE readers it was a preview of what was to come.

• In the mid-50s, a regular section of the MAGAZINE was called "Teaching of Mathematics." The article "The group method," by S. Birnbaum and K. Ommidvar, **28** (1954/55), 277–79, advocated a method of instruction combining "dynamic group activity with individual responsibility" that is now known as the method of small groups. A commentator in the next issue wrote that the method was "standard equipment" in his electronics lab courses, where he found the optimum group size to be 3. In the 1970s small-group instruction in college-level mathematics was rediscovered by writers in other MAA journals like Neil Davidson [**2**] and Julian Weissglass [**14**].

The Problems section

Problems have been presented in the MAGAZINE almost since its inception. A separate "Problem Department" was initiated in issue 6 of Volume **5** (1930/31).

Skimming over the Problems sections of the MAGAZINE, one finds the names of many well-known problem solvers or problem posers. But we also find some names we might not expect: David Blackwell, Eugenio Calabi, George B. Dantzig, Jr., Louis de Branges, and Mark Kac.

In 1954, the master problem poser Murray Klamkin asked readers to show that

$$F(x, y) = \sum_{n=0}^{\infty} \frac{(-1)^n x^n}{a^{n+1} + y}$$

is symmetric in x and y. George Pólya, by then a professor emeritus at Stanford, sent in a solution (the only one, except for the proposer's), **28** (1954/55), 235–36. It is not surprising that the author of *How To Solve It* should break down his solution into

three parts: (1) a heuristic consideration in which he outlines a natural approach to the problem, (2) a proof, and (3) a critique in which he discusses values of a, such as $a = 0$, for which the statement is not true.

Twenty-five years later, another Klamkin problem had only one noncomputer solution submitted, and this time the solver was none other than Paul Erdős! The problem asked if there exists a prime number such that if any digit (in base 10) were changed to any other digit then the resulting number would be composite. Erdős answered yes and proved a slightly stronger result. Further, he proposed several questions of his own. The solution, listed as "Erdős and the computer," **52** (1979), 180–82, included 294,001 as the smallest such prime that was found by a computer search.

Epilogue Will we see another paper like Stone's in the MAGAZINE? Probably not. With the transfer of the MAGAZINE to the MAA in 1960, the Board of Governors specified that the "level of the MAGAZINE shall be below that of the *Monthly* but above that of the *Mathematics Teacher*." We hope that the articles highlighted here will encourage readers to delve into the back issues of the MAGAZINE and find for themselves some additional "gems." A collection of papers from the MAGAZINE, with commentary, will be published in 2005 in the MAA's Spectrum Series under the title *A Mathematics Magazine Reader*.

REFERENCES

1. Augustin Cauchy, *Oeuvres Complètes, Sér. 2.*, vol. 3, Gauthier-Villars, Paris, 1897, pp. 375–377.
2. Neil Davidson, The small group-discovery method as applied in calculus instruction, *Amer. Math. Monthly* **78** (1971), 789–791.
3. Donald E. Knuth, *The Art of Computer Programming*, 3rd edition, vol. 1 (Fundamental Algorithms), Addison-Wesley, Reading, MA, 1997, p. 86.
4. Steven G. Krantz, *Mathematical Apocrypha*, Mathematical Association of America, Washington, DC, 2002, p. 168.
5. Helmut Maier and Carl Pomerance, Unusually large gaps between consecutive primes, *Trans. Amer. Math. Soc.* **322** (1990), 201–237.
6. Udi Manber, *Introduction to Algorithms/A Creative Approach*, Addison-Wesley, Reading, MA, 1989, pp. 244–246.
7. Roger B. Nelsen, *Proofs without Words, Exercises in Visual Thinking*, Mathematical Association of America, Washington, DC, 1993.
8. ———, *Proofs without Words II, More Exercises in Visual Thinking*, Mathematical Association of America, Washington, DC, 2000.
9. Marshall H. Stone, Applications of the theory of Boolean rings to general topology, *Trans. Amer. Math. Soc.*, **41** (1937), 375–481.
10. ———, The generalized Weierstrass approximation theorem, this MAGAZINE **21** (1947/48), 167–184, 237–254. (Reprinted in *Studies in Modern Analysis*, R. C. Buck, ed., vol. 1 in the *Studies in Mathematics* series, Mathematical Association of America, Washington, DC, 1962.)
11. ———, A reminiscence on the extension of the Weierstrass approximation theorem, *Historia Math.* **3** (1976), 328.
12. G. C. Watson, A note on indeterminate forms, *Amer. Math. Monthly* **68** (1961), 490–492.
13. Karl Weierstrass, *Mathematische Werke*, Band 3, Abhandlungen III, pp. 1–37, especially p. 5 (= *Sitzungsberichte, Kön. Preussischen Akademie der Wissenschaften*, July 9 and July 30, 1885).
14. Julian Weissglass, Small groups: an alternative to the lecture method, *Two-Year College Math. J.* **7** (1976), 15–20.

NOTES

Transposition Graphs: An Intuitive Approach to the Parity Theorem for Permutations

DEAN CLARK
Department of Mathematics
University of Rhode Island
Kingston, RI 02881
dclark@uri.edu

Understanding the classical parity theorem for permutations—that a permutation can be written as an odd or even product of transpositions, but not both—requires little, if any, mathematical background. Our approach will be familiar to anyone who has played the game of unscrambling jumbled words. For instance, can you transform the scrambled word *urcle* into an everyday English word, step-by-step, where each step consists of a single switch of a pair of letters? To introduce notation, here is a way that it can be done in two steps: $(1, 4)urcle = lrcue$, $(2, 4)lrcue = lucre$ (that is, profit or gain). The notation $Y = (i, j)X$ will be used henceforth to indicate that the letters of X in positions i and j are exchanged to produce word Y. In that case, of course, $X = (i, j)Y$.

There is something mysterious about the way the human brain accomplishes such pattern recognition. Sometimes it happens in a stroke, without any sense of intermediate steps like $(1, 3)urcle = crule$, $(4, 5)crule = cruel$. We'll probe this mystery further in *Application 2*, below. For now, the meaning of the parity theorem becomes easy to state: There are infinitely many ways to transform *urcle* into *cruel* using a sequence of *transpositions* (pair-switches). Efficiency and intelligence are irrelevant. Monkeys making haphazard pair-switches will certainly create the word cruel, over and over, long before monkey typists create a work of Shakespeare. The method above used two transpositions: an *even* number. The parity theorem says that any other method, however many steps it takes, must also use an even number of transpositions. The analogous result holds true in all cases requiring an *odd* number of pair-switches.

Here is another puzzle: As accessible as the meaning of the parity theorem is, the mainstream proofs are tricky, unintuitive, and nonconstructive—in other words, *elegant* in a way that only a mathematician can appreciate. There are two standard approaches. The first requires careful examination of the product $\prod_{1 \le j < k \le n} (x_j - x_k)$ for a given ordering $x_1, x_2, \ldots, x_n$ of the integers $1, 2, \ldots, n$ [**5, 6, 11**]. The second uses the fact that a transposition acting on a product of disjoint cycles changes the number of disjoint cycles in the product by exactly 1 [**2, 4, 7**]. Both approaches presuppose some knowledge of basic group theory and terminology. But what about "the rest of us?" For example, students in the typical liberal arts mathematics course sometimes need a little dynamite to rouse their interest in things mathematical. We provide some in the form of a baffling new take on the card game *Three Card Monte*, as our *Application 1*.

One of our purposes is to provide a proof of the parity theorem that is pictorial, constructive, and immediate in a way that the traditional proofs are not. We assume no knowledge of group theory, graph theory, or combinatorics; however, we won't hesitate to point out connections where appropriate.

Transposition graphs The treatment of the six permutations of a three-letter word contains all the essentials. Let this word be *abc*. FIGURE 1 shows a diagram with features of both a *graph* and an *array*. It contains six *vertices* (points, the $3! = 6$ permutations of *abc*) arranged in rows, which are numbered 0, 1, and 2. The vertices are labeled by the permutations of *abc*. Two vertices are joined by an *edge* (line) if and only if one word is obtained from another by a transposition. Each edge is labeled with the transposition that activates this switch. We call this type of diagram a *transposition graph.*

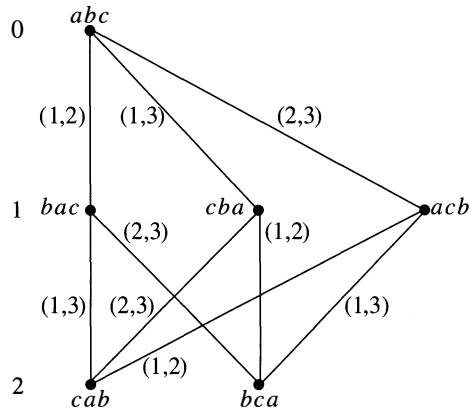


Figure 1 The transposition graph G_3

A *product* of transpositions corresponds to a path in this graph starting from *abc* and ending at another word, or returning to *abc*. An *even* (*odd*) permutation is defined as one that can be written as a product of an even (odd) number of transpositions. Such a path may wander arbitrarily before ending at one of the six vertices. However, as we follow the path step-by-step from row 0 we may say, with each move along an edge to a new row, "odd, even, odd, even, etc." according to the parity of the row number. The parity of the row number matches the parity of the number of transpositions (edges) in any product. Consequently, it is impossible for one path to a given word to use an odd number of edges while a second path to the same word uses an even number of edges. This is the complete argument, *even for the general case* to follow.

Remark 1 Readers familiar with graph theory will note that FIGURE 1 shows a redrawing of $K_{(3,3)}$, the *complete bipartite graph*. The *bipartition* consists, respectively, of the vertices in the even- and odd-numbered rows. The "first theorem of bipartite graph theory" (that every cycle in such a graph has even length [**2**]) is illustrated by the "odd, even, odd, even, ..." procedure described above. Group theorists will see a picture of the *alternating group* A_3 in the even-numbered rows of FIGURE 1.

Since the parity theorem is an immediate corollary of the fact that the general version of FIGURE 1 has a similar structure, it makes sense to define a few graph-theoretic terms. A *graph* is a structure like FIGURE 1 consisting of *vertices*, pairs of which are connected by *edges*. Two vertices are called *adjacent* if they are joined by an edge. A *path* is a sequence of vertices such that every pair of successive vertices is adjacent. The *degree* of a vertex is the number of edges issuing from it. A graph is called *regular* when all the vertices have the same degree. Finally, a graph is *bipartite* when its vertex set can be partitioned into two nonempty subsets such that each edge joins a vertex in one subset to a vertex in the other subset.

First, we give an informal description of how the method used to create FIGURE 1 is extended to create a transposition graph for a word with n distinct letters. Position this word in row 0. Apply every possible transposition once and write the resulting $n(n-1)/2$ words as vertices in row 1 located below row 0. Connect each of the vertices in row 1 to the word in row 0 with an edge labeled by the appropriate transposition. Next, apply the same process to each word in row 1, writing only the new (not-yet-written) words in row 2, below row 1. If transposition τ applied to word X in row 1 results in word Y in row 2, draw an edge from X to Y and label the edge with τ. If τ returns X to row 0, do nothing (the labeled edge has already been drawn). Continue this process, filling in the rows and labeling the edges, until no new words can be created. The graph so obtained will be denoted by G_n. Let R_n denote the array of words of G_n, that is, G_n with all vertices, edges, and edge labels removed (as shown FIGURE 2).

A systematic method Drawing G_4, with its 24 vertices and 72 edges, will be very tedious if you proceed by the informal instructions of the last paragraph. However, the object of immediate interest is R_4 in FIGURE 2, below, and it turns out that the arrays R_n can be constructed recursively in a straightforward way. For a concrete example, consider R_4:

0	***abcd***										
1	***bacd***	***cbad***	***acbd***	*dbca*	*adcb*	*abdc*					
2	***cabd***	***bcad***	*dacb*	*bdca*	*badc*	*dbac*	*cdab*	*cbda*	*dcba*	*adbc*	*acdb*
3	*dabc*	*cdba*	*cadb*	*dcab*	*bdac*	*bcda*					

Figure 2 The array R_4

First, the fourth letter d is appended to each of the six original words in R_3. Three new positions are created in row 1 to allow for the three transpositions $(1, 4)$, $(2, 4)$, $(3, 4)$ operating on the word in row 0. This is where *dbca*, *adcb*, and *abdc* come from. Next, nine new positions are created in row 2 to allow for the same three transpositions acting, respectively, on *bacd*, *cbad*, and *acbd*. Finally, a new row 3 is added to allow for the three transpositions acting on *cabd* and *bcad*. The result is the array R_4 with four rows containing all $4! = 24$ permutations of *abcd*.

The following general construction lists all the permutations of a word with n distinct letters classified by the rows according to the minimum number of transpositions in the product. Suppose that R_n has been constructed. Let $L(n, k)$ denote the number of words in row k of R_n, $k = 0, 1, \ldots, n-1$. (We'll also define $L(n, -1) \equiv L(n, n) \equiv L(n, n+1) \equiv \cdots \equiv 0$.) To construct R_{n+1}:

(1) Introduce the $(n+1)$st letter and append it to every word in R_n.

(2) Augment each row k of R_n with $nL(n, k-1)$ additional positions.

(3) Introduce row n containing $nL(n, n-1)$ positions.

(4) Insert into the new positions of row k the words obtained by applying the n transpositions $(1, n+1)$, $(2, n+1)$, $\ldots$, $(n, n+1)$ to each of the $L(n, k-1)$ words in row $k-1$, for $k = 1, \ldots, n$.

Remark 2 The recurrence relation used in the construction is based on the one associated with the *unsigned Stirling numbers of the first kind* $\sigma(n, k) = (-1)^{n+k} s(n, k)$. The *signed* Stirling numbers $s(n, k)$ are the coefficients of x^k in the product

$$x(x-1)\cdots(x-n+1).$$

Much less well known than the binomial coefficients, the $\sigma(n, k)$ have a similar recurrence relation $\sigma(n+1, k) = \sigma(n, k-1) + n\sigma(n, k)$. For instance, if you were wondering how many words are in row k of R_n, the answer is $\sigma(n, n-k)$, which is not surprising in view of the fact that $L(n+1, k) = L(n, k) + nL(n, k-1)$. Comtet [**3**] gives a discussion of Stirling numbers of the first and second kinds, including numerical tables. For a combinatorial approach to Stirling numbers, see the article by Benjamin et al. in the April 2002 MAGAZINE [**1**].

The graph G_n is recovered from R_n by drawing at each vertex $n(n-1)/2 = \binom{n}{2}$ edges labeled by the appropriate transpositions. The following theorem says that G_n is very much like G_3.

THEOREM. *The transposition graphs G_n generated by the preceding recursive construction have $n!$ distinct words as vertices and $\binom{n}{2}^2(n-2)!$ edges. Each edge connects a vertex in row k to a vertex in row $k+1$ for $k = 0, 1, \ldots, n-2$. Therefore, G_n is a bipartite graph with bipartition consisting, respectively, of the odd- and even-numbered rows. G_n is a regular graph of degree $\binom{n}{2}$.*

Proof. The first interesting case is G_3, for which all the claims are easy to verify. To proceed by induction, we assume that G_n has all the characteristics stated in the theorem and exemplified by G_3 in FIGURE 1. Then we prove that G_{n+1} inherits this same structure from G_n. Let z be the $(n+1)$st letter added to G_n in order to create G_{n+1} and let P, Q, X, and Y denote words in G_{n+1}.

We'll deal with some routine issues first, like the number of vertices in G_{n+1}. Let H_{n+1} denote the newly created part of G_{n+1}, including edges connected to G_n. The G_n piece contributes $n!$ vertices by itself and the construction supplies $n \cdot n!$ new vertices to H_{n+1}. In all, $n! + n \cdot n! = (n+1)!$

All the words in G_n are unique by assumption and remain so when z is appended to each one. But, are all the new words in H_{n+1} unique? To show that they are, suppose X and Y are in H_{n+1} and $X = Y$. Then $X = (i, n+1)P$ for some P in G_n, and thus X has z in position i. If $X = Y = (j, n+1)Q$, then we must have $i = j$ and thus $P = Q$. It follows from the induction hypothesis that X and Y cannot be words located at distinct vertices; hence, all the words of G_{n+1} are unique. This uniqueness, now established in general, allows us to say that the vertex degrees of any transposition graph G_n are all $\binom{n}{2}$, since this is the number of ways of choosing a pair of letters to transpose. By the handshaking lemma (*a.k.a.* "the first theorem of graph theory") the sum of the degrees is twice the number of edges [**2**]. Solving for e in $n!\binom{n}{2} = 2e$ yields the number of edges $e = \frac{n!}{2}\binom{n}{2} = \binom{n}{2}^2(n-2)!$.

That takes care of the routine issues. It remains to show that each edge in G_{n+1} connects row k to row $k+1$ for some $k = 0, 1, \ldots, n-1$. Edges in G_n are assumed to connect adjacent rows, and appending z does not change this. The uniqueness argument offered above proves that every X in H_{n+1} is adjacent to a *unique* P in G_n. Therefore, every edge from G_n to H_{n+1} connects a vertex in row $k+1$ to a unique vertex in row k for some $k = 0, 1, \ldots, n-1$. Finally, consider edges from H_{n+1} to H_{n+1}. We show that X and Y adjacent in H_{n+1} are matched by P and Q adjacent in G_n. In fact, the adjacency of P and Q is necessary and sufficient for the adjacency of X and Y. Consider three exhaustive cases, in which X and Y belong to H_{n+1} and P and Q belong to G_n:

(i) If $X = (i, n+1)P$, $Y = (j, n+1)Q$, then $Y = (i, j)X$ if and only if $Q = (i, j)P$.

(ii) If $X = (i, n+1)P$, $Y = (i, n+1)Q$, then $Y = (k, n+1)X$ if and only if $Q = (i, k)P$, $i \neq k$.

(iii) If $X = (i, n+1)P$, $Y = (i, n+1)Q$, then $Y = (k, m)X$ if and only if $Q = (k, m)P$, $i \neq k$.

This matching of edges is illustrated in FIGURE 3, below. In general, observations (i)–(iii) assure that edges in H_{n+1} connect adjacent rows. Were it not so, (i)–(iii) would imply that there are corresponding adjacent vertices in G_n that are not a row apart, contradicting the induction hypothesis.

