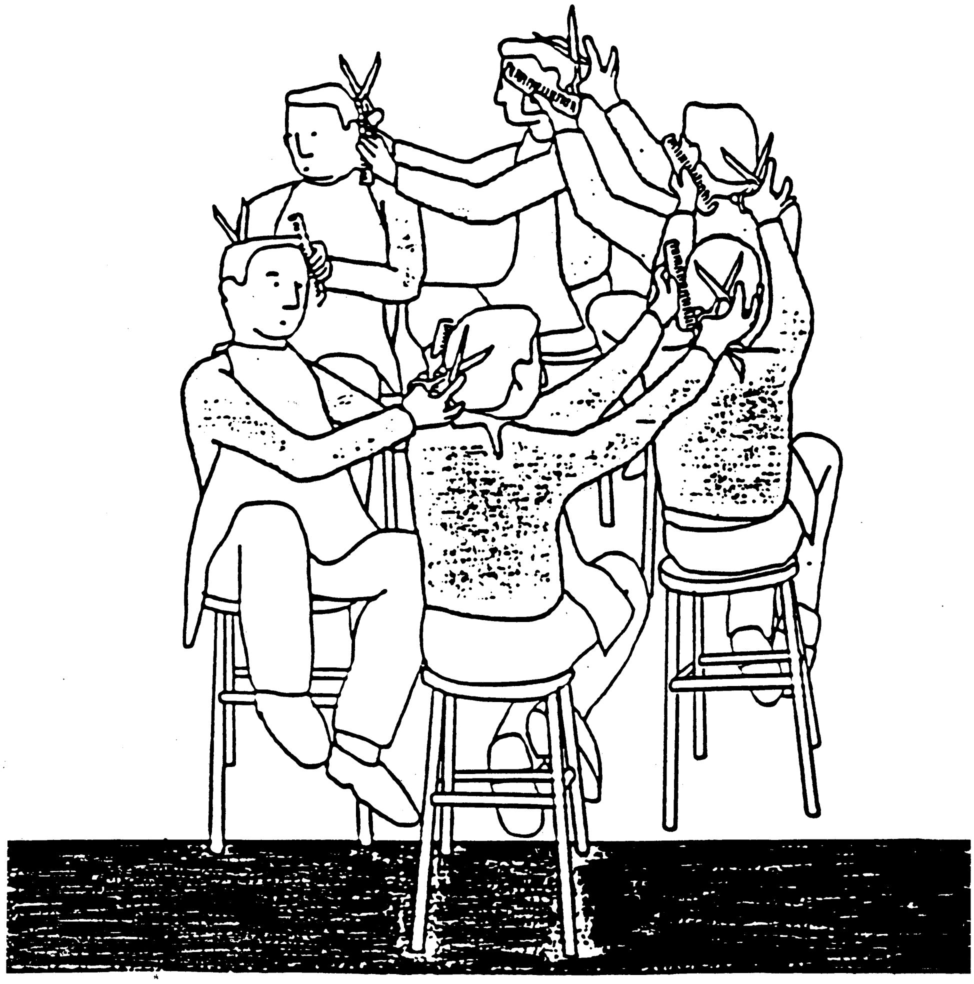
Vol. 65 No. 3, June 1992

MATHEMATICS MAGAZINE



- Reflection Properties of Curves and Surfaces
- Are These Figures Oxymora?
- More on the Four-Numbers Game

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

The aim of *Mathematics Magazine* is to provide lively and appealing mathematical exposition. This is not a research journal and, in general, the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for an article 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. Articles on pedagogy alone, unaccompanied by interesting mathematics, are not suitable. Neither are articles consisting mainly of computer programs unless these are essential to the presentation of some good mathematics. Manuscripts on history are especially welcome, as are those showing relationships between various branches of mathematics and between mathematics and other disciplines.

The full statement of editorial policy appears in this *Magazine*, Vol. 64, pp. 71–72, and is available from the Editor. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, nor published by another journal or publisher.

Send new manuscripts to: Martha Siegel, Editor, Mathematics Magazine, Towson State University, Towson, MD 21204. Manuscripts should be typewritten and double spaced and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should submit the original and two copies and keep one copy. In addition, authors should supply the full five-symbol Mathematics Subject Classification number, as described in *Mathematical Reviews*, 1980 and later. Illustrations should be carefully prepared on separate sheets in black ink, the original without lettering and two copies with lettering added. Do not use staples.