This completes the proof of the theorem. ■

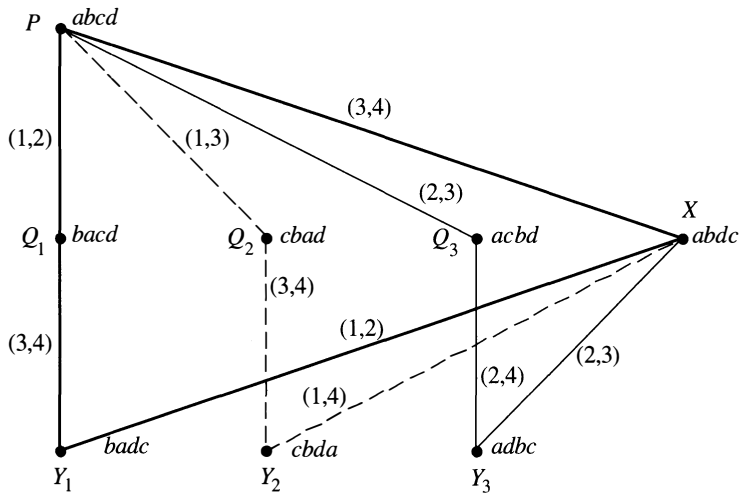


Figure 3 *plain*: case (i); *dashed*: case (ii); *bold*: case (iii)

Remark 3 Early in the proof we determined the number of edges of G_n. A more interesting question concerns the number of edges joining two adjacent rows. Concretely, in how many ways can all the words obtained by a minimum of m transpositions be transformed into all the words obtained by a minimum of $m+1$ transpositions by applying one more transposition? Using the regularity of G_n, it is not hard to show that the number of edges connecting row $n-k$ to row $n-k-1$ is

$$(-1)^{n+k}\binom{n}{2}\sum_{j=1}^{k} s(n, j).$$

The $s(n, k)$ are the signed Stirling numbers of the first kind cited in Remark 2, above.

Transposition graphs are nested structures for visualizing permutations and the way they stand in relation to each other by way of transpositions. For example, the standard result concerning the number of elements in the alternating group A_n of even permutations is available immediately by induction: The even rows of G_n contribute $n!/2$ vertices to the even rows of G_{n+1}, while the odd rows of G_n contribute $n \cdot n!/2$ vertices to the even rows of G_{n+1}. Thus, there are

$$\frac{n!}{2} + n \cdot \frac{n!}{2} = \frac{(n+1)!}{2}$$

even permutations in G_{n+1}.

And, finally, there is the classical parity theorem for permutations, which requires no proof since it has already been given—replace 6 by n—in the paragraph preceding Remark 1, above. We state it as a corollary of the theorem.

COROLLARY. *Every permutation on n distinct symbols a, b, c, ... can be written as a product of either an even number of transpositions or an odd number of transpositions, but not both.*

Application 1—Three Card Monte We'll assume that the following demonstration is taking place in an undergraduate abstract algebra or liberal arts math course. Explain to the class that to win the shady card game *Three Card Monte* a player must simply pick out the single black card, say, the ace of spades, from the other two red cards after the three cards have been rearranged repeatedly, face down. In full view, the three cards are placed face down in a row. One volunteer will be the *player* who thinks she can successfully track the black card. The second volunteer, a *monitor*, will verify that the switching is done according to two rules: First, the cards must always remain in a row. Second, there are only three legal switches: first and second cards, first and third cards, and second and third cards. The same move can be repeated, and the moves need not be done in any particular order. *Without making it conspicuous, you place the black card farthest to the left, with a red two in the middle, and a red three last. You refer to them simply as the black and red cards, but you must remember the original positions of all three.*

Ask a student to choose a number between 10 and 20 to serve as the total number of switches made during the game. The monitor will verify that the moves are legal and that the total number of switches is correct. You begin moving the cards nice and easy, reminding the player to follow the black card... then, suddenly, well before the required number of switches, you stop. Turn your back to the cards and ask the *player* to make two switches. It gets more diabolical. Turn around and ask the player to turn *her* back in order to let *you* make two switches. Then, both of you face the cards and you complete the remaining switches according to the chosen number. Finish by telling the player that you *know* where the black card is and that she must pick it up. But, this is preposterous since each of you failed to see two switches! Nevertheless, insist that you are communicating its position to her telepathically. Her hand may be *guided to the ace by telepathic force!* She turns over a card. If it is the ace of spades, she wins and you are indeed a telepath, which will probably happen about a third of the time. If the chosen card is not the ace of spades, you pick the ace up and show it to the class as promised.

How is it done? Consider FIGURE 1 where the roles of a, b, and c are played, respectively, by the *ace of spades*, *two of hearts*, and *three of diamonds*. There is a simple mnemonic for discriminating the even-rowed permutations from the odd-rowed permutations. The permutations in row 1 have *exactly one* card in its original position. The permutations in the even rows have either all the cards in their original position or *none* of the cards in their original position. Hearing the announced number of switches (even or odd) and seeing one card enables you to deduce the identities of both face-down cards.

Application 2—JUMBLE© We return to an example akin to the *urcle* of the introduction, but this time with a nine-letter word (which would appear in row 5 of G_9):

dieslcamp.

According to Remark 3, there are 827,856 edges above row 5 in G_9. In years of classroom experience no one has ever unscrambled *dieslcamp*, a quite ordinary English word, at this initial stage. (I tell students that *dieslcamp* is a bucolic place where tired train engines are sent in the summer for recuperation.) On the other hand, a computer could easily solve the problem. The *Steinhaus-Johnson-Trotter algorithm* [**8, 9, 10**] lists all the permutations using transpositions of adjacent letters. The list corresponds

to a path through G_n that visits every vertex exactly once (a *Hamiltonian path*). Armed with a sufficiently large dictionary, a computer could check all $9! = 362,880$ permutations of *dieslcamp* and find the matching word, possibly before you read to the end of this sentence.

As with chess-playing computers, what is interesting is that human brains do not seem to solve the problem this way. Certainly, transposition graphs will add no insight into how human brains solve the problem. However, we *will* let you experience for yourself the highly discontinuous "aha!" moment—the quantum leap—of recognition by playing the following game. Since I myself know the mystery word, I'll switch pairs of letters one at a time so that the correct letter is put in the correct place moving from left to right. Here's the first switch:

mieslcadp.

The first letter of the unscrambled word is *m*. We have just moved *up* one row in G_9 (if not, four more switches will create the mystery word but it won't be in row 0!). If I continue to transpose letters to their correct positions, then at some point the mystery word will snap into focus. Incidentally, the second letter *i* is already in the correct position. Do you see it now with a sudden flash of insight? If not, we make the third letter correct:

miselcadp.

In a typical classroom experiment this is the point at which two or three hands shoot up, signaling recognition of the mystery word. Why not the whole class? Is the reader not yet at cognitive breakthrough? With one more switch, see how the cognitive blind spot vanishes:

misplcade.

You might have guessed it would be this word, as so many letters have been treated this way from the beginning of our discussion.

REFERENCES

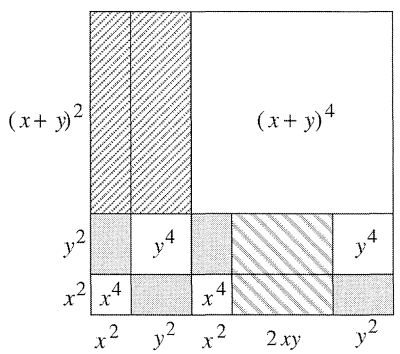
1. Arthur T. Benjamin, Gregory O. Preston, and Jennifer J. Quinn, A Stirling enCOUNTer with harmonic numbers, this MAGAZINE, **72**:2, (2002), 95–103.
2. Norman L. Biggs, *Discrete Mathematics*, revised edition, Oxford University Press, Oxford, 1989.
3. Louis Comtet, *Advanced Combinatorics*, Reidel, Dordrecht, 1974.
4. John B. Fraleigh, *A First Course in Abstract Algebra*, 2nd ed., Addison-Wesley, Reading, 1976.
5. Pierre Antoine Grillet, *Algebra*, Wiley, New York, 1999.
6. Thomas W. Hungerford, *Algebra*, Springer-Verlag, New York, 1990.
7. I. Martin Isaacs, *Algebra—A Graduate Course*, Brooks/Cole, Pacific Grove, CA, 1994.
8. S. M. Johnson, Generation of permutations by adjacent transpositions, *Math. Comp.* **17** (1963), 282–285.
9. E. Reingold, J. Nievergelt, N. Deo, *Combinatorial Algorithms*, Prentice Hall Professional, New Jersey, 1977.
10. Hugo Steinhaus, *One Hundred Problems in Elementary Mathematics*, Dover, New York, 1990.
11. John Stillwell, *Elements of Algebra*, Springer-Verlag, New York, 1994.

Proof Without Words: Candido's Identity

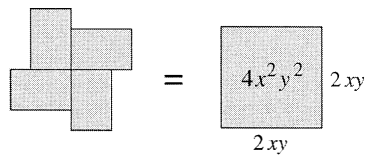
Giacomo Candido, 1871–1941

$$[x^2 + y^2 + (x + y)^2]^2 = 2[x^4 + y^4 + (x + y)^4]$$

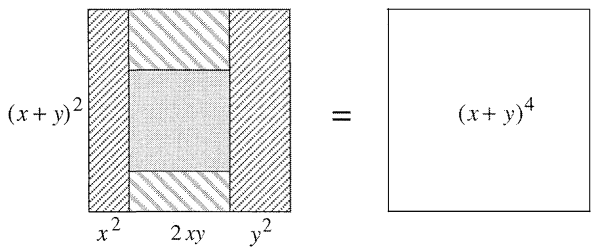
1.



2.



3.



Note: Candido employed this identity to establish

$$[F_n^2 + F_{n+1}^2 + F_{n+2}^2]^2 = 2[F_n^4 + F_{n+1}^4 + F_{n+2}^4],$$

where F_n denotes the nth Fibonacci number.

——ROGER B. NELSEN
LEWIS & CLARK COLLEGE
PORTLAND OR 97219

Maximizing the Chances of a Color Match

RAMIN NAIMI
Occidental College
Los Angeles, CA 90041
rnaimi@oxy.edu

ROBERTO CARLOS PELAYO
California Institute of Technology
Pasadena, CA 90035
roberto@caltech.edu

A classroom experiment In a classroom with seats arranged in a rectangular grid, each student is given an individual list of n colors, from which each student is to choose one color at random. The lists may vary from student to student. Any pair of lists may overlap and could even have all n colors in common. According to your intuition, are the following statements true or false?

1. The probability that some pair of students will pick matching colors is greatest if all lists are identical.
2. The probability that some pair of *adjacent* students will pick matching colors is greatest if all lists are identical. (Two students are said to be adjacent if they are next to one another in a row or one is immediately in front of the other.)

For most of us, our intuition says that both are true. But, surprisingly, only the first one is! Try to find a counterexample to the second statement before reading any further. Hint: there is a simple one with only six students and $n = 2$. As a further challenge, prove that the first statement is true.

A counterexample The situation described in our experiment can be represented by a graph, in which the vertices (dots) represent students and edges (lines) indicate adjacent students. For example, FIGURE 1 shows a class of six students (a, b, c, d, e, f) seated in two rows of three each. We first give all six students the same list of two colors (which we'll represent as numbers), say $\{1, 2\}$, and compute the probability that two adjacent students will pick the same color. After this, we will see that it is possible to improve upon this probability by giving the students different lists.

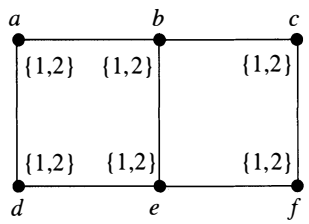


Figure 1 Six students with identical lists

Before doing so, however, it is convenient to introduce a bit more terminology. A *coloring* is an assignment of one color to each vertex from its list. In our example, this corresponds to each student picking a color from his or her list. If two adjacent vertices are assigned the same color, we say there is a *match*. A *proper coloring* is a coloring

with no matches. So our goal is to find a set of lists that maximizes the probability of obtaining a match, or, equivalently, has as few proper colorings as possible.

Now let's go back to FIGURE 1 and compute the probability of getting a match. There are 2^6 possible ways to color the graph. Out of these, there are exactly two proper colorings: assign color 1 to a, c, e and color 2 to b, d, f, or vice versa. So the probability of obtaining a match is $(2^6 - 2)/2^6$. What matters here is that it's less than 1.

In FIGURE 2, on the other hand, the probability of getting a match is 1! To see why, try to find a proper coloring, and you'll find that it's impossible. Start with vertex b, and assign color 1 or 2 to it:

Case 1. Pick 1 for b. Then, to avoid a match, we must pick 3 for a, and therefore 2 for d, and 1 for e, which gives a match between b and e.

Case 2. Pick 2 for b. Then, to avoid a match, we must pick 3 for c, and therefore 1 for f, and 2 for e, which again gives a match between b and e.

So, if there are six students seated in two rows of three each, using this set of lists *guarantees* that two adjacent students will pick the same color! We would have no such guarantee if we gave every student the same list.

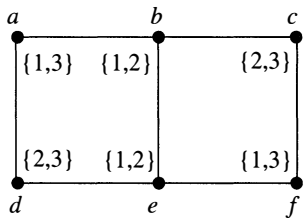


Figure 2 A counterexample

Other graphs A natural question to ask now is: *What about other class sizes and seating arrangements?* If we are trying to maximize the probability that two adjacent students pick the same color, when is it wise to give everyone the same list of colors? Or, more generally, *for which graphs does having identical lists maximize the probability of getting a match?* Let's give such graphs a name: a graph is *n-monophilic* if no set of n-color lists has a greater probability of yielding a match than the set with all the lists equal.

With this definition, we can restate our question as:

QUESTION 1. *For $n \geq 2$, which graphs are n-monophilic?*

The following theorems offer some partial answers. Only Theorem 1 is proved in the printed version of this paper; proofs of the remaining theorems can be found online at the MAGAZINE website, `www.maa.org/pubs/mathmag.html`.

Theorems 2 and 3 have short and easy proofs, and we encourage you to prove them yourself; hints are provided after their statements. The proofs of Theorem 4 (perhaps surprisingly) and Theorem 5 are more involved.

Some of the results in this paper are similar to those of Erdős, Rubin, and Taylor [2]. We thank Thomas Hull at Merrimack College and Douglas West at the University of Illinois at Urbana-Champaign for bringing this similarity to our attention.

THEOREM 1. *For every $n \geq 2$, there exists a graph that is not n-monophilic.*

If you are not familiar with common graph theory terminology, the following definitions will be helpful for the remaining theorems. A graph is said to be *complete* if every pair of vertices is connected by an edge. A *path* is a sequence of distinct vertices $v_1, v_2, \ldots, v_k$ such that every pair of consecutive vertices $\{v_i, v_{i+1}\}$ is connected by an edge. A graph is *connected* if for every pair of vertices $\{v, w\}$ there is a path $v_1, v_2, \ldots, v_k$ with $v_1 = v$ and $v_k = w$. A *tree* is a graph in which every pair of vertices $\{v, w\}$ is connected by exactly one path. A *cycle* is a sequence of distinct vertices $v_1, v_2, \ldots, v_k$ such that every pair of consecutive vertices $\{v_i, v_{i+1}\}$, as well as $\{v_k, v_1\}$, is connected by an edge. A cycle is *even* if it has an even number of vertices. For example, the graph in FIGURE 2 contains three even cycles, two of length 4 and one of length 6.

THEOREM 2. *Every complete graph is n-monophilic for all* $n \geq 2$.

Hint for proof: Do a direct counting argument to show that the graph K_m has at least $n(n-1)\cdots(n-m+1)$ proper colorings for any set of n-color lists, and using identical lists for all vertices achieves this lower bound.

As a fun exercise, find the fallacy in the following wrong proof:

> For any given pair of vertices, the probability of getting a match between those two vertices is greater when they have identical lists than when they don't. Therefore, the probability of getting a match for the graph is greater when all the lists are identical.

To convince yourself that the above argument is incorrect, just note that the same proof would apply to any graph; but we have already seen that not every graph is n-monophilic. Can you pinpoint exactly which step in the above argument is incorrect?

If, in the classroom experiment, we are looking to maximize the probability that two students—any two, not just adjacent ones—will pick the same color, then, according to Theorem 2, we should do exactly as our intuition says: give everyone identical lists.

THEOREM 3. *Every tree is n-monophilic for all* $n \geq 2$.

Hint for proof: Show that if G' is obtained by adding one new vertex and connecting it to a vertex of a graph G, then G' is n-monophilic if and only if G is.

THEOREM 4. *Every cycle is n-monophilic for all* $n \geq 2$.

Thus, if students are seated around a round table, giving everyone the same list maximizes the probability of two adjacent students getting a match.

For the next theorem, we need to refer to a particular graph called $K_{2,3}$, which is depicted in FIGURE 3. This nomenclature is explained in the last section, where we prove Theorem 1.

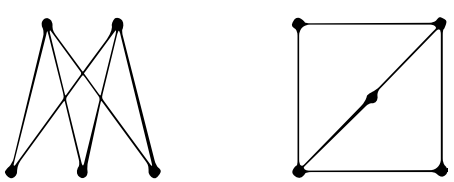


Figure 3 Two diagrams of the graph $K_{2,3}$

THEOREM 5. *A connected graph is* not 2*-monophilic if and only if all its cycles are even and it contains at least two cycles whose union is not* $K_{2,3}$.

With regard to seating arrangements, Theorem 5 implies that in a typical class, where the students are seated in rows and columns, giving everyone the same pair of colors will often *not* maximize the probability of getting a match! If the students aren't trying to avoid sitting next to each other, then there likely exist at least two cycles. And the students' being seated in rows and columns guarantees that all cycles are even and no two cycles have $K_{2,3}$ as their union (can you see why?).

The Dinitz Problem The Dinitz Problem, which was open for fifteen years before finally being solved by Galvin [**3**] in 1994, is as follows.

Suppose there are n^2 students sitting in a rectangular grid of n rows and n columns, each with a list of n colors. Is it always possible for each student to pick a color from his or her list in such a way that no two students in the same row or the same column end up with the same color?

As an example, consider the special case when all the color lists are identical, say $\{1, 2, \ldots, n\}$. Then there is a simple solution, as shown in FIGURE 4. Our intuition might suggest that if the students can pick distinct colors in each row and each column when all the lists are identical, then they should also be able to do so when the lists aren't all identical. Galvin showed that this is indeed the case.

$$\begin{array}{ccccc} 1 & 2 & \cdots & n-1 & n \\ 2 & 3 & \cdots & n & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n & 1 & \cdots & n-2 & n-1 \end{array}$$

Figure 4 A special case of the Dinitz Problem

Let's represent the students with a graph G_n, with n^2 vertices in a rectangular grid and edges connecting any two vertices in the same row or the same column. Readers may recognize G_n as the Cartesian product $K_n \times K_n$. FIGURE 5 shows G_3 as an example. Then the Dinitz Problem asks whether or not we can find a proper coloring of G_n for any set of n-color lists assigned to the vertices.