AUTHORS

Daniel Drucker has an S.B. from M.I.T. and received his Ph.D. from the University of California at Berkeley in 1973. His mathematical interests include Lie theory, differential geometry, linear algebra, and number theory, but he also ekes out time to play violin in a local orchestra and piano at home. While explaining the reflection properties of conic sections to his classes at Wayne State University in Detroit, he wondered whether any other curves had such properties. Those musings produced the present article.

Harold L. Dorwart received A.B. from Washington and Jefferson College in 1924. His graduate studies at Yale were interrupted by a teaching position at Williams College. In 1931, he earned his Ph.D. at Yale, under the guidance of Oystein Ore. He then taught at Williams, at Washington and Jefferson, and at Trinity College (Hartford), where he served as chairman of the mathematics department and as dean of the college, retiring in 1968. He is the author of numerous articles on irreducibility criteria, the Tarry-Escott problem, configurations and other geometric topics, and a book, *The Geometry of Incidence*. His interest in configurations has continued since publishing *Configurations* in 1968. Dr. Dorwart has served the MAA on the Board of Governors, as Secretary of the Allegheny Mountain Section and Chairman of the New England Section. He is delighted that a West Coast mathematician—well known for his clear and careful writing—shares this interest to the extent of being co-author of this article.

Branko Grünbaum received his Ph.D. from Hebrew University in Jerusalem in 1958, and has been at the University of Washington since 1966. Most of his research is guided by the conviction that elementary geometry is not only an inexhaustible source of new ideas and insights, but also great fun. The present article is the outcome of collaboration between authors who never met and are separated by a continent and by many years of age, but united in their love for geometry.

Vol. 65 No. 3, June 1992

MATHEMATICS MAGAZINE

The *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 the MATHEMATICS MAGAZINE to an individual member of the Association is $16 included as part of the annual dues. (Annual dues for regular members, exclusive of annual subscription prices for MAA journals, are $64. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 40% dues discount for the first two years of membership.) The nonmember/library subscription price is $68 per year.

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 Ms. Elaine Pedreira, Advertising Manager, The Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

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

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

PRINTED IN THE UNITED STATES OF AMERICA

ARTICLES

Reflection Properties of Curves and Surfaces

DANIEL DRUCKER
Wayne State University
Detroit, MI 48202

Introduction

The reflection properties of parabolas, ellipses, and hyperbolas are well known and have many practical applications. What does it mean in general for a curve or surface to have a reflection property? Which curves and surfaces have them? The purpose of this article is to answer both questions. The results show that the conic sections are quite special, and the proofs include a demonstration of the reflection properties of parabolas, ellipses, and hyperbolas all at once, rather than in the usual case-by-case fashion.

Conic sections

The reflection properties of conic sections are often stated and proved as examples of the application of calculus techniques to analytic geometry. (See Edwards/Penney [**5**, Ch. 10], Stein [**16**, §9.6], and Olmsted [**11**, Ch. 13], for example.) For a parabola, the tangent line at a point P makes the same angle with the line joining P and the focus as it does with the line through P that is parallel to the axis (FIGURE 1).

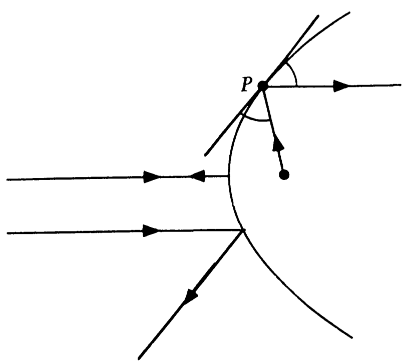


FIGURE 1

Physically, this means that revolving the parabola about its axis produces an excellent shape for a reflecting surface. Light or sound emanating from the focus will be reflected parallel to the axis, a property useful in the design of searchlights, flashlights, headlights, megaphones, etc. Also, light rays from a distant source will be very nearly parallel and can be reflected to the focus for collection, as in a reflecting telescope. We have been assuming that the side of the surface facing the focus is the reflector. If the other side is the reflecting surface, then rays parallel to the axis approaching that side will be reflected *away* from the focus along lines that pass through the focus, creating a simulated point source of light (with a "blind spot" behind the reflector in the direction of the axis). Similarly, rays approaching that side and heading toward the focus will be reflected parallel to the axis. These "back side" versions of the reflection property are mathematically equivalent to the usual versions, but so far as I know, convex parabolic reflectors are never used.