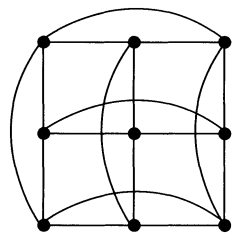


Figure 5 The graph $G_3 = K_3 \times K_3$

The idea of coloring all the vertices of a graph all from a single list has been around for a quite a while (it is, for example, related to the Four Color Problem). In 1979, Erdős, Rubin, and Taylor [**2**] generalized the Dinitz Problem to the question of coloring vertices of arbitrary graphs from *different* lists of colors. They called a graph *n-choosable* if, given any set of n-color lists for its vertices, one can always find a proper coloring. "List coloring," as it's sometimes called nowadays, has become a very active field since then.

It was also the Dinitz Problem that led us to the notion of n-monophilic graphs. Recall that in the special case when all the color lists are identical, there *is* a proper coloring of G_n—a "matchless coloring," so to speak. So if we randomly assign a color from the same list to each vertex, the probability of getting a match is less than 1. Thus, if one could prove that G_n is n-monophilic, it would follow that for every set of n-color lists the probability of getting a match is less than 1. This would mean there always exists a proper coloring, that is, G_n is n-choosable, which would answer the Dinitz Problem. By the same reasoning one sees that for any n-colorable graph (that is, a graph with a proper coloring from one n-color list), n-monophilic implies n-choosable. The converse, however, is not true in general. Galvin [3] proved that G_n is n-choosable. But this doesn't tell us whether G_n is n-monophilic. This is a special case of Question 1:

QUESTION 2. *Is the graph G_n n-monophilic?*

Examples of non-n-monophilic graphs We now prove Theorem 1, which asserts the existence of non-n-monophilic graphs for every $n \geq 2$. To construct examples of such graphs, we need the following definition.

The *complete bipartite graph* on m, n vertices, denoted $K_{m,n}$, is a graph with vertices $v_1, \ldots, v_m$ and $w_1, \ldots, w_n$, where every v_i is connected to every w_j by an edge, and there are no other edges. FIGURE 6 shows a picture of $K_{2,4}$ (ignore the color lists for now) and FIGURE 3 shows $K_{2,3}$.

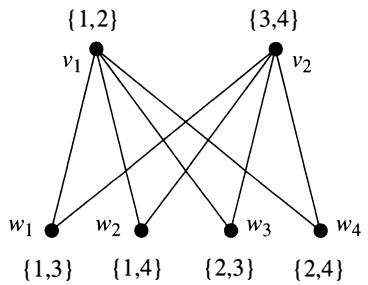


Figure 6 $K_{2,4}$ is not 2-monophilic

We will show that K_{n,n^n} is not n-monophilic for $n \geq 2$.

Assign mutually disjoint lists of n colors to the top row vertices $v_1, \ldots, v_n$. There are n^n different ways to make a list of n colors by choosing one color from each of the n top row lists. Assign these n^n lists to the bottom row vertices $w_1, \ldots, w_{n^n}$.

For any assignment of colors to the n top vertices, there is a bottom vertex whose list consists of those n colors. So there will be a match between that bottom vertex and some top vertex. Thus the probability of getting a match with these color lists is 1. On the other hand, if all the lists are identical, say $\{1, \ldots, n\}$, then the probability of getting a match is less than 1: assigning color 1 to the top vertices and color 2 to the bottom vertices gives a proper coloring. This shows that K_{n,n^n} is not n-monophilic.

Remark You may be wondering: n^n is a very large number; aren't there any smaller examples? Yes, there are; for $n \geq 2$ there is a non-n-monophilic graph with only $3n$ vertices. We don't use this graph in our proof above because constructing it, as well as proving that it is not n-monophilic, is more elaborate and involved (for details, email `rnaimi@oxy.edu`). It would be interesting to try to find the minimum number of vertices of a graph that is not n-monophilic.

Acknowledgment. We thank Tamas Lengyel (of Occidental College) and the referees for many helpful suggestions. Naimi expresses gratitude to Caltech for its hospitality while he worked there on this paper during a sabbatical leave. Pelayo wishes to thank the California Alliance for Minority Participation and the Undergraduate Research Center at Occidental College for supporting this work. Theorem 4 was obtained by Radoslav Kirov and Naimi as part of a summer undergraduate research program at Occidental College.

REFERENCES

1. Reinhard Diestel, *Graph Theory,* Springer, Graduate Texts in Mathematics **173** (1997).
2. Paul Erdős, Arthur L. Rubin, Herbert Taylor, Choosability in graphs, *Congr. Numer.* **26** (1980), 125–157.
3. Fred Galvin, The list chromatic index of a bipartite multigraph, *J. of Combinatorial Theory, Series B* **63** (1995), 153–158.
4. Doron Zeilberger, The method of undetermined generalization and specialization, illustrated with Fred Galvin's amazing proof of the Dinitz Conjecture, *Amer. Math. Monthly*, **103** (1996), 233–239.

Why Euclidean Area Measure Fails in the Noneuclidean Plane

DIETER RUOFF
University of Regina
Regina, SK, Canada S4S 0A2
ruoff@math.uregina.ca

One of the central and most interesting themes of noneuclidean (hyperbolic) geometry concerns the angle sum and area of polygons. A triangle—so we learn—has an angle sum of less than π, a quadrilateral one of less than 2π, and generally a n-gon one of less than $n \cdot \pi - 2\pi$. More specifically, one can establish that the number n of vertices of a polygon does not determine its angle sum, which can be anything between 0 and $n \cdot \pi - 2\pi$. This prepares the way for the remarkable conclusion that the *defect* of a n-gon, the difference between its angle sum and $n \cdot \pi - 2\pi$, has all the desired properties of an area measure.

But where does this leave conventional Euclidean area measure with the formula (base $\times$ altitude)/2 for triangles? Is the defect simply a convenient alternative for measuring area in noneuclidean geometry, or do we have to use it because the Euclidean area measure is not applicable? The second is true, and the reason is that an indispensable but often neglected property of the formula for the area measure of a triangle in a Euclidean plane cannot be carried over to the noneuclidean plane. What we refer to is the *well-definedness* of Euclidean area measure, in particular the fact that in a triangle with sides a, b, c and related altitudes h_a, h_b, h_c, the area can be calculated as

$$\frac{ah_a}{2}, \quad \frac{bh_b}{2}, \quad \frac{ch_c}{2}$$

with the results in all three cases invariably being equal.

For the investigation of the same formula on the noneuclidean side, we choose a path that contrasts the two theories clearly, and which can be taken before or after one discusses the defect.

We first turn to a figure consisting of a triangle $\triangle OAB$ with a right angle at B, and points A' on ray $\overrightarrow{OA}$ such that $\overrightarrow{OA'} = 2 \cdot \overrightarrow{OA}$, and B' on ray $\overrightarrow{OB}$ such that $\triangle OA'B'$ is a

triangle with right angle at B', as in FIGURE 1. In addition we define A'' as the point on ray $\overrightarrow{B'A'}$ that satisfies $BA \equiv B'A''$, and B'' as the image of A'' under the translation $[A' \to A] = [A \to O]$. It follows that $\triangle AA'A'' \equiv \triangle OAB''$. (Note: Readers may be unfamiliar with translations along a line l in noneuclidean geometry; they can be easily explained in terms of reflections. To construct the translation $[A \to O]$, create line r perpendicular to $\overrightarrow{OA}$ through A and line s perpendicular to $\overrightarrow{OA}$ through the midpoint of OA; reflecting through r and then through s produces the desired translation. [**3**, p. 326])

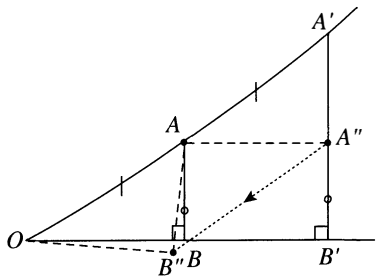


Figure 1

Remembering that triangles have angle sums $< \pi$ and quadrilaterals have angle sums $< 2\pi$, we recognize that in triangle $\triangle AOB$, $\angle AOB + \angle OAB < \pi/2$, and in quadrilateral $BB'A''A$, $\angle BAA'' = \angle B'A''A < \pi/2$. Consequently,

$$\angle AOB < \frac{\pi}{2} - \angle OAB < \pi - \angle OAB - \angle BAA'' \equiv \angle A'AA'' \equiv \angle AOB'',$$

which means that B'' lies outside $\angle AOB$ and $A''A' \equiv B''A > BA$. As a result

$$B'A' \equiv B'A'' + A''A' \equiv BA + A''A' > 2 \cdot BA.$$

Expressed intuitively: *A point that moves with a constant velocity along ray $\overrightarrow{OA}$ distances itself at an accelerating rate from ray $\overrightarrow{OB}$.*

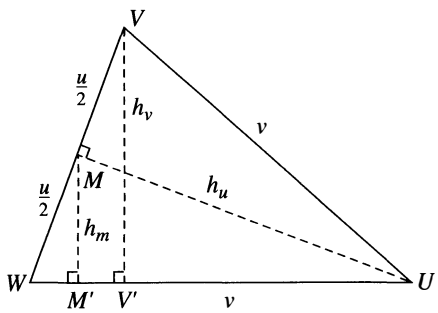


Figure 2

Consider now an acute-angled, isosceles triangle $\triangle UVW$ with apex U, having M as the midpoint of the base, and orthogonal projections M', V' of M, V on WU, as in FIGURE 2. The length of UM, the principal altitude of $\triangle UVW$, will be denoted by h_u and the lengths of UW, VW, VV', and MM' by v, u, h_v, and h_m respectively. Now, if

Euclidean area measure were well-defined in the noneuclidean plane we would have

$$\text{in} \quad \triangle UVW, \qquad \frac{uh_u}{2} = \frac{vh_v}{2},$$

$$\text{and in} \quad \triangle UMW, \qquad \frac{(u/2)h_u}{2} = \frac{vh_m}{2},$$

and so $h_v = 2h_m$. However, we should have $h_v > 2h_m$, as we showed before in FIGURE 1, which means that at least one of the above two equations is false. Hence, Euclidean area measure is not applicable in noneuclidean geometry.

How does one prove that (base $\times$ altitude)/2 is well-defined in Euclidean geometry, and what accounts for the difference in noneuclidean geometry? As is often the case in Euclidean geometry, one verifies the equation between two products of segments by transforming it into one between two quotients and then applies a proportionality theorem [**3**, § 20]. And that is exactly what does not work in noneuclidean geometry; in FIGURE 1

$$\frac{B'A'}{OA'} > \frac{BA}{OA}.$$

REFERENCES

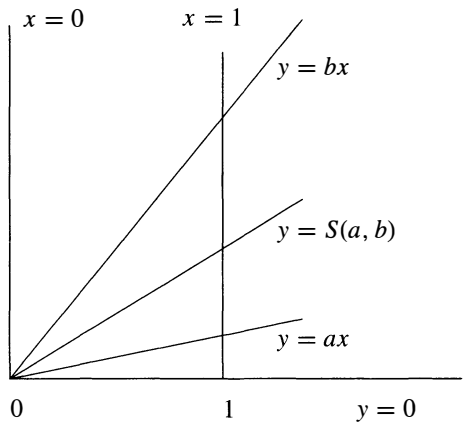
1. H. S. M. Coxeter, *Non-Euclidean Geometry*, 6th ed., The Mathematical Association of America, Washington, D.C., 1998.
2. D. Gans, *Introduction to Non-Euclidean Geometry*, Academic Press, Inc., Orlando, 1970.
3. M. J. Greenberg, *Euclidean and Non-Euclidean Geometry*, W. H. Freeman and Co., San Francisco, 1974.
4. D. Hilbert, *Foundations of Geometry*, The Open Court Publishing Co., La Salle, IL, 1971.
5. J. R. Smart, *Modern Geometries*, 4th ed., Chapter 9, Brooks/Cole Publishing Company, Pacific Grove, CA, 1993.

The Slope Mean and Its Invariance Properties

JUN JI
Kennesaw State University
Kennesaw, GA 30144
jji@kennesaw.edu

CHARLES KICEY
Valdosta State University
Valdosta, GA 31698
ckicey@valdosta.edu

For $a, b > 0$, we know that the arithmetic mean $A(a, b) = (a + b)/2$ produces the midpoint of the segment $[a, b]$ on the real line. But what if we interpret a and b as slopes? A more natural mean in this context could be the "intermediate" slope, specifically, the positive slope $S(a, b)$ of the line $y = S(a, b)x$ that bisects the angle formed by the lines $y = ax$ and $y = bx$. As $a, b > 0$ vary in the figure, one senses that $S(a, b)$ is different from $A(a, b)$, but nonetheless has characteristics often associated with a mean.



Originally we chanced upon this mean while statistically comparing various linear regression methods [**9**]. We randomly perturbed a set of points on a line of slope m, repeating many times. For a particular method, we computed an average slope $\hat{m}$ to compare with the underlying slope m. Once in a while a random sample of perturbed points produced a near vertical regression line; this was a problem: since we used the arithmetic mean to compute $\hat{m}$, the corresponding near-infinite slope would not be canceled by the near-zero slope of a near horizontal line. We felt that it would be more meaningful to compute $\hat{m}$ by identifying slopes to angles, which led us to consider the mean

$$S(x_1, x_2, \ldots, x_n) = \tan\left((\tan^{-1} x_1 + \tan^{-1} x_2 + \cdots + \tan^{-1} x_n)/n\right). \tag{1}$$

Before proceeding, we must carefully consider what we mean by *mean*. Typically a mean M is a function from $(0, \infty) \times (0, \infty) \times \cdots \times (0, \infty)$ to $(0, \infty)$ satisfying $\min\{x_1, x_2, \ldots, x_n\} \leq M(x_1, x_2, \ldots, x_n) \leq \max\{x_1, x_2, \ldots, x_n\}$ (intermediacy) and with an output value independent of the arrangement of input values (symmetry).

The classic means—the arithmetic, geometric, and harmonic—are defined respectively by

$$\begin{aligned} A(x_1, x_2, \ldots, x_n) &= (x_1 + x_2 + \cdots + x_n)/n, \\ G(x_1, x_2, \ldots, x_n) &= \sqrt[n]{x_1 x_2 \cdots x_n}, \\ H(x_1, x_2, \ldots, x_n) &= n/(1/x_1 + 1/x_2 + \cdots + 1/x_n). \end{aligned}$$

It is easy to see that these, as well as (1), are means as defined above. Many other means and families of means can be found in the vast literature on means [**1, 3, 6, 7, 10**].

Depending on focus and context, there is some variation in definition of *mean*. For example, sometimes continuity is assumed and other times symmetry is not required. We contend that often the definition is restricted to positive numbers for convenience and out of geometric tradition [**4**]. This requirement avoids problems in, for example, G and H, but is arguably unnaturally restrictive for A.

Define $D = \{(x_1, x_2, \ldots, x_n) : x_1, x_2, \ldots, x_n > 0 \text{ or } x_1, x_2, \ldots, x_n < 0\}$, and, for the purposes of this note, define a *mean* to be a function $M : D \to \mathbb{R}$ satisfying intermediacy and symmetry and

$$M(x_1, x_2, \ldots, x_n) = -M(-x_1, -x_2, \ldots, -x_n) \quad \text{for} \quad x_1, x_2, \ldots, x_n < 0. \tag{2}$$

If necessary, we extend a mean on $(0, \infty) \times (0, \infty) \times \cdots \times (0, \infty)$ to D by requiring (2).

Returning to the definition of S given by (1), we see that $S(x_1, x_2)$ returns the slope of the line that bisects the angle formed by lines with slopes x_1 and x_2. Because of this interpretation, we call S the *slope mean*, though the "slope to arc back to slope mean" would be a more apt description. In the literature, a special focus has been placed on homogeneous means. Although the slope mean is not homogeneous, we will see how it is closely related to the three classic means.

The notion of invariance Much attention is given to means M that are *homogeneous*, meaning that $M(\xi x_1, \ldots, \xi x_n) = \xi M(x_1, \ldots, x_n)$, for $\xi > 0$. We generalize this notion, and say that a mean M on D is *invariant* under a real valued function f if $M(f(x_1), f(x_2), \ldots, f(x_n)) = f(M(x_1, x_2, \ldots, x_n))$ for all $(x_1, x_2, \ldots, x_n) \in D$ for which both sides of the equality are defined. Thus, we will also call a homogeneous mean *scalar invariant*, due to the fact that it is invariant under $f_\xi(x) = \xi x$ for all $\xi \neq 0$.

It is straightforward to verify that the arithmetic mean is scalar invariant and invariant under translation given by $g_\tau(x) = x - \tau$, $\tau \in \mathbb{R}$. Moving on, it is also easily seen that the geometric mean is also scalar invariant. Moreover, it is invariant under reciprocation $r(x) = 1/x$. Other well-known means (see Eves' list [**4**, p. 200]) do not have this invariance, but the slope mean shares such a property with the geometric mean, as we will now show:

Fix $x_1, x_2, \ldots, x_n > 0$ and choose $\theta_1, \theta_2, \ldots, \theta_n \in (0, \pi/2)$ such that $\tan\theta_i = x_i$ for all $i = 1, 2, \ldots, n$. Then $\tan^{-1}(1/x_i) = \pi/2 - \theta_i$ for each i, which, together with the relations between $\tan x$ and $\cot x$, leads to

$$S(r(x_1), \ldots, r(x_n)) = \tan\Big(\big(\tan^{-1}(1/x_1) + \cdots + \tan^{-1}(1/x_n)\big)/n\Big)$$

$$= \cot\left(\left(\sum_{i=1}^{n} \tan^{-1} x_i\right)\Big/ n\right) = r\big(S(x_1, x_2, \ldots, x_n)\big).$$

We leave to the reader to check that the harmonic mean is scalar invariant but the slope mean is not. The relation between the arithmetic and harmonic means

$$A\big(r(x_1), r(x_2), \ldots, r(x_n)\big) = r\big(H(x_1, x_2, \ldots, x_n)\big) \tag{3}$$

will be used in the remainder of the note.

Comparison with the classic means The three classic means given in the introduction can be compared by the most frequently proven inequalities of classical analysis [**2, 5**]:

$$H(x_1, x_2, \ldots, x_n) \leq G(x_1, x_2, \ldots, x_n) \leq A(x_1, x_2, \ldots, x_n) \quad \text{for all} \quad x_i > 0. \tag{4}$$

The second inequality in (4) is the celebrated *Geometric-Arithmetic Mean Inequality*. The slope mean, like the geometric mean, is also trapped between A and H.

THEOREM 1. *For all $x_1, x_2, \cdots, x_n > 0$, we have*

$$H(x_1, x_2, \ldots, x_n) \leq S(x_1, x_2, \ldots, x_n) \leq A(x_1, x_2, \ldots, x_n).$$

Proof. The second inequality follows from the fact that $f(x) = \tan^{-1} x$ is concave down for $x > 0$. The first inequality follows from the invariance of S under $r(x) = 1/x$ and (3). ∎

The slope mean S is not scalar invariant, which in turn implies that a nontrivial family of means S_ξ can be introduced by

$$S_\xi(x_1, x_2, \ldots, x_n) = (1/\xi)S(\xi x_1, \xi x_2, \ldots, \xi x_n), \quad \text{for all} \quad \xi > 0. \tag{5}$$

The following result will further extend the assertion in Theorem 1 and connect the slope mean to the arithmetic and harmonic means.

THEOREM 2. *Let S_ξ be defined by* (5). *Then, for all $x_1, x_2, \cdots, x_n > 0$,*

(a) $H(x_1, x_2, \ldots, x_n) \le S_\xi(x_1, x_2, \ldots, x_n) \le A(x_1, x_2, \ldots, x_n)$ *for all* $\xi > 0$.

(b) $\lim_{\xi \to 0^+} S_\xi = A$ *and* $\lim_{\xi \to \infty} S_\xi = H$.

Proof. Part (a) follows from Theorem 1 and the fact that A and H are scalar invariant. Part (b) is easily verified by L'Hospital's rule. ∎

Throughout the rest of the note we will concentrate on means with $n = 2$. Let $a, b > 0$, thinking of a and b as slopes. In this case, (1) simplifies to give the mean $m = S(a, b)$ of a and b as $m = \tan((\tan^{-1} a + \tan^{-1} b)/2)$. Therefore, m satisfies $\tan(2\tan^{-1} m) = \tan(\tan^{-1} a + \tan^{-1} b)$. Applying the angle sum identity for tangent we end up with a quadratic equation in m whose positive root can be expressed as

$$m = S(a, b) = \left(ab - 1 + \sqrt{(a^2 + 1)(b^2 + 1)}\right) / (a + b), \tag{6}$$

under the assumption that $ab \neq 1$. If $ab = 1$, then (6) yields $S(a, b) = S(a, 1/a) = 1$, which is consistent with (1) and the geometric interpretation of S. Therefore, for all $a, b > 0$ (and in fact for all $(a, b) \in D$) the formula (6) provides the slope of the line that bisects the angle formed by two lines of slopes a and b.

Characterization by invariance The classic means A, G, and H are scalar invariant. Mathematically, it is interesting to determine the class of functions under which a given mean is invariant; we will also see that a mean is uniquely determined by the class of functions under which it is invariant. We will study characterizations by invariance for arithmetic, geometric, harmonic, and slope means.

As pointed out earlier, A is invariant under all the functions

$$f_\xi(x) = \xi x, \quad \xi \neq 0 \quad \text{and} \quad g_\tau(x) = x - \tau, \ \tau \in \mathbb{R}. \tag{7}$$

More importantly, A is determined uniquely by these two sets of invariances as follows.

We assume that M is any mean on D invariant under the functions given by (7). Fix $(a, b) \in D$, and let $m = M(a, b)$. Using symmetry and scalar invariance, we have $m = M(b, a) = -M(-b, -a)$. On the other hand, the translation invariance gives $M(-b, -a) = M(a - (a + b), b - (a + b)) = M(a, b) - (a + b) = m - (a + b)$. Therefore $2m = a + b$ or $M(a, b) = A(a, b)$. Thus, the arithmetic mean is the only mean invariant under scaling and translation.

Next, we turn our attention to the geometric mean G. It turns out that the invariances under $f_\xi(x) = \xi x, \xi \neq 0$ and $r(x) = 1/x$ are enough to determine G. We encourage the reader to prove this result by adapting the argument for the arithmetic mean, or otherwise.

Moving on, it is not hard to verify that H is invariant under $f_\xi(x) = \xi x, \xi \neq 0$ and $g_\tau(x) = x/(1 - \tau x), \tau \in \mathbb{R}$. (We used the relation (3) to find the second invariance family.) The proof of Theorem 3 below suggests a method to show that these two families of invariances in fact determine H.

To finish, we now turn our attention to the slope mean. Since we have already shown that S is invariant under $r(x) = 1/x$, we know that S cannot be invariant under scaling, otherwise $S = G$. Is there a second invariance that determines S?

Now is a good time to think geometrically. Let $a, b > 0$ and consider the three lines through the origin and containing the points $A(1, a)$, $B(1, b)$, and $C(1, S(a, b))$. It is easily seen that $1/a$, $1/b$ and $1/S(a, b)$, respectively, are the slopes of the same lines, but taken with respect to the y-axis. The geometric meaning of the slope mean indicates $S(1/a, 1/b)$ is also the slope of the line containing C with respect to the y-axis. Hence, invariance under r, $S(1/a, 1/b) = 1/S(a, b)$, is now geometrically obvious.

Continuing to think this way, we find another natural invariance for S: rotation by a fixed angle. After a little work, we have that S is invariant under

$$f_\rho(x) = (x + \rho)/(1 - \rho x),\ \rho \in \mathbb{R} \quad \text{and} \quad r(x) = 1/x. \tag{8}$$

Moreover, S is determined by these two invariances.

THEOREM 3. *If a mean $M : D \to \mathbb{R}$ is invariant under all the functions in* (8), *then $M = S$ on D.*

Proof. Assume $M : D \to \mathbb{R}$ is invariant under the functions in (8). Fix $a, b > 0$ and let $m = M(a, b)$. The key observation is that the system of equations $f_\rho(a) = r(b)$ and $f_\rho(b) = r(a)$ admits a solution, namely $\rho = (1 - ab)/(a + b)$. Using symmetry and the invariances r and f_ρ (where $\rho = (1 - ab)/(a + b)$), we have

$$m = M(a, b) = M(b, a) = \frac{1}{M(1/b, 1/a)} = \frac{1}{M(f_\rho(a), f_\rho(b))} = \frac{1}{f_\rho(m)},$$

where the last equality is valid provided that $f_\rho(m)$ is defined. But if $m = 1/\rho$, then the intermediacy of $m = M(a, b)$ leads either to $b^2 \le -1$ (when $a < b$) or $a^2 \le -1$, both of which are impossible.

Thus, $m = 1/f_\rho(m)$ or $m = -\rho \pm \sqrt{\rho^2 + 1}$. Applying intermediacy one more time and substituting for ρ yields

$$m = \frac{ab - 1 + \sqrt{(ab - 1)^2 + (a + b)^2}}{a + b}.$$

Thus, $m = S(a, b)$ given by (6). ∎

Note that the technique employed in Theorem 3 can also be used to prove those previously mentioned characteristic results through invariance for arithmetic, geometric, and harmonic means. Also note that the families of invariances are algebraic subgroups under composition of the group of fractional linear transformations (provided functions are considered equal if they differ at a finite number of points).

Further notes We have focused on quasi-arithmetic means, means of the form $f^{-1}((f(x_1) + f(x_2) + \cdots + f(x_n))/n)$. Whenever f is a monotone function defined on $(0, \infty)$, this generates a mean. In particular, taking $f(x)$ to be x, $\ln x$, $1/x$, and $\tan^{-1} x$ generates A, G, H, and S, respectively. It should be noted that some of the general theory of quasi-arithmetic means, developed in Hardy, Littlewood, and Pólya's *Inequalities* [**5**], can be applied to obtain and extend the comparison results above.

Besides the quasi-arithmetic means, there are other families of means containing the slope mean. A beautifully simple class of means, going back to 1933 [**8**], is generated by any function $f : (0, \infty) \to (0, \infty)$ as $(f(a)b + f(b)a)/(f(a) + f(b))$. Taking $f(x)$ to be 1, $\sqrt{x}$, and x generates A, G, and H, respectively. Here the slope mean is generated by $f(x) = \sqrt{x^2 + 1}$. More recently, Dietel and Gordon [**3**] generated

means from functions $f : (0, \infty) \to (0, \infty)$ and their tangent lines, which is a special case of the means in Horwitz [**6**]. Under the assumption that $f''(x)$ is nonzero and continuous, a mean is given by the x-coordinate of the intersection of the tangent lines to $y = f(x)$ at $x = a$ and $x = b$. In this case, taking $f(x)$ to be x^2, $\sqrt{x}$, and $1/x$ generates A, G, and H, respectively. The slope mean also belongs to this family, generated by $f(x) = \sqrt{x^2 + 1}$.

Some families of means do not contain the slope mean because the means are homogeneous, and yet there are still other families [**10**] where it is not clear whether or not the slope mean is a member.

In their study of the three classic means, Bullen, Mitrinović, and Vasić implicitly characterized these means through a family of functions as follows. Let $\{f_\lambda(x) : \lambda \in \mathbb{R}\}$ be a family of functions indexed by λ such that $f_\lambda^{-1}(x) = f_\lambda(x)$ and suppose that for every pair $(a, b) \in (0, \infty) \times (0, \infty)$, there exists a unique index $\lambda \equiv \lambda(a, b)$ such that $f_\lambda(a) = b$. Then $m = M(a, b)$ can be defined by $f_\lambda(m) = m$. It can be seen that our characterization of means by two sets of functions leads to such a characterization using a single family of functions.

Acknowledgment. We would like to thank all the anonymous referees for their comments and suggestions, which certainly improved the presentation of the paper.

REFERENCES

1. J. M. Borwein and P. B. Borwein, The way of all means, *Amer. Math. Monthly* **94** (1987), 519–522.
2. P. S. Bullen, D. S. Mitrinović, and P. M. Vasić, *Means and Their Inequalities*, D. Reidel Publishing Company, 1988.
3. B. C. Dietel and R. A. Gordon, *Using tangent lines to define means*, this MAGAZINE **76** (2003), 52–61.
4. H. Eves, *An Introduction to the History of Mathematics*, 6th ed., Saunders College Publishing, 1992.
5. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge, 1952.
6. A. Horwitz, Means and Taylor polynomials, *J. Math. Anal. Appl.* **149** (1990), 220–235.
7. M. E. Mays, Functions which parameterize means, *Amer. Math. Monthly* **90** (1983), 677–683.
8. D. Moskovitz, An alignment chart for various means, *Amer. Math. Monthly* **40** (1933), 592–596.
9. P. Sprent, *Models in Regression and Related Topics*, Methuen and Co., LTD, 1969.
10. K. B. Stolarsky, Generalizations of the logarithmic mean, this MAGAZINE **48** (1975), 87–92.

A Carpenter's Rule of Thumb

ROBERT FAKLER
University of Michigan-Dearborn
Dearborn, MI 48128
rfakler@umd.umich.edu

In an episode of the PBS television series "The New Yankee Workshop," host and master carpenter Norm Abram needed to construct a rectangular wooden frame as part of a piece of furniture he was building. After gluing and clamping four pieces of wood together to form a rectangle, he checked the rectangle for squareness by measuring the two diagonals to determine whether or not they were of equal length. Upon finding a small difference in the two measurements, he announced that he would "split the difference." He proceeded to carefully nudge the top corner of the frame at the end of the longer diagonal until his measuring tape indicated that its length was the average of his two original diagonal measurements. He then said he was satisfied that the frame

was square. After hearing this, I wondered if the frame was indeed square, that is, whether there was a right angle at each of the four corners of the frame.

To answer the above question, we need to solve the following problem: Suppose we have a parallelogram with sides of lengths a and b. Let x and y be the lengths of the two diagonals. If we square up this parallelogram by transforming it into a rectangle with side lengths a and b, what is the common length d for the two diagonals? FIGURE 1 shows our parallelogram and its squared up version.

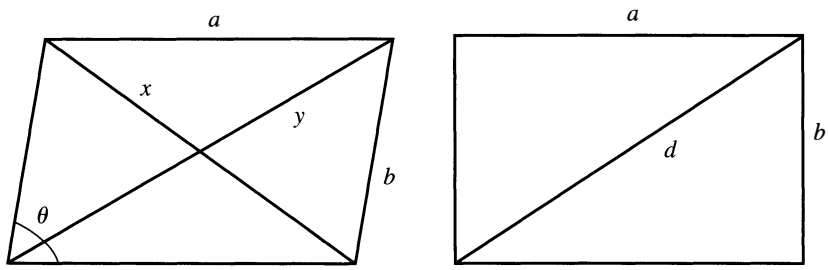


Figure 1 A wooden frame before and after straightening

From the Law of Cosines, we see that

$$x^2 = a^2 + b^2 - 2ab\cos\theta \quad \text{and} \quad y^2 = a^2 + b^2 - 2ab\cos(\pi - \theta).$$

Therefore

$$x^2 + y^2 = 2a^2 + 2b^2 \quad \text{and so} \quad d^2 = a^2 + b^2 = (x^2 + y^2)/2.$$

Thus

$$d = \sqrt{(x^2 + y^2)/2}.$$

This is evidently not the same as the average of the two original diagonal measurements, $(x + y)/2$, the length that Abram recommended for the new diagonal. His algorithm does not exactly produce the length needed to turn the parallelogram into a rectangle, but we will see that it is a good approximation.

What is the mathematical basis for this approximation? Consider the function $d(x, y) = \sqrt{(x^2 + y^2)/2}$, where $x > 0$ and $y > 0$. Suppose $L(x, y)$ is the linearization of d at the point (x_0, x_0), where $x_0 > 0$ is the correct measurement of the diagonal. (Note that $d(x_0, x_0) = x_0$.) Then

$$L(x, y) = d(x_0, x_0) + d_x(x_0, y_0)(x - x_0) + d_y(x_0, x_0)(y - x_0).$$

Since $d_x(x, y) = x/\sqrt{2x^2 + 2y^2}$ and $d_y(x, y) = y/\sqrt{2x^2 + 2y^2}$, we have $d_x(x_0, x_0) = d_y(x_0, x_0) = 1/2$. Also, $d(x_0, x_0) = x_0$. Therefore

$$L(x, y) = x_0 + (x - x_0)/2 + (y - x_0)/2 = (x + y)/2.$$

In the "The New Yankee Workshop" episode, Abram took the diagonal measurements x and y of the wooden frame. Instead of the exact diagonal length $d(x, y)$ that would square up the frame, which is difficult to compute mentally, he used the linear approximation $L(x, y)$, which is easy to calculate. In practice, the frame would be squared up by eye before its diagonals were measured so that x and y would be nearly equal and nearly equal to the exact diagonal x_0. In this case, the linearization $L(x, y)$ is a quite good approximation for $d(x, y)$, as the following example shows.

Example Suppose our diagonal measurements are $x = 36$ inches and $y = 37$ inches. Then $d(36, 37) = \sqrt{(36^2 + 37^2)/2} = 36.503424$ and $L(36, 37) = 36.5$.

The reader may wish to estimate the error in this linear approximation using Taylor's theorem in two variables [**1**]. If y is the longer diagonal and x the shorter, and if both $|x - x_0|$ and $|y - x_0|$ are known to be less than h, the error can be seen to be less than $y^2h^2/(2x^3)$.

REFERENCE

1. Jerrold E. Marsden and Michael J. Hoffman, *Elementary Classical Analysis*, 2nd ed., W. H. Freeman and Co., New York, 1993.

Chess: A Cover-Up

ERIC K. HENDERSON
DOUGLAS M. CAMPBELL
DOUGLAS COOK
ERIK TENNANT
Computer Science Department
Brigham Young University
Provo, Utah 84602-6576

The game of chess has always proved a rich source of interesting combinatorial problems to challenge mathematicians, logicians, and computer scientists. Apart from playing strategies and end-games, many chess-based problems have been posed over the centuries that tax the limits of symbolic reasoning, such as the n-queens and re-entrant knight's tour problems. However, the modern computer has enabled new approaches to these types of problems, and some of these questions have been explored (and even decided) in ways not previously possible.

One such problem has been attributed to Joseph Kling [**8**], a music producer who operated a chess-oriented coffee house in London from 1852. Kling, who migrated to England from Germany, is described as "a pioneer of the modern style of chess" [**5**]; he published several studies on the game including the popular but short-lived journal *Chess Player*, co-edited with the chess professional Bernhard Horwitz. Kling posed the following question: using a player's eight major chess pieces and no pawns, can all squares on the chessboard be covered (attacked)? At first glance the problem looks easy—the eight pieces collectively have more than enough attack power. Determining whether there is a solution, however, is nontrivial. No simple logical argument has yet been discovered.

Because a chessboard is small, the combinatorial size of chess problems is theoretically tractable. However, only the recent advances in computing power have made this true in practice. In 1989, Robison, Hafner, and Skiena [**8**] applied an exhaustive search to prove that no solutions exist to the Kling cover problem. To accomplish this using the computing power available at the time, they developed a novel technique for reducing, or "pruning," the number of solutions that need to be searched. Their approach, although not immediately intuitive, reduced the search space by more than 99.9%.

With the computing power available today, problems such as this can now be exhaustively searched in a reasonable time with no need for creative pruning of the search space. However, there will always be larger problems pushing the limits of computing power, and methods for reducing the search space help put more of these within reach.

To illustrate the application of computing techniques to perform exhaustive search, we revisit the Kling cover problem in depth. We first attempt to formalize it and present a clear definition of the scope of the problem. We then apply a standard computing algorithm to organize the search space, and present an analysis of three alternative pruning techniques. These techniques, although much less complex computationally and conceptually, give comparable reductions in the size of the search space. In our conclusions we discuss the limitations of exhaustive search techniques in the context of traditional proofs, and the role these methods may play in the future of scientific progress.

Bounding the search space In order to enumerate all the possible ways that the eight pieces may be positioned on the chessboard, we must first select an ordering of the pieces, that is, which piece to place first, second, and so on. Although all orderings produce equivalent sets of final chessboard configurations, certain orderings will prove to allow better optimization of the solution search. We use three different orderings that facilitate our pruning methods:

$$O_1 = [q, r_1, r_2, b_w, b_b, h_1, h_2, k]$$
$$O_2 = [q, r_1, r_2, h_1, h_2, b_w, b_b, k]$$
$$O_3 = [q, b_w, b_b, h_1, h_2, k, r_1, r_2],$$

where q is the position of the queen, r_1 and r_2 the position of the rooks, b_w the position of the bishop on a white square, b_b the position of the bishop on a black square, h_1 and h_2 the position of the knights, and k the position of the king. We indicate the position of a piece on the chessboard using a number from 1 to 64 that corresponds to the occupied square, as shown in FIGURE 1.