Next consider ellipses. The tangent line at a point P on an ellipse makes equal angles with the lines that join P to the foci (FIGURE 2). Physically, this means that if a reflecting surface is generated by revolving the ellipse about its major axis, then sound or light originating at one focus is reflected through the other. "Whispering galleries," such as the rotunda of the Capitol in Washington, are shaped like the upper half of such a surface. The "back side" version for a convex ellipsoidal reflector says that light headed toward one focus is reflected directly away from the other.

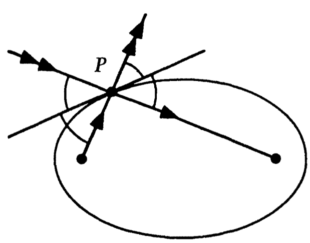


FIGURE 2

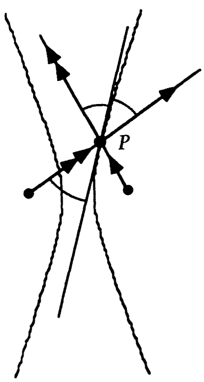


FIGURE 3

The hyperbola, like the ellipse, has the property that the tangent line at a point P makes equal angles with the lines joining P to the foci (FIGURE 3). For a concave reflector obtained by revolving one branch of the hyperbola about its transverse axis (i.e., the line through the foci), this means that light originating at the near focus will be reflected directly away from the other focus. An interesting but not so well-known consequence is that a point source of light at the near focus produces a cone of light whose vertex is the other focus. By contrast, a point source of light at the focus of a parabolic reflector produces a cylindrical beam of light. The "back side" reflection property for a convex hyperbolic reflector says that light directed at the near focus is reflected through the far focus.

For an introduction to the huge variety of applications of conic sections and their reflection properties, see the series of articles by Whitt **[17, 18, 19]**. For further information, consult some of the many books and articles in his bibliographies. Of those, the books by Hilbert/Cohn-Vossen **[10]** and Brueggeman **[2]**, and the articles by Flanders **[7]** and Foster/Pederson **[8]** are especially recommended. Good references not listed by Whitt include Greenstein's article[1] **[9]**, the clear and concise (but

[1]Greenstein considers the sum l of the distances from a variable point on a curve to fixed points p_1 and p_2 on one side of the curve. By viewing l as a function of arc length, he characterizes points of the curve where the lines to the foci make equal angles with the tangent as critical points of l. All the points on the curve have that property exactly when l is constant—that is, when the curve is (part of) an ellipse with foci at p_1 and p_2.

hard to find) book by Bridge [1], and Zwikker's book [20], which is based on the arithmetic of complex numbers rather than on coordinate geometry. These references present a variety of different proofs that parabolas, ellipses, and hyperbolas have reflection properties. (Flanders' proof is remarkable for its simplicity and brevity.[2]) It is worth mentioning some of the many fine books on conic sections, such as those by Coolidge [3], Salmon [12], Smith [13], Sommerville [14], and Spain [15]. These books are not especially concerned with reflection properties, but they contain a wealth of information for the interested reader.

We can unify our view of reflection properties for conic sections by viewing a parabola as an ellipse with one focus at "infinity." To justify this view algebraically, consider the ellipse in FIGURE 4 with foci at $(0,0)$ and $(2c,0)$ and with vertices at $(-d,0)$ and $(2c+d,0)$. Its equation is

$$\frac{(x-c)^2}{(c+d)^2}+\frac{y^2}{(c+d)^2-c^2}=1.$$

If we multiply this equation by c and subtract $c^3/(c+d)^2$ from both sides, we can rewrite the equation as

$$\frac{cx^2-2c^2x}{(c+d)^2}+\frac{cy^2}{d(2c+d)}=\frac{cd(2c+d)}{(c+d)^2}.$$

As $c\to+\infty$ this becomes $-2x+y^2/2d=2d$ or

$$y^2=4d(x+d).$$

This represents a parabola with vertex $(-d,0)$ and focus $(0,0)$.

Now the reflection property for all three types of conic sections is characterized by saying that a ray from either focus is reflected along a line that passes through the other focus. Mathematically, the tangent line at any point of the curve makes equal angles with the lines joining the point to the two foci. (In the case of the parabola, the line joining the point to the "focus at infinity" is parallel to the axis of symmetry.)

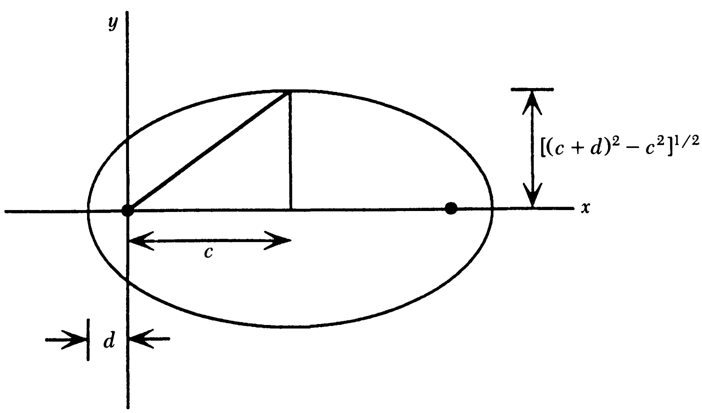


FIGURE 4

[2]There is a hidden assumption in Flanders' proof of the reflection property for a parabola. For the equation $\mathbf{x}\cdot\mathbf{u}=|\mathbf{x}-\mathbf{a}|$ to hold, the origin must be on the directrix.

Candidates for curves with reflection properties

How special are the conic sections? Do other curves have reflection properties? To find out, we must first decide what we mean by a "reflection property." The definition we use will be worded so as to accommodate all the conic sections—even circles, whose "foci" F and F' coincide.

Let us agree that a smooth connected plane curve **C** has a *reflection property* if there are points F and F', not on **C**, such that the tangent line at any point P of **C** bisects one of the pairs of opposite angles formed by the intersection of the lines joining P to F and F'. A "point at infinity" is specified by means of a line through the origin. The line joining P to a point at infinity is defined to be the line through P parallel to the given line. Points at infinity are not considered to be on **C**.

If **C** has a reflection property given by points F and F' at infinity, then for each point P on **C**, the tangent line at P makes equal angles with the lines from P to F and F'. Evidently the tangent line either bisects the acute angle between the lines joining P to F and F', or else it is perpendicular to the angle bisector. (When $F = F'$, the tangent line at P is either parallel or perpendicular to the line through P and F.) The directions of the lines from P to F and F' do not depend on the choice of P. Thus the smoothness of **C** ensures that all the tangent lines to **C** are parallel. It follows that **C** is (part of) a straight line. **C** does not determine F and F' uniquely.

Next suppose that **C** has a reflection property given by F and F', where F, F' are not points at infinity. We are only concerned with the shape of **C**, not with its position or orientation, so we can assume without loss of generality that F is the origin and that F' has coordinates $(s, 0)$. Suppose that points P of **C** are described by polar coordinates (r, θ), where r is a positive-valued function of θ. Consider such a point P, which we can take to be in the upper half plane. Let α be the angle of inclination of the tangent line l to **C** at P, and let ψ be the angle of inclination for the line through P and F'. Of course θ is the angle of inclination for the line through P and F (see FIGURE 5). As P moves along **C**, the coordinates r and θ change but s remains constant. We will now exploit that fact by solving for s in terms of r and θ, and investigating the condition that $ds/d\theta = 0$.

By applying the law of sines to triangle FPF', we obtain

$$\frac{s}{\sin(\psi - \theta)} = \frac{r}{\sin(\pi - \psi)} = \frac{r}{\sin \psi},$$

so

$$s = \frac{r(\sin \psi \cos \theta - \cos \psi \sin \theta)}{\sin \psi} = r\left(\cos \theta - \frac{\sin \theta}{\tan \psi}\right). \tag{1}$$

(If $\psi = \pi/2$, then $s = r \cos \theta$.)