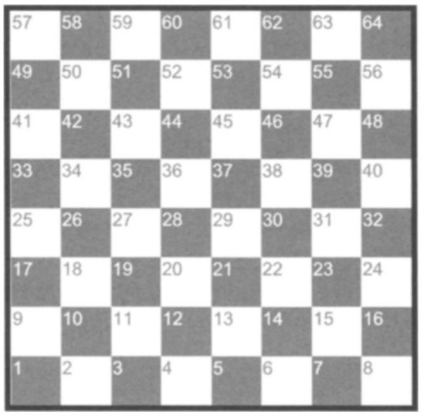


Figure 1 Numbered chessboard

We simplify the implementation by initially removing the restriction that pieces occupy distinct squares, allowing what we refer to as *superpositioning*. We define a *configuration vector* $\hat{c}$ as a vector that holds the numbers indicating the positions of each of the eight pieces relative to a given ordering, which we call a *configuration*. We call the set of all possible configuration vectors for a given ordering the *search space*. Clearly, the size of the search space is independent of the ordering used.

Let us first *bound* the size of the search space by determining the number of possible configuration vectors. Since there are 64 possible positions for each of the eight

pieces (a naïve upper bound that allows for superpositioning), the total number of configuration vectors is bounded by $64^8 = 2^{48}$. Kling's problem restricts one bishop to the 32 black squares and the other bishop to the 32 white squares (following the rules of chess), dropping the bound on the number of configurations to $32 \times 32 \times 64^6$. (FIGURE 2 shows a solution to Kling's problem if both bishops can be on the same color.)

Figure 2 A solution with both bishops on the same color

We can further reduce the bound on the number of configurations by noting that due to the symmetry of the board, some configurations are *equivalent*. Two configurations $\hat{c}_1$ and $\hat{c}_2$ are equivalent if one or more of the following three transformations takes $\hat{c}_2$ to $\hat{c}_1$:

1. *Interchanging* the position of r_1 and r_2 or h_1 and h_2 in $\hat{c}_2$
2. *Rotating* the position of all the pieces in $\hat{c}_2$ through angles of 90°, 180°, or 270° about the center of the board
3. *Reflecting* the position of all the pieces in $\hat{c}_2$ about one of the diagonal axes

Equivalent configurations attack the same number of squares, so it is only necessary to include one configuration in a set of equivalent configurations in the search space. By restricting the positions of certain pieces, we can prevent equivalent configurations from appearing in our search space.

First, we eliminate configurations equivalent due to interchanging rooks and knights by limiting the position of the second rook and second knight: we require that $r_1 \leq r_2$ and $h_1 \leq h_2$. This produces

$$\sum_{k=1}^{64} k = 65 \times 32$$

distinct positions for each of the two pairs of interchangeable pieces and reduces the upper bound on the number of configurations in the search space to

$$32 \times 32 \times (65 \times 32) \times (65 \times 32) \times 64^2.$$

Second, we eliminate configurations equivalent due to rotation. Each configuration has three equivalent forms (90°, 180°, and 270° rotations). We divide the board into four quadrants and limit the position of one of the pieces—in our experiments, the queen—to the lower-left quadrant. This restriction on the queen eliminates all

rotationally equivalent configurations and reduces the upper bound on the number of configurations by a factor of four to $32 \times 32 \times (65 \times 32) \times (65 \times 32) \times 16 \times 64$.

Third, we eliminate the configurations equivalent due to reflections. Observe: limiting the queen to the lower-left quadrant (as described above) also eliminated the diagonal symmetry along the upper-left to lower-right axis. We then limit the white bishop to the 16 white squares underneath the lower-left to upper-right diagonal to eliminate the diagonal symmetry along the lower-left to upper-right axis and reduce the upper bound on the number of configurations by a factor of two to

$$16 \times 32 \times (65 \times 32) \times (65 \times 32) \times 16 \times 64 = 2^{29} \times 65^2 = 2{,}268{,}279{,}603{,}200 \quad (1)$$

or about 2.27 trillion possible configurations in the search space.

Assuming 8 CPU cycles to check a configuration, a computer from the late 1980s running at 25 MHz would take on the order of one week to exhaustively search each of these configurations. Today, a computer running at 2 GHz could finish in less than three hours. If a pruning technique is employed that achieves a 99.9% prune rate, the running time can be reduced to less than 10 *seconds*.

Backtracking The search space can be organized as a tree with eight levels. We place the ith piece (relative to an ordering O_1, O_2, or O_3) on the ith level of the tree. The ith level contains *nodes* representing the positions that the ith piece can occupy on the chessboard. Each node is the parent of a sub-tree containing descendant nodes representing all the possible positions of the remaining unplaced pieces. A leaf node occupies the last level (level eight) in the tree. A *leaf node* represents a configuration where all pieces have been placed. FIGURE 3 illustrates the tree organization relative to O_1.

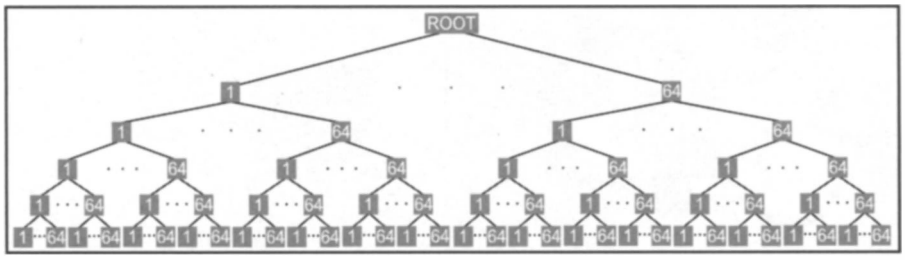


Figure 3 The first five levels of the tree organization for the search space. To simplify the illustration, pieces are allowed to occupy any of the 64 squares.

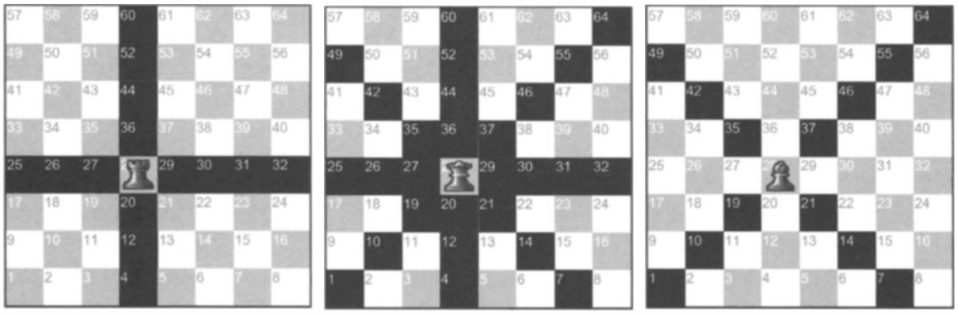
Let $\hat{c}[i]$ be a *partial configuration vector*, by which we mean a placement of the first i pieces according to one of the orderings O_1, O_2, or O_3. In the tree search space, a partial configuration is represented as an *internal node*, by which we mean any nonleaf node. The level of an internal node corresponds to the index i in its partial configuration vector $\hat{c}[i]$. Note that $\hat{c}[8]$, which is $\hat{c}$, can be interpreted as a path from the root to a leaf on the tree (relative to O_1, O_2, or O_3).

An algorithm commonly applied to a tree search space is *recursive backtracking* (also called *depth-first* traversal). Depth-first traversal means that we probe the children of a node x before we probe x's siblings. When we can probe no other child of a node, we backtrack and try the next child of its parent node, that is, its nearest sibling.

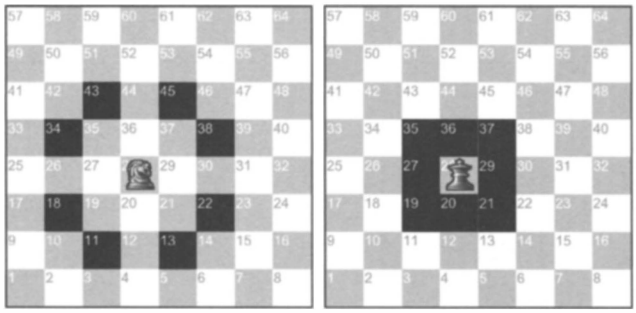
Backtracking itself is simply an ordered traversal of the search space, allowing for an exhaustive search of all nodes in the tree. However, by organizing the tree intelligently, it is possible to develop criteria for skipping the evaluation of a node's descendants. We refer to these as *prune rules*. If a prune rule for a node n determines that no

child of n can possibly be a solution, then node n is pruned and none of its descendants are visited. Effective prune rules can yield dramatic reductions in the number of nodes that need to be traversed.

Strong vs. weak solutions We define $a_p[s]$, the *attack pattern of a piece p relative to position s* to be the set of squares attacked, or "covered," when chess piece p is at position $s \in \{1 \ldots 64\}$, without regard to other pieces that may be on the board. FIGURE 4 shows an example of the attack pattern for each piece (relative to position 28). Observe that neither a king nor a knight's attack set is affected by intervening pieces. In contrast, the queen, bishop, and rook's attack patterns may be affected by an intervening piece, that is, a part of the attack pattern may be *blocked*. FIGURE 5 shows an example of a knight blocking a bishop's attack to square 15 and square 8.



$a_R[28]$ (Rook) $a_Q[28]$ (Queen) $a_B[28]$ (Bishop)



$a_H[28]$ (Knight) $a_K[28]$ (King)

Figure 4 Attack patterns for each piece relative to position 28

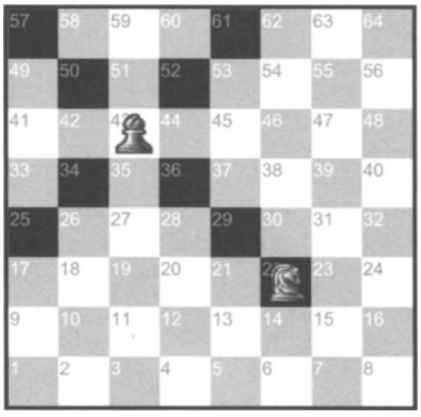


Figure 5 A knight blocking a bishop's attack on squares 15 and 8

Following Robison et al. [**8**], we say that configuration $\hat{c}$ *strongly covers* square s if and only if there is at least one piece in $\hat{c}$ whose attack pattern covers square s and whose attack pattern for square s is not blocked by any other piece. We say that configuration $\hat{c}$ *weakly covers* square s if and only if there is at least one piece in $\hat{c}$ whose attack pattern covers square s regardless of blocking.

A configuration $\hat{c}$ is a *strong (weak) solution* to the Kling problem if and only if (1) each of the 64 squares (including the eight squares occupied by the eight pieces) is *strongly (weakly) covered*, and (2) each piece is on a distinct square (no superpositioning).

A strong solution will also necessarily be a weak solution. It is much easier computationally to check for weak solutions then strong solutions, since blocking effects can be ignored. Thus, to search for strong solutions, it suffices to first compute the set of weak solutions, and then examine this much smaller set for strong solutions.

Three prune rules Let $|\hat{c}[i]|$ denote the cardinality of the set of chessboard squares *weakly* covered by the partial configuration $\hat{c}[i]$. Let $|\hat{c}[i]|_b$ ($|\hat{c}[i]|_w$) denote the cardinality of the set of black (white) chessboard squares weakly covered by $\hat{c}[i]$. Clearly, $|\hat{c}[i]|$ is a nondecreasing function of i, and

$$\left|\hat{c}[i]\right|_b + \left|\hat{c}[i]\right|_w = \left|\hat{c}[i]\right|.$$

PRUNE RULE 1. Order the pieces relative to O_1. Define $mwa[j]$ to be the maximum possible attack potential of the jth piece placed *anywhere* on the board. TABLE 1 gives these values for each piece. At node $\hat{c}_1[i]$, let

$$r[i] = \sum_{j=i+1}^{8} mwa[j].$$

Prune rule 1 is: if

$$\left|\hat{c}_1[i]\right| + r[i] < 64,$$

then prune the node. (That is, the node should be pruned if, at the ith level, the weakly attacked squares and the squares that the remaining pieces can attack cannot cover the board.)

TABLE 1: Maximum weak attack *mwa* potential per level j for pieces ordered as in O_1

Piece	j	$mwa[j]$
Queen	1	27
Rook	2	14
Rook	3	14
Bishop	4	13
Bishop	5	13
Knight	6	8
Knight	7	8
King	8	8

PRUNE RULE 2. Order the pieces relative to O_2, (the last three pieces being white bishop b_w at level 6, the black bishop b_b at level 7, and the king at level 8). Let $m_{black}[i]$ be the maximum number of black squares that can be attacked by the unplaced pieces *after* level i. Let $m_{white}[i]$ be the maximum number of white squares that can be attacked by the unplaced pieces *after* level i (see Table 2). Prune rule 2 is: if

$$\left|\hat{c}_2[i]\right|_b + m_{black}[i] < 32 \quad or \quad \left|\hat{c}_2[i]\right|_w + m_{white}[i] < 32,$$

then prune the node. (That is, the node should be pruned if, at the ith level, the black (white) squares weakly attacked plus the largest possible number of black (white) squares that the remaining pieces can attack is less than 32.)

TABLE 2: Maximum black and white attack potential after positioning level i. Note the asymmetry.

	Attack Potential	
Level i	$m_{black}[i]$	$m_{white}[i]$
1	49	49
2	41	41
3	33	33
4	25	25
5	17	17
6	17	4
7	4	4
8	0	0

PRUNE RULE 3. Order the pieces relative to O_3 (the king at level 6 and the rooks at levels 7 and 8). Let $m_{rows}[i]$ be the number of rows in $\hat{c}_3[i]$ containing three or more nonattacked squares (counting rows with three or more nonattacked squares ensures that the nonattacked squares are in more than two different columns). Since each rook can attack only one row, prune rule 3 is: if

$$m_{rows}[6] > 2 \quad or \quad m_{rows}[7] > 1$$

then prune the node.

Results We measured the effectiveness of our prune rules by running backtracking to traverse the entire search space four times: once with no prune rules, once with only prune rule 1, once with only prune rule 2, and once with only prune rule 3. All four times we found 813 weak solutions to the chess cover problem. It is easy to determine if any of the 813 weak solutions are strong solutions. In fact, none were, confirming that there are no strong solutions to the Kling Chess problem.

One interesting trend to observe is the position taken by the queen. Of the 813 weak solutions the queen occupies one of the four central squares 19, 20, 27, 28 (see FIGURE 1) in 71% of the cases. This central location maximizes the initial attack potential of the queen, and the trend appears to be even stronger when superpositioning is allowed. There are 8,715 weak solutions allowing for superpositioning, with the queen occupying one of the four central squares in 87% of the cases.

Of the 8,715 weak superposition solutions, 1,984 used only seven of the eight pieces. In each of these 1,984 cases, one knight was superimposed on the queen and the

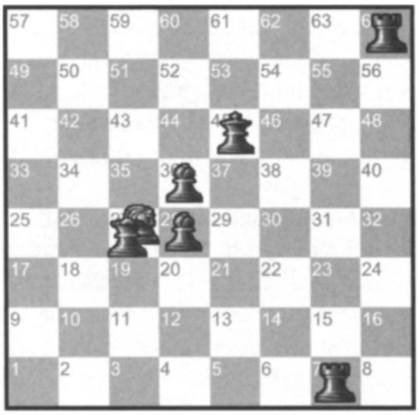


Figure 6 A weak superposition solution using only seven pieces. Notice the queen and a knight both occupy square 27.

second knight was not necessary. FIGURE 6 shows an example of a weak seven-piece superposition solution.

When we checked the 8,715 weak solutions in which superpositioning is allowed, 350 strong solutions were found. Of these 350 strong superposition solutions, the queen occupies one of the four central squares 19, 20, 27, 28 (see FIGURE 1) 97% of the time. In addition, each rook belongs to the outer perimeter (the set of squares that border the edge of the chessboard) 97% of the time. FIGURE 7 shows an example of a strong superposition solution with the queen on square 28 and the rooks on the outer perimeter.

Figure 7 A strong solution allowing superpositioning. Notice the queen and the knight both occupy square 28.

Upon examining the weak solutions, we discovered that the knights sometimes block more of the covered squares than they themselves attack (that is, the addition of a knight *reduced* the total number of squares strongly covered). This phenomenon occurred in 35% of the 813 weak solutions. The effective attack potential of the knights is much less than first appears and may offer the key to a traditional proof.

Effectiveness of the prune rules Equation (1) gives about 2.27 trillion complete configurations in the search space that we must explore. These correspond to leaf nodes

in the tree search space. We measure the effectiveness of a prune rule by determining how many of these complete configurations (leaf nodes) the prune rule removes from the search space. Recall that when a node is pruned, all of its children are pruned from the tree. Since each node is a progenitor to some number of leaf nodes (and internal nodes), pruning the node removes these leaf nodes (and internal nodes) from the search space.

Table 3 shows the results of using each of our prune rules on the Kling Chess Problem. For each prune rule we list the prune results for the three levels 5, 6, and 7. No prunes were recorded for the previous levels. For each level we record the number of prunes and the number of leaf nodes removed from the search space as a result of these prunes. The final column shows the number of leaf nodes removed as a percentage of the total number of leaf nodes in the search space (calculated from (1)). The last row under each prune rule shows the total results for all prunes.

TABLE 3: Results of the prune rules

Tree Level	Prunes	Leaf Nodes Pruned	% of Total Leaf Nodes
		Prune Rule 1	
5	500,796	66,665,963,520	2.94
6	565,321,622	1,211,853,924,352	53.43
7	15,346,724,150	982,190,345,600	43.3
Total	15,912,546,568	2,260,710,233,472	99.67
		Prune Rule 2	
5	2,551,017	83,591,725,056	3.69
6	979,235,545	2,005,474,396,160	88.41
7	2,757,235,103	176,463,046,592	7.78
Total	3,739,021,665	2,265,529,167,808	99.89
		Prune Rule 3	
5	0	0	0
6	1,069,115,100	2,223,759,408,000	98.04
7	706,198,686	24,748,310,328	1.09
Total	1,775,313,786	2,248,507,718,328	99.13

In all three cases our prune rules were most effective at level 6. At this level, prune rule 1 removed 53% of the leaf nodes, prune rule 2 removed 88% of the leaf nodes, and prune rule 3 removed 98% of the leaf nodes. Although each of the three prune rules removed very close to the same number of total leaf nodes (99%), they had very different numbers for total prunings. Prune rule 1 had 15.9 billion prunes, prune rule 2 had 3.7 billion prunes, and prune rule 3 had just 1.7 billion prunes.