The angle of inclination of the normal to **C** is $\alpha - \pi/2$ or $\alpha + \pi/2$, whichever is in $[0, \pi)$. If l bisects angle FPF', then $\alpha = \frac{1}{2}(\theta + \psi)$. Otherwise $\alpha \pm \pi/2 = \frac{1}{2}(\theta + \psi)$. (In FIGURE 5, $\alpha + \pi/2 = \frac{1}{2}(\theta + \psi)$.) Thus

$$\psi = 2\alpha - \theta + \sigma\pi, \quad \text{where } \sigma \in \{-1, 0, 1\},$$

and

$$\tan \psi = \tan(2\alpha - \theta) = \frac{\tan 2\alpha - \tan \theta}{1 + \tan 2\alpha \tan \theta}. \tag{2}$$

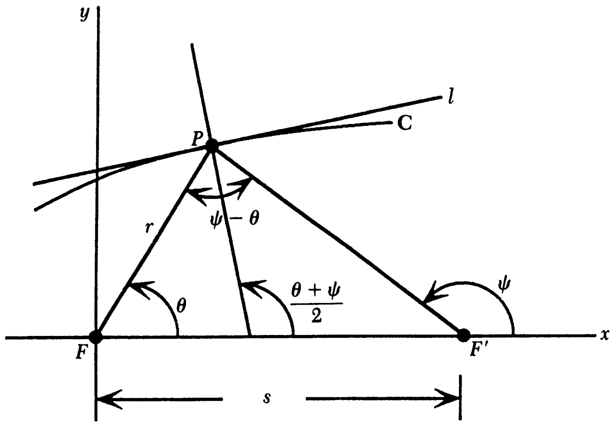


FIGURE 5

Here

$$\tan 2\alpha = \frac{2\tan\alpha}{1-\tan^2\alpha} \tag{3}$$

and

$$\tan\alpha = \frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta} = \frac{(r\sin\theta)'}{(r\cos\theta)'} = \frac{r'\sin\theta + r\cos\theta}{r'\cos\theta - r\sin\theta}, \tag{4}$$

where the primes denote differentiation with respect to θ. Substitution of (4) in (3) yields the formula

$$\tan 2\alpha = \frac{\left[(r')^2 - r^2\right]\sin 2\theta + 2rr'\cos 2\theta}{\left[(r')^2 - r^2\right]\cos 2\theta - 2rr'\sin 2\theta} \tag{5}$$

and substitution of (5) in (2) shows that

$$\tan\psi = \frac{\left[(r')^2 - r^2\right]\sin\theta + 2rr'\cos\theta}{\left[(r')^2 - r^2\right]\cos\theta - 2rr'\sin\theta}. \tag{6}$$

Finally, substitution of (6) in (1) gives the desired formula for s:

$$s = \frac{2r^2r'}{\left[(r')^2 - r^2\right]\sin\theta + 2rr'\cos\theta}. \tag{7}$$

Setting $s' = 0$ (i.e., using the quotient rule and setting the numerator of the resulting expression for s' equal to 0), we obtain

$$0 = 2r\left[(r')^2 + r^2\right]\left\{-\left[rr'' - 2(r')^2\right]\sin\theta + rr'\cos\theta\right\}.$$

Since r and $(r')^2 + r^2$ are positive, this yields the differential equation

$$\left[rr'' - 2(r')^2\right]\sin\theta = rr'\cos\theta. \tag{8}$$

This equation can be simplified by setting $\rho = 1/r$, so that

$$r = 1/\rho, \qquad r' = -\rho'/\rho^2, \quad \text{and} \quad r'' = \left[-\rho\rho'' + 2(\rho')^2\right]/\rho^3.$$

Then (8) becomes

$$\rho'' \sin\theta = \rho' \cos\theta. \tag{9}$$

Solving (9) for ρ' by separation of variables, we get

$$\rho' = b \sin\theta \quad \text{and} \quad \rho = a - b\cos\theta$$

for some constants a and b, not both zero (since $\rho \neq 0$). Thus

$$r = \frac{1}{a - b\cos\theta} \quad (a, b \text{ not both zero}). \tag{10}$$

Using (7) to calculate s, we get

$$s = 2b/(a^2 - b^2). \tag{11}$$

Since $s < \infty$, we must have $|a| \neq |b|$.

If a and b are positive, we can set $e = b/a$ and $p = 1/b$, so that (10) becomes

$$r = \frac{pe}{1 - e\cos\theta}, \tag{12}$$

the polar equation of a conic section with focus at the origin, directrix $x = -p$, and eccentricity $e > 0$. If a and b are nonzero, but not both positive, then (10) still represents a conic section with focus at the origin, directrix $x = -1/b$, and eccentricity $e = |b/a|$. Since $|a| \neq |b|$, the conic section is an ellipse or hyperbola, not a parabola. (The parabolic case $e = 1$ is the case in which the denominator in (7) vanishes.)

If $a \neq 0$ and $b = 0$, then $s = 0$ and $r = 1/a$, so $F = F'$ and **C** is a circle. If $a = 0$ and $b \neq 0$, then $s = -2/b$ and $r = -1/b\cos\theta$, so **C** is the vertical line $x = -1/b$, that is, the perpendicular bisector of the line segment FF'.

There are also two degenerate cases. If **C** is part of the x-axis, then we cannot take our points P to lie in the upper half plane, so our derivation does not apply. **C** will have a reflection property relative to any points F, F' on the x-axis that do not lie on **C**. Also, our derivation only applied to points P where $\psi \neq \pi/2$. If **C** is part of the vertical line through F', then $\psi = \pi/2$ for *all* points of **C** and we cannot solve for s. The points F and F' must coincide for the reflection property to hold. We see that in both cases, **C** is part of a line and F and F' are points of the same line, but they are not on **C**.

We still have to consider the case in which just one of F, F' is a point at infinity. Let F' be the point at infinity defined by the x-axis. Then the reflection property of **C** implies that $2\alpha = \theta$ or $2\alpha = \theta + \pi$ (in FIGURE 6, $2\alpha = \theta$), so

$$\tan 2\alpha = \tan\theta. \tag{13}$$

Formula (5) still holds. When we substitute (5) in (13), we get

$$\left[(r')^2 - r^2\right]\sin\theta + 2rr'\cos\theta = 0. \tag{14}$$

As we remarked earlier, this differential equation is satisfied only by parabolas in polar form:

$$r = \frac{p}{1 \pm \cos\theta}. \tag{15}$$

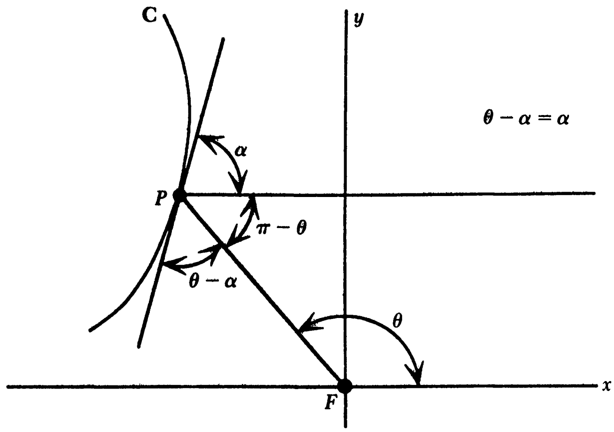


FIGURE 6

To see this directly, set $\xi = r(1-\cos\theta)$, so that

$$r = \xi/(1-\cos\theta) \text{ and } r' = [\xi'(1-\cos\theta) - \xi\sin\theta]/(1-\cos\theta)^2.$$

Then (14) becomes (after some manipulations)

$$\xi'(\xi'\sin\theta - 2\xi)/(1-\cos\theta)^2 = 0. \tag{16}$$

It follows that $\xi' = 0$ or $\xi'/\xi = 2\csc\theta$, so

$$\xi = p \quad \text{or} \quad \xi = \frac{p(1-\cos\theta)}{1+\cos\theta}$$

for some constant p. This gives (15). Again straight lines occur as a degenerate case (when **C** is part of the x-axis).

We now have the full slate of candidates for reflection properties, namely straight lines, circles, ellipses, hyperbolas, parabolas, and pieces of those curves.

Checking the candidates' credentials

It is obvious that straight lines and circles have a reflection property. This is not obvious for ellipses, hyperbolas, and parabolas. By going through our calculations backwards, we can replace the usual case-by-case treatment with a single proof.