The reason prune rule 3 could remove the same number of leaf nodes with much fewer prunes is because it was more effective at level 6, and could therefore remove more leaf nodes with fewer prunes than the other prune rules that had prunes at level 7. The higher a node is on the tree, the more leaf nodes it has as children. Prunes that occur at a higher level remove a larger number of leaf nodes than prunes at lower levels. For instance, prune rule 1 had 565 million prunes at level 6 and 15.3 billion prunes at level 7, yet the prunes at level 6 removed 19% more leaf nodes than the prunes at level 7. Prune rule 2 had 1,000 times more prunes at level 7 than at level 5, but these prunes amounted to just 2 times the number of leaf nodes removed. We were unable to devise a computationally feasible prune rule that could prune a node higher up in the tree than level 5.

A greedy solution The odds of randomly picking a weak solution are 813 in 2,268,279,603,200 or about 1 in 2.79 billion. One might wonder if a simple algorithmic design paradigm such as greedy, which is designed to maximize some parameter at each step in the algorithm, can stumble upon a weak solution. The answer is yes:

> *Take pieces in the order of configuration O_3 and greedily place each piece on a square that maximizes the number of weakly covered squares (break ties by choosing squares closest to the center).*

This greedy formulation produces the weak solution of FIGURE 8. Notice that the queen lies in the four central squares 19, 20, 27, 28 (see FIGURE 1) and the rooks lie on the perimeter.

Figure 8 The weak solution found using the greedy algorithm

Implementation Even with the massive amount of pruning we achieved, it was critical to have an optimized implementation to keep the runtime of the program reasonable. In our implementation, the data representation of a *board* $b[i]$, $1 \leq i \leq 64$, corresponding to the attack pattern $a_p[s]$ (piece p on square s) is an array of Booleans that represents the state of the squares on a chessboard numbered from 1 to 64 as in FIGURE 1. The Boolean $b[i]$ is true (one) if square i is attacked by piece p on square s; $b[i]$ is false (zero) if square i is not attacked by piece p on square s. We implement a *board* as eight contiguous bytes where each Boolean value is represented by one bit. To optimize the search space computation, we initially generate and store as *boards* the *attack patterns* for each of the pieces at every possible position.

Given a configuration $\hat{c}$ we define its *cover pattern*, as the union of the *attack patterns* in its configuration. The *cover pattern* is the result of logically OR-ing the appropriate *boards* representing these *attack patterns*. FIGURE 9 illustrates a cover pattern for the partial configuration with $a_K[18]$, $a_B[16]$, and $a_R[64]$. The chessboard in the figure shows graphically the squares that are attacked by this partial configuration. Next to this is the corresponding cover pattern shown numerically, where each 1 represents an attacked square and each 0 represents a nonattacked square. A configuration is a weak solution if and only if all squares in the representation of the corresponding *cover pattern* have a value of one ($|\hat{c}| = 64$).

To optimize the computation of $|\hat{c}|$, we interpret each row of the chessboard as a single byte. For a given configuration, each bit will be 0 or 1 depending on whether the corresponding chess square is safe or attacked. The value of this byte is then used as an index in a lookup table containing a count of the number of attacked squares,

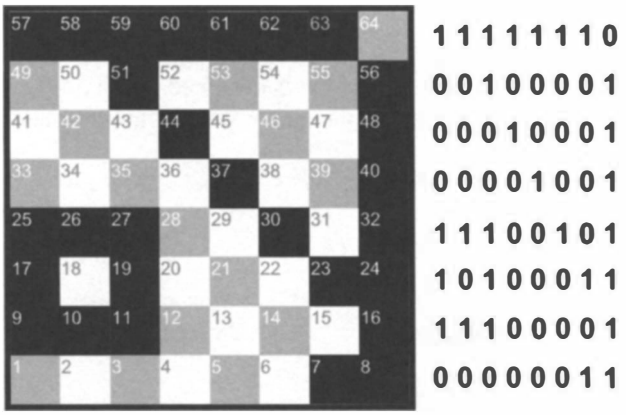


Figure 9 A cover pattern shown graphically and numerically. Attacked squares are shown in black.

which amounts to the number of ones for that particular binary pattern. For example, the top row in FIGURE 9 has all but the last square attacked and the corresponding byte value in this instance would be 254 (FE_{HEX}). This value is then used as an index to the lookup table entry containing the number 7, which corresponds to the number of binary ones in the byte value FE_{HEX}. The computation of $|\hat{c}|$, the number of attacked squares for a given configuration, can thus be computed with eight lookups and a sum, significantly reducing the computational complexity of the operation.

Conclusion We applied backtracking to exhaustively compute the Kling Cover Problem and presented three rules for reducing the search space. Our prune rules, though effective, used simple principles and did not operate close to the root of the tree. But our results revealed interesting trends that might be exploited to create more effective prune rules. Finding effective prune rules requires insight into the nature of the problem. The more effective a prune rule is, the more it begins to resemble a logical argument. A traditional proof that a problem has no solutions is essentially a prune rule that operates on the first node of the tree! Because effective prune rules are difficult to create for the Kling chess problem, we do not expect a general proof to be found easily. But finding more sophisticated prune rules that operate closer to the root may help provide the insight necessary for a traditional proof.

For some combinatorial problems like this one, applying the computing power available today requires little more than the most basic implementation, allowing us to construct an exhaustive proof with a limited understanding and no real insight into the problem itself—in short, using "brute force." In contrast, a traditional proof reveals the nature of a problem using symbols, words, and logic that can be verified by other human beings. A traditional proof appeals to us because it elegantly captures the truth we are attempting to explain. Exhaustive computation may provide us with an answer we can believe, but it leaves the question in some sense unresolved. In spite of this, it is a powerful tool and some classic problems have already benefited from this relatively new technique.

One of the earliest examples of incorporating exhaustive computation into a proof is the Haken-Appel proof of the four-color problem. The problem, which dates to 1852 [**9**], asks whether one can color the regions on an arbitrary two-dimensional map using only four colors, such that no two regions that share a border have the same color. This remained an open problem until 1976, when Wolfgang Haken and Kenneth Appel proved that any map could be shown to be equivalent to one of 1900 special

cases. Then, using a computer, they checked each special case [**1**]. Even though their proof consisted of a traditional logical argument, it made essential use of exhaustive computation.

Haken-Appel's proof has been disparagingly referred to as a "silicon proof," a type of proof for which only another computer can carry out the customary validity check. The mathematical community raised two objections. One objection was aesthetic: the proof failed to reveal a simple, single, fundamental understanding of the problem. A second objection was analytical: the program was thousands of lines long and depended on a compiler and operating system, which themselves had not been proven correct. Despite these limitations, the Haken-Appel proof remains the accepted solution to the four-color problem.

Applying exhaustive techniques to study, and perhaps solve, combinatorial problems is becoming more feasible with each advance in computing power. A traditional proof is naturally preferred to a proof by exhaustion, but some problems remain unsolved using traditional means leaving us to wonder if there are limits to purely symbolic methods. The computer presents us with a tool for attacking problems by exhaustion although its use is somewhat unsatisfying because it involves unverified components such as operating systems, compilers, and chip design. (The floating point error found in Intel's Pentium III processor is a good example of the problems unverified chip design can bring.) Are we as mathematicians going to be forced to give up verifying proofs by hand, just as scientists have been forced to accept from microscopes and telescopes evidence that cannot be directly perceived?

REFERENCES

1. K. Appel and W. Haken, Every planar map is four colorable, *Contemporary Mathematics* **98** (1989), American Mathematical Society, 1–741.
2. ———, The four color proof suffices, *Mathematical Intelligencer* **8** (1986), 10–20.
3. M. J. Beeson, *Foundation of Constructive Mathematics*, Springer-Verlag, Heidelberg, 1985.
4. Bledsoe and Loveland, *Automated Theorem Proving: After 25 Years*, American Mathematical Society, 1989.
5. F. Boase, *Modern English Biography*, Vol. 3, Truro, UK, 1897.
6. M. Gardner, *The Unexpected Hanging and Other Mathematical Diversions*, Simon and Schuster, New York, 1969.
7. ———, *Wheels, Life, and Other Mathematical Amusements*, W. H. Freeman, New York, 1983, pp. 183–193.
8. A. D. Robison, B. J. Hafner, and S. S. Skiena, Eight pieces cannot cover a chess board, *Comp. J.* **32** (1989) 567–570.
9. T. Saaty and P Kainen, *The Four Color Problem*, McGraw Hill, 1977.

50 Years Ago in the MAGAZINE

From "Science in the Modern World," by Marston Morse, Vol. **28**, No. 4, (Mar.–Apr., 1955), 209–211:

> Small wonder, then, that a large proportion of the young mathematicians become technicians in limited fields mostly connected with the foundations. Some leap over the foundations and proceed at once to the front as represented by the material world; these are the ones whom we call applied mathematicians.... Then there are the few—all too few—who aim to build the whole edifice of mathematics, neither lingering too long over the foundations, nor too hastily testing their strength at the front. Such a mathematician was Riemann, who, fifty years before Einstein, built the structure in mathematics whose counterpart in physics is relativity.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

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

Proposals

To be considered for publication, solutions should be received by September 1, 2005.

1716. *Proposed by Café Dalat Problem Solving Group, Washington D.C.*

Let k, n be integers with $n \geq 1$ and $0 \leq k \leq n$. Prove that there is an $n \times n$ matrix A of 0s and 1s with $\operatorname{per}(A) = k$. (Here $\operatorname{per}(A)$ denotes the permanent of A.)

1717. *Proposed by Mohammed Aassila, Strasbourg, France*

Let ABC be a triangle, and let A_1, B_1, C_1 be on BC, CA, AB, respectively, with none of A_1, B_1, C_1 coinciding with a vertex of ABC. Show that if

$$AB + BA_1 = AC + CA_1, \quad AB + AB_1 = BC + CB_1, \quad \text{and} \quad AC + AC_1 = BC + BC_1,$$

then

$$\operatorname{Area}(ABC) \geq 4 \cdot \operatorname{Area}(A_1B_1C_1).$$

1718. *Proposed by David Callan, Madison, WI.*

Let k, n be integers with $1 \leq k \leq n$. Prove the identity

$$\sum_{i=0}^{k-1} \binom{k-1}{i} \binom{n-(k-1)}{k-i} 2^{k-i-1} = \sum_{i=0}^{k-1} \binom{k-i}{i} \binom{n-i}{k}.$$

1719. *Proposed by G.R.A. 20 Problems Group, Università di Roma, Tor Vorgata, Rome, Italy.*

From an $(n+4) \times (n+4)$ checkerboard of unit squares, the central $n \times n$ square is removed to leave a square frame of width 2. In how many ways can the frame be

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 readers name, full address, and an e-mail address and/or FAX number.

tiled with 1×2 dominos? (Two different tilings that can be made identical through a rotation of the frame are considered different.)

1720. *Proposed by Stephen J. Herschkorn, Highland Park, NJ.*

Let X be a standard normal random variable and let a be a positive number. Show that the conditional expectation $E[X \mid |X - a| < t]$ is strictly decreasing in nonnegative t.

Quickies

Answers to the Quickies are on page 164.

Q949. *Proposed by Vasyl Dmytrenko and Felix Lazebnik, University of Delaware, Newark, DE.*

Let j, k, n be integers with $j, k \geq 0$ and $n \geq 1$. Prove that

$$\sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} \binom{ki}{j} = \begin{cases} 0, & 0 \leq j < n. \\ k^n, & j = n \end{cases}$$

Q950. *Proposed by William P. Wardlaw, U. S. Naval Academy, Annapolis, MD.*

Let F be a field and let B be a matrix over F. Suppose that B has characteristic polynomial $p_B(x) = x^3 - x$. Prove that there is no matrix A with entries in F such that B is the classical adjoint of A. (By classical adjoint we mean the transpose of the matrix of cofactors.)

Solutions

Extremes by Majorization **April 2004**

1691. *Proposed by Murray S. Klamkin, University of Alberta, Edmunton, Alberta, Canada.*

Let p, r, and n be integers with $1 < r < n$, and let k be a positive constant. Determine the maximum and minimum values of

$$\sum_{j=1}^{n} \frac{t_j^p}{1 + kt_j},$$

where $x_i \geq 0$, $1 \leq i \leq n$ with $x_1 + x_2 + \cdots + x_n = 1$, and $t_j = x_j + x_{j+1} + \cdots x_{j+r-1}$, where $x_{i+n} = x_i$.

Solution by the proposer.

The second derivative of $F(t) = t^p/(1 + kt)$ is

$$F''(t) = \frac{t^{p-2}\left(p(p-1) + 2kp(p-2)t + k^2(p-1)(p-2)t^2\right)}{(1+kt)^3}.$$

Thus F is concave for $p = 1$ and convex for $p \geq 2$ and $p \leq 0$.

Noting that for any choice of x_js we have $t_1 + t_2 + \cdots + t_n = r$, we apply a majorization result due to Hardy, Littlewood, and Pólya (see A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, NY, 1979).

Given a vector $\mathbf{y} = (y_1, \ldots, y_n)$, let $y_{[1]}, y_{[2]}, \ldots, y_{[n]}$ be the components of $\mathbf{y}$ in decreasing order. For vectors $\mathbf{y}$ and $\mathbf{z}$, write $\mathbf{y} \prec \mathbf{z}$ if $\sum_{j=1}^{k} y_{[j]} \le \sum_{j=1}^{k} z_{[j]}$ for $1 \le k < n$ and $\sum_{j=1}^{n} y_{[j]} = \sum_{j=1}^{n} z_{[j]}$. If g is convex on $[a, b]$, $\mathbf{y}, \mathbf{z} \in [a, b]^n$, and $\mathbf{y} \prec \mathbf{z}$, then $\sum_{j=1}^{n} g(y_j) \le \sum_{j=1}^{n} g(z_j)$.

For all choices of x_js we have

$$\left(\frac{r}{n}, \frac{r}{n}, \ldots, \frac{r}{n}\right) \prec (t_1, t_2, \ldots, t_n) \prec (1, 1, \ldots, 1, 0, \ldots, 0),$$

where the last n-tuple consists of r 1s followed by $(n - r)$ 0s. Thus, if $p = 1$ (so F is concave), we have

$$\sum_{i=1}^{n} F(t_i) \le nF\left(\frac{r}{n}\right) = \frac{nr}{n + kr} \quad \text{and} \quad \sum_{i=1}^{n} F(t_i) \ge rF(1) + (n - r)F(0) = \frac{r}{1 + k}.$$

If $p \ge 2$ or $p \le 0$ (so F is convex) then

$$\sum_{i=1}^{n} F(t_i) \ge nF\left(\frac{r}{n}\right) = \frac{r^p}{n^{p-2}(n + kr)},$$

and for $p \ge 2$ or $p = 0$,

$$\sum_{i=1}^{n} F(t_i) \le rF(1) + (n - r)F(0) = \frac{r}{1 + k}.$$

If $p < 0$, the sum is not bounded above.

Also solved by Chip Curtis, Minh Can, Enkel Hysnelaj (Australia), Stephen Kaczkowski, Elias Lampakis (Greece), and Li Zhou.

Fibonacci and Lucas Polynomials **April 2004**

1692. *Proposed by Mario Catalani, Department of Economics, University of Torino, Torino, Italy.*

Let $F_n = F_n(x, y)$ and $L_n = L_n(x, y)$ be the bivariate Fibonacci and Lucas polynomials, defined by

$$F_0 = 0, \quad F_1 = 1, \quad F_n = xF_{n-1} + yF_{n-2}, \quad n \ge 2$$
$$L_0 = 2, \quad L_1 = x, \quad L_n = xL_{n-1} + yL_{n-2}, \quad n \ge 2.$$

Assume that $x \ne 0$, $y \ne 0$, and $x^2 + 4y \ne 0$. Prove that

$$F_n\left(L_{2m+1}, y^{2m+1}\right) = \frac{F_{n(2m+1)}(x, y)}{F_{2m+1}(x, y)} \quad \text{and} \quad F_n\left(L_{2m}, -y^{2m}\right) = \frac{F_{2mn}(x, y)}{F_{2m}(x, y)}.$$

Solution by Daniele Donini, Bertinoro, Italy.

We first prove, by induction, that

$$F_{i+j} = L_i F_j + (-1)^{i+1} y^i F_{j-i} \quad \text{for} \quad j \ge i \ge 0. \tag{1}$$

For $i = 0$ and any $j \ge 0$, equation (1) reduces to $F_j = 2F_j - F_j$. For $i = 1$ expression (1) is $F_{j+1} = xF_j + yF_{j-1}$, which is also true for all $j \ge 1$. Now assume that (1) is true for $i = k - 1$ and $i = k$ for some $k \ge 1$. Then for $j \ge k + 1$ we have

$$\begin{aligned} L_{k+1}F_j + (-1)^{k+2}y^{k+1}F_{j-(k+1)} &= (xL_k + yL_{k-1})F_j + (-1)^k y^k (yF_{j-k-1}) \\ &= (xL_k + yL_{k-1})F_j + (-1)^k y^k (F_{j-k+1} - xF_{j-k}) \\ &= x(L_k F_j + (-1)^{k+1} y^k F_{j-k}) + y(L_{k-1}F_j + (-1)^k y^{k-1} F_{j-(k-1)}) \\ &= xF_{k+j} + yF_{k+j-1} = F_{k+j+1}. \end{aligned}$$

This establishes (1) for $i = k + 1$. Substituting $i = k$ and $j = k(n + 1)$ into (1) we obtain

$$F_{k(n+2)} = L_k F_{k(n+1)} + (-1)^{k+1} y^k F_{kn}, \quad k, n \geq 0. \tag{2}$$

Now fix k. Consider the sequences $\{G_n\}$ and $\{H_n\}$, defined by

$$G_n = F_k F_n \left(L_k, (-1)^{k+1} y^k\right) \quad \text{and} \quad H_n = F_{kn},$$

respectively. These sequences satisfy the same recurrence relation,

$$\begin{aligned} G_0 = 0, \quad G_1 = F_k, \quad G_{n+2} = L_k G_{n+1} + (-1)^{k+1} y^k G_n, \quad n \geq 0, \\ H_0 = 0, \quad H_1 = F_k, \quad H_{n+2} = L_k H_{n+1} + (-1)^{k+1} y^k H_n, \quad n \geq 0, \end{aligned}$$

where the last equality follows from (2). It follows that $G_n = H_n$ for all n, that is

$$F_{kn} = F_k F_n \left(L_k, (-1)^{k+1} y^k\right).$$

This is equivalent to the both of the desired equalities.