Using translation and rotation if necessary, we can assume that our conic section **C** has a polar equation of the form

$$r = \frac{pe}{1-e\cos\theta},$$

where $p > 0$ and $e > 0$. (See [6, §11.5], for example.) Let F be the origin. When $e \neq 1$, let F' be the point with coordinates $(s, 0)$, where $s = 2pe^2/(1-e^2)$. (This expression for s comes from formula (11), with $b = 1/p$ and $a = b/e = 1/pe$.) When $e = 1$, let F' be the point at infinity determined by the x-axis.

Initially, let's suppose $e \neq 1$, so that the formula for s makes sense. If points P on **C** have polar coordinates (r, θ), where r is a function of θ, then we see that

(i) the slope of the line through P and F is $\tan\theta$,
(ii) the slope of the tangent line to $\mathbf{C}$ at P is given by (4):

$$\tan\alpha = \frac{r'\sin\theta + r\cos\theta}{r'\cos\theta - r\sin\theta},$$

(iii) the slope of the line through P and F' is given by

$$\tan\psi = \frac{r\sin\theta}{r\cos\theta - s}. \tag{17}$$

Substituting $r = pe/(1 - e\cos\theta)$ and $s = 2pe^2/(1 - e^2)$ in (17), we get

$$\tan\psi = \frac{(1-e^2)\sin\theta}{(1+e^2)\cos\theta - 2e}. \tag{18}$$

This formula is correct even when $e = 1$, so we now drop the assumption that $e \neq 1$. By the computations in equations (2) through (6), we have

$$\tan(2\alpha - \theta) = \frac{\left[(r')^2 - r^2\right]\sin\theta + 2rr'\cos\theta}{\left[(r')^2 - r^2\right]\cos\theta - 2rr'\sin\theta}. \tag{19}$$

Substituting for r in (19) and simplifying gives

$$\tan(2\alpha - \theta) = \frac{(1-e^2)\sin\theta}{(1+e^2)\cos\theta - 2e}. \tag{20}$$

Comparison of (18) and (20) shows that $2\alpha - \theta = \psi + \sigma\pi$, where $\sigma \in \{-1, 0, 1\}$. Thus $\frac{1}{2}(\theta + \psi) = \alpha + \sigma\pi/2$, showing that the tangent or normal line to $\mathbf{C}$ at P bisects the angle between the lines joining P to F and F'. This proves that $\mathbf{C}$ has a reflection property relative to F and F'.

We have shown that the conic sections are *very* special:

THEOREM. *A smooth connected plane curve has a reflection property if and only if it is a connected subset of a circle, ellipse, hyperbola, parabola, or straight line.*

Algebraically, the plane curves with a reflection property are those that (after translation and rotation if necessary) can be represented by a polar equation of the form $r = 1/(a - b\cos\theta)$, where a and b are nonnegative, but not both zero. (Replacing r and θ by $-r$ and $\theta + \pi$ changes the sign of a without changing the curve; replacing θ by $\pi - \theta$ changes the sign of b by reflecting the curve about the y-axis. These maneuvers enable us to ensure that a and b are nonnegative.)

Surfaces with reflection properties

Earlier in this article, reflection properties of conic sections were described in terms of surfaces of revolution they generate. It seems appropriate to inquire whether other surfaces might also have reflection properties. With that purpose in mind, we now propose a definition of "reflection property" for surfaces: Let $\mathbf{S}$ be a smooth connected surface and let F and F' be points not in $\mathbf{S}$. We say that $\mathbf{S}$ has a *reflection property* relative to F and F' if, for each point P in $\mathbf{S}$,

(i) the unit surface normal $\vec{N}$ at P lies in the subspace spanned by the vectors $\overrightarrow{FP}$ and $\overrightarrow{F'P}$, and

(ii) the line through $\vec{N}$ at P bisects one of the pairs of opposite angles formed by the intersection of the lines joining P to F and F'.

When $F \neq F'$, the first condition says that $\vec{N}$, $\overrightarrow{FP}$, and $\overrightarrow{F'P}$ are coplanar. When $F = F'$, it requires that $\vec{N}$ and $\overrightarrow{FP}$ be parallel.[3]

If F and F' are points at infinity, then an argument analogous to the one for curves shows that the surface normals at points of **S** are all parallel, so **S** is (part of) a plane.