Also solved by Michel Bataille (France), Jany C. Binz (Switzerland), Brian Bradie, Enkel Hysnelaj (Australia), Harris Kwong, Rolf Richberg (Germany), Heinz-Jürgen Seiffert (Germany), Ricardo M. Torrejón, Chu Wenchang and Maglio Maria Rosaria (Italy), Li Zhou, and the proposer.

Rarely Equilateral **April 2004**

1693. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College of the City University of New York, Bronx, NY.*

Let $A = (p, q)$, $B = (p^2, q^2)$, and $C = (p^3, q^3)$ be the vertices of a nondegenerate triangle.

a. For how many pairs (p, q) is triangle ABC equilateral?

b. If p or q is rational, can triangle ABC be equilateral?

Solution by Robert L. Doucette, McNeese State University, Lake Charles, LA.

We show that there are only two pairs (p, q) for which the triangle can be equilateral and that there are no such pairs with p or q rational.

Triangle ABC is nondegenerate if and only if $p, q \notin \{0, 1\}$ and $p \neq q$. We have $AB = BC$ and $AC = BC$ if and only if (p, q) is a solution of the system

$$\begin{aligned} x^2(x-1)^2(x^2-1) + y^2(y-1)^2(y^2-1) = 0 \\ x^2(x-1)^2(2x+1) + y^2(y-1)^2(2y+1) = 0. \end{aligned}$$

If (p, q) is a solution to this system with $p, q \notin \{0, 1\}$, then

$$(p^2-1)(2q+1) = (q^2-1)(2p+1) \quad \text{from which} \quad (2pq + p + q + 2)(p - q) = 0.$$

Hence, $\triangle ABC$ is a nondegenerate equilateral triangle if and only if (p, q) is a solution of the system

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

For convenience, let $u = 2x + 1$ and $v = 2y + 1$. The system becomes

$$\begin{aligned} uv + 3 &= 0 \\ (u-1)^2(u-3)^2u + (v-1)^2(v-3)^2v &= 0. \end{aligned} \tag{1}$$

Letting $f(x) = (x-1)^2(x-3)^2x$, we see that solutions to (1) arise from solutions to

$$f(x) + f\left(-\frac{3}{x}\right) = 0. \tag{2}$$

For $x < -1$, $f(x) < -64$ and for $0 < x < 3$, $f(x) < 3$. It follows that for $x \in (-\infty, -1) \cup (0, 3)$ we have $f(x) + f(-3/x) < -64 + 3 < 0$. Working with $f'(x)$, it is not difficult to see that f is increasing on $(-\infty, 0)$ and on $(3, \infty)$. It follows that as a function of x, $f(-3/x)$ is increasing on $(-1, 0)$ and $(0, \infty)$. Therefore $f(x) + f(-3/x)$ is increasing on $(-1, 0)$ and $(3, \infty)$. Note that $f(-1) + f(3) = -64$ and that $f(x) + f(-3/x) \to \infty$ as $x \to 0^-$ or $x \to \infty$. Thus (2) has exactly two solutions x_0 and x_1 with $-1 < x_0 < 0$, $x_1 > 3$, and $x_0x_1 = -3$. It follows that (1) has exactly two solutions, $(u, v) = (x_0, x_1)$ and $(u, v) = (x_1, x_0)$.

Triangle ABC is a nondegenerate equilateral triangle if and only if

$$(p, q) = \left(\frac{1}{2}(x_0 - 1), \frac{1}{2}(x_1 - 1)\right) \approx (-0.87791, 1.4846)$$

or

$$(p, q) = \left(\frac{1}{2}(x_1 - 1), \frac{1}{2}(x_0 - 1)\right).$$

Equation (2) is equivalent to the equation $p(x) = 0$, where p is a monic polynomial of degree 10 with integer coefficients and constant term -243. By the rational root theorem, $p(x) = 0$ has no rational solution in the interval $(-1, 0)$. It follows that both x_0 and x_1 are irrational. Hence if p or q is rational, then $\triangle ABC$ cannot be equilateral.

Also solved by Roy Barbara (Lebanon), Michel Bataille (France), Jany C. Binz (Switzerland), John Christopher, Chip Curtis, Knut Dale (Norway), Daniele Donini (Italy), G.R.A.20 Problems Group (Italy), Mike Hitchman, Peter W. Lindstrom, H. T. Tang and T. Tsang, Ajaj A. Tarabay and Bassem B. Ghalayini, Dave Trautman, Li Zhou, and the proposer. There were two incorrect submissions.

A Radius of Convergence **April 2004**

1694. *Proposed by Óscar Ciaurri, Universidad de La Rioja, La Rioja, Spain.*

For $\alpha \geq 1$, a sequence $\{b_n\}_{n\geq 0}$ is defined by

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

A sequence $\{a_n\}_{n\geq 0}$ is then defined by $a_0 = 1$ and, for $n \geq 1$, by $\sum_{k=0}^{n} a_{n-k}b_k = 0$. Find the value of $\limsup_{n\to\infty} \sqrt[2n]{|a_n|}$.

Solution by Michael Goldenberg and Mark Kaplan, The Ingenuity Project, Baltimore Polytechnic Institute, Baltimore, MD.

For $-1 < x < 1$ let

$$h(x) = \left(\frac{1+ix}{1-ix}\right)^{\alpha} = \left(\sum_{p=0}^{\infty}\binom{\alpha}{p}(ix)^p\right)\left(\sum_{q=0}^{\infty}(-1)^q\binom{-\alpha}{q}(ix)^q\right)$$
$$= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}(-1)^{n-k}\binom{\alpha}{k}\binom{-\alpha}{n-k}\right)(ix)^n.$$

Define g for $-1 < x < 1$ by

$$g(x) = \frac{1}{2}(h(x) + h(-x)) = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{2n}(-1)^{2n-k}\binom{\alpha}{k}\binom{-\alpha}{2n-k}\right)(ix)^{2n} = \sum_{n=0}^{\infty} b_n x^{2n}.$$

Now let $x = \tan(\theta/2)$, where $-\pi/2 < \theta < \pi/2$. Then $h(x) = (e^{i\theta})^{\alpha}$ and, because g is real,

$$g(x) = \frac{e^{i\theta\alpha} + e^{-i\theta\alpha}}{2} = \cos(2\alpha \arctan x).$$

Because g is analytic and nonzero in a neighborhood of 0, $1/g(x)$ is also analytic in a neighborhood of 0. From the definition of a_n we have

$$\sum_{n=0}^{\infty} a_n x^{2n} = \frac{1}{g(x)} = \sec(2\alpha \arctan x).$$

The power series about 0 for $\sec z$ has radius of convergence $\pi/2$. It follows that the power series about 0 for $\frac{1}{g(x)}$ has radius of convergence $R = \tan(\frac{\pi}{4\alpha})$, and hence that

$$\limsup_{n\to\infty} \sqrt[2n]{|a_n|} = \frac{1}{R} = \cot\left(\frac{\pi}{4\alpha}\right).$$

Also solved by Michel Bataille (France), Daniele Donini (Italy), Rolf Richberg (Germany), Li Zhou, and the proposer.

Is it Real? **April 2004**

1695. *Proposed by Shalom Feigelstock, Bar-Ilan University, Ramat-Gan, Israel.*

A field F is a real field if -1 cannot be written as a sum of squares. Prove that a field F is a real field if and only if for every $n \times m$ matrix A with entries from F, $\text{rank}(A) = n$ implies that AA^T is invertible.

Solution by Jim Delany, California Polytechnic State University, San Luis Obispo, CA.

Let A be an $n \times m$ matrix of rank n. If the $n \times n$ matrix AA^T is not invertible, then there is a nonzero vector $\mathbf{u}$ for which $\mathbf{u}AA^T = \mathbf{0}$. Because A has rank n, $\mathbf{u}A \neq \mathbf{0}$. Let $\mathbf{v} = \mathbf{u}A = (x_1, \ldots, x_m)$ and find k such that $x_k \neq 0$. Now

$$\mathbf{v}\mathbf{v}^T = \mathbf{u}AA^T\mathbf{u}^T = 0,$$

so $0 = \mathbf{v}\mathbf{v}^T = \sum_{j=1}^{m} x_j^2$. Dividing by x_k^2 we find

$$\sum_{\substack{j=1\\ j\neq k}}^{m}\left(\frac{x_j}{x_k}\right)^2 = -1,$$

and the field is not a real field.

Conversely, if $x_1^2 + \cdots + x_m^2 = -1$, then the matrix $A = (x_1, \ldots, x_m, 1)$ has rank one while $AA^T = 0$.

Also solved by Abhishek Banerjee (India), Roy Barbara (Lebanon), Michel Bataille (France), Daniele Donini (Italy), Robert L. Doucette, Erin Emerson, Michael Goldenberg and Mark Kaplan, Eugene Herman, David E. Manes, Northwestern University Math Problem Solving Group, Li Zhou, and the proposer.

Answers

Solutions to the Quickies from page 159.

A949. *Solution 1, by the poser.*

We expand $((1+x)^k - 1)^n$ in two ways. We first write

$$\begin{aligned}\left((1+x)^k - 1\right)^n &= \sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}(1+x)^{ki} = \sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}\left(\sum_{j=0}^{ki}\binom{ki}{j}x^j\right)\\ &= \sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}\left(\sum_{j=0}^{kn}\binom{ki}{j}x^j\right)\\ &= \sum_{j=0}^{kn}\left(\sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}\binom{ki}{j}\right)x^j. \qquad (1)\end{aligned}$$

For the second expansion,

$$\left((1+x)^k - 1\right)^n = \left(\sum_{m=0}^{k}\binom{k}{m}x^m - 1\right)^n = x^n\left(\sum_{m=1}^{k}\binom{k}{m}x^{m-1}\right)^n. \qquad (2)$$

From (2) we see that in the expansion of $((1+x)^k - 1)^n$, the coefficient of x^k is 0 for $0 \le k \le n-1$, and k^n for $k = n$. The desired identity follows by comparing with the coefficients of these powers of x in (1).

Solution 2, from the editors.

Take balls numbered $1, 2, \ldots, kn$. Divide them into n sets with consecutive integer labels, $\{1, 2, \ldots, k\}$, $\{k+1, \ldots, 2k\}$, and so forth. Call these sets *flocks*.

We count the number of ways to pick j balls from exactly n different flocks; call this number W. The top term on the left is $\binom{kn}{j}$. This is the number of ways to pick j balls without regard to flock, and gives an overcount of W. To compensate for the overcount, subtract the term $\binom{n}{n-1}\binom{k(n-1)}{j}$, the number of ways to pick j balls from at most $n-1$ different flocks. However, this adjustment overcompensates for the number of ways to pick balls from $n-2$ flocks. To account for this add $\binom{n}{n-2}\binom{k(n-2)}{j}$. This, in turn, overcounts the number of ways to pick balls from at most $n-3$ flocks. Continuing with the inclusion/exclusion argument, we see that the lefthand side equals W.

On the other hand, if $j < n$, then there are no ways for each of the n flocks to be represented, so $W = 0$. If $j = n$, then there are k choices for the ball from each of the n flocks, so $W = k^n$.

A950. Because $p_B(x) = (x+1)x(x-1)$, B has characteristic roots $-1, 0$ and 1. Hence B is a singular 3×3 matrix of rank 2. Suppose that $B = \mathrm{adj}(A)$ for some matrix A. Then $AB = (\det A)I$ cannot be invertible, so $AB = 0$. Since $B = \mathrm{Adj}(A) \neq 0$, the rank of the 3×3 matrix A must be 2. This implies that the columns of B are in the one dimensional null space of A, which implies that the rank of B must be at most 1. This contradiction shows that B is not the classical adjoint of any matrix.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

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

Mandelbrot, Benoit B., and Richard L. Hudson, *The (mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward*, Basic Books, 2004; xxiv + 328 pp, $27.50. ISBN 0–465–04355–0. Mandelbrot, Benoit B., *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, Springer, 1997; x + 551 pp, $54.95. ISBN 0–387–98363–5. Olson, Steve, The genius of the unpredictable, *Yale Alumni Magazine* (November/December 2004) 36–43; `http://yalealumnimagazine.com/issues/current/mandelbrot.html` .

We celebrate this year the 100th anniversary of Einstein's marvelous year of papers on the size of molecules, light as composed of particles, special relativity theory, and the existence of atoms by measuring Brownian motion of particles in solution. Just five years earlier, Louis Bachelier had presented his thesis that financial markets can be described by the laws of Brownian motion. Bachelier's ideas sank from sight but were resurrected half a century later in what has become financial orthodoxy, with its key assumption that price changes in financial markets are normally distributed and statistically independent of past changes. This orthodoxy culminated in the Nobel Prize-winning Black-Scholes method of valuing options. But Benoit Mandelbrot, father of fractal geometry, strongly disputes those assumptions and asserts that price changes are correlated (with a long memory) and follow a power law of scaling (longer tails). The results, he says, are that "trouble runs in streaks," markets are riskier than portrayed, and financial "bubbles" are inevitable. He gives data to support his claims and offers a "multifractal" model instead. His new book is eminently readable and engaging; the older book reprints the original papers on which his theories of scaling in finance are based; and the *Yale Alumni Magazine* article is based on an interview with Mandelbrot.

Brodie, Josh, and Elyse Graham, Math profs link particle actions, human free will, *Daily Princetonian* (24 November 2004) 1–2; `http://www.dailyprincetonian.com/archives/2004/11/24/news/11569.shtml`. Collins, Simon, We're not alone in the universe of free will, *New Zealand Herald*, (26 January 2005), `http://www.nzherald.co.nz/index.cfm?c_id=5&ObjectID=10008051` .

Mathematics has a way of concretizing intuition about the everyday world; to take food as an example, we have the Pancake Theorem, the Ham Sandwich Theorem, and the No Free Lunch Theorem. Simon Kochen and John H. Conway (Princeton) have added another (non-food) instance, the Free Will Theorem: Given three assumptions, if even one person has free will, then the behavior of all particles in the universe is indeterminate. The first two assumptions are restatements of known phenomena in quantum mechanics, and the third is that information cannot cause instantaneous change at a distance. (Unfortunately, there is already a (conflicting?) No Free Will Theorem: To any nondeterministic finite automaton corresponds a deterministic one that accepts exactly the same input strings.) "[Kochen and Conway] have not yet decided when or where they will publish their results": Such inaction may be either an application of or a counterexample to the theorem, depending on whether they are waiting to exercise free will or are not free to do so!

Sharp, John, *Surfaces: Explorations with Sliceforms*, QED Books, 2004; x + 281 pp, $34.95 (P). ISBN 1–85853–201–9.

The Danish mathematician Olaus Henrici invented in the 19th century a type of model of a geometric surface that displays the surface in terms of slices. The simplest example is a model of two planes made by slicing each of two rectangles halfway through and slotting them together; a multi-plane version is often used to separate bottles in boxes. Author Sharp calls such models "sliceforms" and devotes this book to telling how to make sliceform models of quadric surfaces, surfaces of revolution, ruled surfaces, algebraic surfaces, and polyhedra, including using software to design the pieces needed.

Gelman, Andrew, and Deborah Nolan, *Teaching Statistics: A Bag of Tricks*, Oxford University Press, 2002; xv + 299 pp, $120, $45 (P). ISBN 0–19–857225–5, 0–19–857224–7.

Many, if not most, of the introductory statistics courses in the U.S. are taught by mathematicians who were not trained as statisticians. They—and the statisticians, too—will find this book an absolute gold mine of tested ideas for in-class activities, demonstrations, and group work for such a course, as well as for student projects. The "tricks" are organized and classified according to course topic. One-fifth of the book is devoted to similar activities for more advanced courses in decision theory and Bayesian statistics, survey sampling, mathematical probability, and mathematical statistics. The authors suggest how to make time for all (or some) of the activities; the activities promote motivation and teach concepts experientially, hence can substitute for some class discussion—but not for the great extent of practice in problem-solving that I find many of my students badly need: translating (from English to statistics), analyzing, calculating, and interpreting. The section of Notes tells where the data sets, data displays, and other materials came from, but beware: Several URLs no longer work, including those for sites of the authors, though you can still find the materials by rooting around.

Polya, G., *How to Solve It: A New Aspect of Mathematical Method*, Princeton University Press, 2004; xxvii + 253 pp, $16.95 (P).

Those of us who have been in the profession for more than a few years may not realize that our students do not always discover on their own what we take for granted in our own backgrounds. And so we must tell them: You need to be aware of this, try that, read a particular book—be aware of *Mathematics Magazine*, join the MAA—and read Polya's classic *How to Solve It*. The book, now with a new foreword by John H. Conway, has sold over a million copies since 1945. It teaches the reader how to understand a mathematical problem, devise a plan for it, carry out the plan, and reflect back on the result. In dialogue format with numerous examples from elementary algebra and geometry, it encourages a heuristic approach, whether the problem is a "problem to find" or a "problem to prove." And it encourages a sense of humor: "If you can't solve the proposed problem, solve an easier one." Every mathematics student should experience and *live* this book.

MacMahon, P.A., *New Mathematical Pastimes*, edited by Paul Garcia, Tarquin Reprints, 2004; xxx + 118 pp, £15 (P). ISBN 1–899618–64–3. Book in CD-ROM form with color versions of the diagrams and additional material, £15. ISBN 1–185853220–5. (From QED Books, Pentagon Place, 195b Berkhamsted Road, Chesham, Bucks HP5 3AP, United Kingdom; `http://www.mathsite.co.uk/Home/Mathematics/Recreational%20Mathematics`.)

Percy A. MacMahon (1854–1929) is known for work in combinatorics (his combinatorics books are still in print). This book was issued on the occasion of a conference commemorating the 150th anniversary of his birth. It returns to print his major work in recreational mathematics, whose main topics are how to construct repeating patterns and how to construct and solve edge-matching puzzles in two and three dimensions (often called Vess puzzles). The book's figures are in black and white; the CD-ROM version contains colorized versions and additional resources. (Thanks to Antony Unwin, Augsburg University.)

33rd United States of America Mathematical Olympiad
April 27 and 28, 2004

edited by Titu Andreescu, Zuming Feng, and Po-Shen Loh

Problems

1. Let $ABCD$ be a quadrilateral circumscribed about a circle, whose interior and exterior angles are at least $60°$. Prove that

$$\frac{1}{3}|AB^3 - AD^3| \leq |BC^3 - CD^3| \leq 3|AB^3 - AD^3|.$$

   When does equality hold?

2. Suppose $a_1, \ldots, a_n$ are integers whose greatest common divisor is 1. Let S be a set of integers with the following properties:
   (a) For $i = 1, \ldots, n$, $a_i \in S$.
   (b) For $i, j = 1, \ldots, n$ (not necessarily distinct), $a_i - a_j \in S$.
   (c) For any integers $x, y \in S$, if $x + y \in S$, then $x - y \in S$.

   Prove that S must be equal to the set of all integers.

3. For what real values of $k > 0$ is it possible to dissect a $1 \times k$ rectangle into two similar, but incongruent, polygons?

4. Alice and Bob play a game on a 6 by 6 grid. On his or her turn, a player chooses a rational number not yet appearing in the grid and writes it in an empty square of the grid. Alice goes first and then the players alternate. When all squares have numbers written in them, in each row, the square with the greatest number in that row is colored black. Alice wins if she can then draw a line from the top of the grid to the bottom of the grid that stays in black squares, and Bob wins if she can't. (If two squares share a vertex, Alice can draw a line from one to the other that stays in those two squares.) Find, with proof, a winning strategy for one of the players.

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

$$(a^5 - a^2 + 3)(b^5 - b^2 + 3)(c^5 - c^2 + 3) \geq (a + b + c)^3.$$

6. A circle ω is inscribed in a quadrilateral $ABCD$. Let I be the center of ω. Suppose that

$$(AI + DI)^2 + (BI + CI)^2 = (AB + CD)^2.$$

   Prove that $ABCD$ is an isosceles trapezoid.

Note: For interested readers, the editors recommend the *USA and International Mathematical Olympiads 2004*. There many of the problems are presented together with a collection of remarkable solutions developed by the examination committees, contestants, and experts, during or after the contests.

Solutions

1. By symmetry, we only need to prove the first inequality. Because quadrilateral $ABCD$ has an incircle, we have $AB - AD = BC - CD$. It suffices to prove that $(AB^2 + AB \cdot AD + AD^2)/3 \le BC^2 + BC \cdot CD + CD^2$. By the given condition, $60^\circ \le \angle A, \angle C \le 120^\circ$, and so $1/2 \ge \cos A, \cos C \ge -1/2$. Applying the Law of Cosines to triangle ABD yields $BD^2 = AB^2 - 2AB \cdot AD \cos A + AD^2 \ge AB^2 - AB \cdot AD + AD^2 \ge (AB^2 + AB \cdot AD + AD^2)/3$. Equality holds if and only if $AB = AD$. On the other hand, applying the Law of Cosines to triangle BCD yields $BD^2 = BC^2 - 2BC \cdot CD \cos C + CD^2 \le BC^2 + BC \cdot CD + CD^2$. Combining the above inequalities gives the desired result. For the equality case, we must have $AB = AD$. This condition is also sufficient, because all the entries in the equalities are 0. Thus, equality holds if and only if $ABCD$ is a *kite* with $AB = AD$ and $BC = CD$.

2. For integers $a_1, \ldots, a_n \in \mathbb{Z}$ with greatest common divisor 1, we say that S *is generated by* $a_1, \ldots, a_n$ if conditions (a), (b), (c) in the problem hold. By the given conditions, we can easily deduce that if S is generated by $a_1, \ldots, a_n$, then

   (d) $0 = a_1 - a_1 \in S$ by (b).

   (e) $-s = 0 - s \in S$ whenever $s \in S$, by (a) and (d).

   It is then not difficult to show that

   - *fact 1.* If S is generated by $a_1, \ldots, a_n$, then S is generated by $a_1, a_2 - a_1, \ldots, a_n - a_1$.
   - *fact 2.* If S is generated by $a_1, \ldots, a_n$, then S is generated by $-a_1, a_2, \ldots, a_n$.

   Now suppose S is generated by $a_1, \ldots, a_n$ (and that none of the a_i are zero, without loss of generality); by fact 2, we may assume without loss of generality that $a_i > 0$ for each i. Choose integers $b_1, \ldots, b_k > 0$ with greatest common divisor 1 such that S is generated by $b_1, \ldots, b_k$ and $b_1 + \cdots + b_k$ is as small as possible. Note that the b_i must all be distinct (otherwise we could have omitted one), so we may assume without loss of generality that b_1 is smaller than the others.

   Suppose $k > 1$, and put $c_1 = b_1$ and $c_s = b_s - b_1$ for $s = 2, \ldots, k$. Then $\gcd(c_1, \ldots, c_k) = \gcd(b_1, \ldots, b_k) = 1$, and S is generated by $c_1, \ldots, c_k$ by fact 1. But $c_1 + \cdots + c_k = (b_1 + \cdots + b_k) - (k-1)b_1 < b_1 + \cdots + b_k$, contradiction. Hence $k = 1$ and $b_1 = 1$.

   All that remains is to check that if S is generated by 1, then $S = \mathbb{Z}$. We show that $0, 1, \ldots, k \in S$ for all positive integers k, by induction on k. Note that $-1, 0, 1 \in S$ by (d) and (e), so the base case $k = 1$ is okay. As for the induction step, if $0, 1, \ldots, k \in S$, then $k + 1 = k - (-1) \in S$ by (c). Thus the induction goes through, and all nonnegative integers are in S. By (e), all negative integers are also is in S. Hence $S = \mathbb{Z}$, and we are done.

3. We will show that a dissection satisfying the requirements of the problems is possible if and only if $k \neq 1$.

   We first show by contradiction that such a dissection is not possible when $k = 1$. Assume that we have such a dissection. The common boundary of the two dissecting polygons must be a single broken line connecting two points on the boundary of the square (otherwise either the square is subdivided in more than two pieces or one of the polygons is inside the other). The two dissecting polygons must have the same number of vertices. They share all the vertices on the common boundary, so they have to use the same number of corners of the square as their own vertices. Therefore, the common boundary must connect two opposite sides of the square

(otherwise one of the polygons will contain at least three corners of the square, while the other at most two). However, this means that each of the dissecting polygons must use an entire side of the square as one of its sides, and thus each polygon has a side of length 1. A side of longest length in one of the polygons is either a side on the common boundary or, if all those sides have length less than 1, it is a side of the square. But this is also true of the other polygon, which means that the longest side length in the two polygons is the same. This is impossible since they are similar but not congruent, so we have a contradiction.

We now construct a dissection satisfying the requirements of the problem when $k \neq 1$. Notice that we may assume that $k > 1$, because a $1 \times k$ rectangle is similar to a $1 \times 1/k$ rectangle.

We first construct a dissection of an appropriately chosen rectangle (denoted by $ABCD$ below) into two similar incongruent polygons. The construction depends on two parameters (n and r below). By appropriate choice of these parameters we show that the constructed rectangle can be made similar to a $1 \times k$ rectangle, for any $k > 1$. The construction follows.

Let $r > 1$ be a real number. For any positive integer n, consider the following sequence of $2n + 2$ points: $A_0 = (0, 0)$, $A_1 = (1, 0)$, $A_2 = (1, r)$, $A_3 = (1 + r^2, r)$, $A_4 = (1 + r^2, r + r^3)$, $A_5 = (1 + r^2 + r^4, r + r^3)$, and so on, until

$$A_{2n+1} = (1 + r^2 + r^4 + \cdots + r^{2n},\ r + r^3 + r^5 + \cdots + r^{2n-1}).$$

Define a rectangle $ABCD$ by $A = A_0$, $C = A_{2n+1}$,

$$B = (1 + r^2 + \cdots + r^{2n},\ 0), \quad \text{and} \quad D = (0,\ r + r^3 + \cdots + r^{2n-1}).$$

The sides of the $(2n + 2)$-gon $A_1A_2 \cdots A_{2n+1}B$ have lengths

$$r,\ r^2,\ r^3,\ \ldots,\ r^{2n},\ r + r^3 + r^5 + \cdots + r^{2n-1},\ r^2 + r^4 + r^6 + \cdots + r^{2n},$$

and the sides of the $(2n + 2)$-gon $A_0A_1A_2 \ldots A_{2n}D$ have lengths

$$1,\ r,\ r^2,\ \ldots,\ r^{2n-1},\ 1 + r^2 + r^4 + \cdots + r^{2n-2},\ r + r^3 + r^5 + \cdots + r^{2n-1},$$

respectively. These two polygons dissect the rectangle $ABCD$ and, apart from orientation, it is clear that they are similar but incongruent, with coefficient of similarity $r > 1$. The rectangle $ABCD$ and its dissection are thus constructed.

The rectangle $ABCD$ is similar to a rectangle of size $1 \times f_n(r)$, where

$$f_n(r) = \frac{1 + r^2 + \cdots + r^{2n}}{r + r^3 + \cdots + r^{2n-1}}.$$

It remains to show that $f_n(r)$ can assume any value $k > 1$ for appropriate choices of n and r. Choose n sufficiently large so that $1 + 1/n < k$. Since

$$f_n(1) = 1 + \frac{1}{n} < k < k1 + k^2 + k^4 + \ldots + k^{2n}k^2 + k^4 + \ldots + k^{2n} = f_n(k)$$

and $f_n(r)$ is a continuous function for positive r, there exists an r such that $1 < r < k$ and $f_n(r) = k$, so we are done.

4. Bob has a winning strategy. Let (i, j) denote the square in the ith row and jth column. Define set $A = \{(4, 1), (4, 2), (5, 1), (5, 2), (5, 3), (6, 1), (6, 2), (6, 3)\}$ and $B = \{(1, 4), (1, 5), (1, 6), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5), (3, 6), (4, 5)\}$. We claim that after each of his moves, Bob can insure that the maximum number in each row is a square in the set $A \cup B$. Based on our claim, Bob can make

sure that when all the numbers are written, the maximum square in row 1 is in B and the maximum square in row 6 is in A. Since there is no path from B to A that stays in $A \cup B$, Bob wins.

Now we prove our claim. Bob pairs each square of $A \cup B$ with a square in the same row that is not in $A \cup B$, so that each square of the grid is in exactly one pair. Whenever Alice plays in one square of a pair, Bob will play in the other square of the pair on his next turn. If Alice moves with x in $A \cup B$, Bob writes y with $y < x$ in the paired square. If Alice moves with x not in $A \cup B$, Bob writes z with $z > x$ in the paired square in $A \cup B$. So after Bob's turn, the maximum of each pair is in $A \cup B$, and thus the maximum of each row is in $A \cup B$.

5. For any positive number x, the quantities $x^2 - 1$ and $x^3 - 1$ have the same sign. Thus, we have $0 \le (x^3 - 1)(x^2 - 1) = x^5 - x^3 - x^2 + 1$, or $x^5 - x^2 + 3 \ge x^3 + 2$. It suffices to show that $(a^3 + 2)(b^3 + 2)(c^3 + 2) \ge (a + b + c)^3$. Expanding both sides of the last inequality and applying $a^3 + a^3b^3 + 1 \ge 3a^2b$ and its analogous forms and $a^3b^3c^3 + a^3 + b^3 + c^3 + 1 + 1 \ge 6abc$ gives the desired result. (We can also using $a^3 + 2 = a^3 + 1 + 1$ and its analogous forms and applying either Hölder's or Cauchy-Schwarz inequalities to finish the proof.)

6. The key is to recognize that the given identity is a combination of equality cases of certain inequalities. By equal tangents, we have $AB + CD = AD + BC$ if only if $ABCD$ has an incenter. We will prove that for a convex quadrilateral $ABCD$ with incenter I, then

$$(AI + DI)^2 + (BI + CI)^2 \le (AB + CD)^2 = (AD + BC)^2, \qquad (*)$$

and equality holds if and only if $AD \parallel BC$ and $AB = CD$.

Because circle ω is inscribed in $ABCD$, we can set $\angle DAI = \angle IAB = a$, $\angle ABI = \angle IBC = b$, $\angle BCI = \angle ICD = c$, $\angle CDI = \angle IDA = d$, and $a + b + c + d = 180°$. Our proof is based on the following key lemma.

LEMMA. *If a circle ω, centered at I, is inscribed in a quadrilateral $ABCD$, then*

$$BI^2 + \frac{AI}{DI} \cdot BI \cdot CI = AB \cdot BC. \qquad (\ddagger)$$

Proof. Construct a point P outside of the quadrilateral such that triangle ABP is similar to triangle DCI. We obtain that $\angle PAI + \angle PBI = \angle PAB + \angle BAI + \angle PBA + \angle ABI = a + b + c + d = 180°$, implying that the quadrilateral $PAIB$ is cyclic. By Ptolemy's theorem, we have $AI \cdot BP + BI \cdot AP = AB \cdot IP$, or

$$BP \cdot \frac{AI}{IP} + BI \cdot \frac{AP}{IP} = AB. \qquad (\dagger)$$

Because $PAIB$ is cyclic, $\angle IPB = \angle IAB = a$, $\angle API = \angle ABI = b$, $\angle AIP = \angle ABP = c$, and $\angle PIB = \angle PAB = d$. Thus triangles AIP and ICB are similar, implying that $AI/IP = IC/CB$ and $AP/IP = IB/CB$. Substituting the above equalities into the identity (†), we arrive at

$$BP \cdot CI + BI^2 = AB \cdot BC. \qquad (\dagger')$$

Note also that triangle BIP and triangle IDA are similar, implying that $BP/BI = IA/ID$, or $BP = (AI/ID) \cdot IB$. Substituting the above identity back into (†′) gives the desired relation (∗), establishing the lemma.

Now we prove our main result. By the lemma and symmetry, we have

$$CI^2 + \frac{DI}{AI} \cdot BI \cdot CI = CD \cdot BC. \qquad (\ddagger')$$

Adding the two identities (‡) and (‡′) gives

$$BI^2 + CI^2 + \left(\frac{AI}{DI} + \frac{DI}{AI}\right) BI \cdot CI = BC(AB + CD).$$

By the AM-GM Inequality, we have $AI/DI + DI/AI \geq 2$. Thus

$$BC(AB + CD) \geq IB^2 + IC^2 + 2IB \cdot IC = (BI + CI)^2,$$

where the equality holds if and only if $AI = DI$. Likewise, we have

$$AD(AB + CD) \geq (AI + DI)^2,$$

where the equality holds if and only if $BI = CI$. Adding the last two identities gives the desired inequality $(*)$ from the very beginning.

By the given condition in the problem, all the equalities in the above discussion must hold, that is, $AI = DI$ and $BI = CI$. Consequently, we have $a = d$, $b = c$, and so $\angle DAB + \angle ABC = 2a + 2b = 180°$, implying that $AD \parallel BC$. It is not difficult to see that triangle AIB and triangle DIC are congruent, implying that $AB = CD$. Thus, $ABCD$ is an isosceles trapezoid.

Poem: Stopping by Euclid's Proof of the Infinitude of Primes

(with apologies to Robert Frost)

Whose proof this is I think I know.
I can't improve upon it, though;
You will not see me trying here
To offer up a better show.

His demonstration is quite clear:
For contradiction, take the mere
n primes (no more), then multiply;
Add one to that . . . the end is near.

In vain one seeks a prime to try
To split this number—thus, a lie!
The first assumption was a leap;
Instead, the primes will reach the sky.

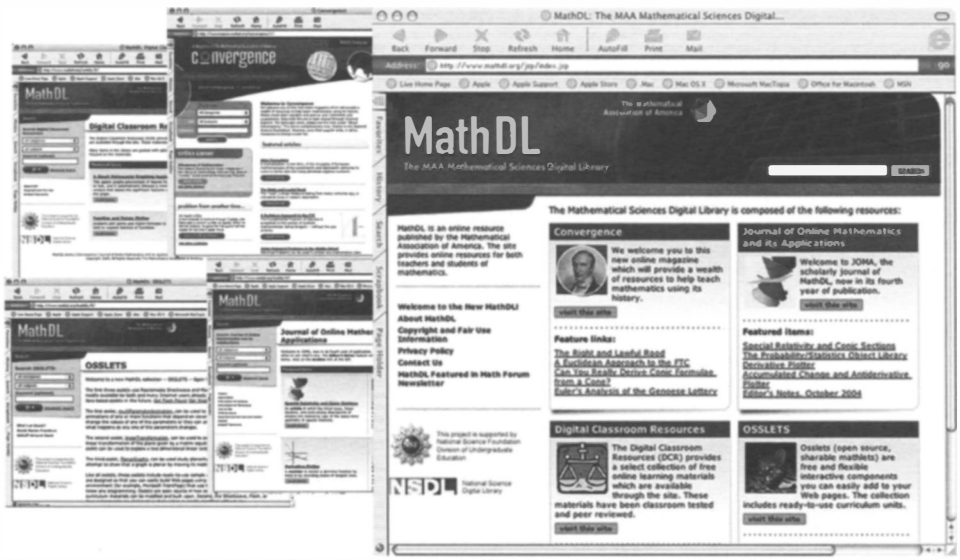
This proof is lovely, sharp, and deep,
But I have promises to keep,
And tests to grade before I sleep,
And tests to grade before I sleep.

——BRIAN D. BEASLEY
PRESBYTERIAN COLLEGE
CLINTON, SC 29325
bbeasley@mail.presby.edu

MathDL www.mathdl.org

The MAA Mathematical Sciences Digital Library

MathDL is an online resource published by the Mathematical Association of America. The web-site provides online resources for teachers and students of mathematics.



Components of the MathDL site:

Convergence - an online magazine in which mathematics, history and teaching interact. It seeks contributions for all departments, and particularly welcomes brief lesson plans demonstrating the use of history in the teaching of grades 9-14 mathematics.The online environment includes interesting graphics and interactive segments.

JOMA (Journal of Online Mathematics and its Applications) - the scholarly journal of MathDL is now in its fourth year. JOMA is a peer-reviewed journal devoted to advances in undergraduate mathematics education. It publishes articles addressed to faculty, modules for student learning, mathlets (single-purpose dynamic interactions) for classroom use or self-study, reviews of web-based materials, and developers' area contributions for those interested in designing and building their own computer-based materials.

DCR (Digital Classroom Resources) - a collection of peer-reviewed computer-based activities for use in the mathematics classroom. The current material spans a wide range of formats and platforms. DCR is always prepared to push the boundaries with new types of resources. The commom ingredients in all Digital Classroom Resources are student-centered interactivity, sound teaching pedagogy, and classroom / peer testing of material.

OSSLETS (Open Source, Sharable Mathlets) - free and flexible interactive components you can easily add to your web pages.

MathDL is a Collections Project supported by the NSF (DUE-0085861).

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

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