If neither F nor F' is a point at infinity, then we can assume without loss of generality that F is the origin and that F' has coordinates $(s, 0, 0)$. Using the somewhat nonstandard spherical coordinate system (with a vertical x-axis) pictured in FIGURE 7, we can describe points P of **S** by spherical coordinates r, θ, and ϕ $(0 \leqslant \theta \leqslant \pi)$, where r is a positive-valued function of θ and ϕ. $\vec{N}$ can be computed in terms of the position vector $\vec{R} = \overrightarrow{FP}$ by means of the formula

$$\vec{N} = \vec{e}_\theta \times \vec{e}_\phi, \tag{21}$$

where $\vec{e}_\theta = (\partial\vec{R}/\partial\theta)/|\partial\vec{R}/\partial\theta|$ and $\vec{e}_\phi = (\partial\vec{R}/\partial\phi)/|\partial\vec{R}/\partial\phi|$. Using the fact that

$$\vec{R} = (x, y, z) = (r\cos\theta, r\sin\theta\cos\phi, r\sin\theta\sin\phi) \tag{22}$$

and writing r_θ, r_ϕ instead of $\partial r/\partial\theta, \partial r/\partial\phi$, we find that

$$\sqrt{r^2 + r_\theta^2}\sqrt{r^2\sin^2\theta + r_\phi^2}\,\vec{N} = (X, Y, Z), \tag{23}$$

where

$$X = r\sin\theta(r_\theta\sin\theta + r\cos\theta),$$
$$Y = r\left[r_\phi\sin\phi + \sin\theta\cos\phi(r\sin\theta - r_\theta\cos\theta)\right], \text{ and}$$
$$Z = r\left[-r_\phi\cos\phi + \sin\theta\sin\phi(r\sin\theta - r_\theta\cos\theta)\right].$$

Suppose $s \neq 0$. Then condition (i) says that

$$\overrightarrow{FP} \times \overrightarrow{F'P} \cdot \vec{N} = 0, \tag{24}$$

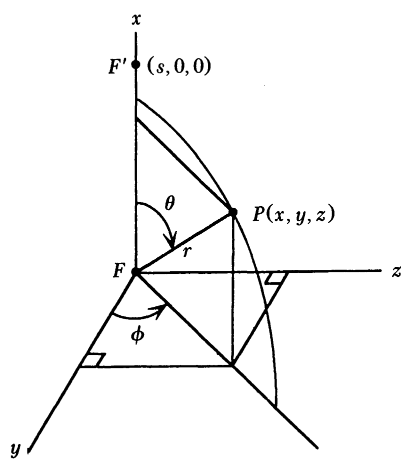


FIGURE 7

[3]The analogous condition for curves, requiring that the normal vector at each point P lie in the span of $\overrightarrow{FP}$ and $\overrightarrow{F'P}$, would be vacuous most of the time but would eliminate the two degenerate cases.

or equivalently,

$$0=\begin{vmatrix} r\cos\theta & r\sin\theta\cos\phi & r\sin\theta\sin\phi \\ r\cos\theta - s & r\sin\theta\cos\phi & r\sin\theta\sin\phi \\ X & Y & Z \end{vmatrix}$$

$$= sr\sin\theta(Z\cos\phi - Y\sin\phi) = sr^2 r_\phi \sin\theta.$$

For P not on the x-axis, $r \neq 0$ and $\sin\theta \neq 0$. Since $s \neq 0$, we conclude that $r_\phi = 0$, so r is constant as a function of ϕ for points of **S** not on the x-axis. In other words, **S** is (part of) a surface of revolution. By condition (ii), the cross section of that surface of revolution sliced out by the xy-plane is a plane curve **C** with reflection property determined by F and F'. The points F and F' are distinct since $s \neq 0$. Thus **C** is (part of) an ellipse or hyperbola with foci at F and F', or else (part of) the perpendicular bisector of the line segment FF'. It follows that **S** is (part of) a plane, or (part of) an ellipsoid or hyperboloid of revolution whose axis of rotation contains the foci of the generating ellipse or hyperbola.

Next suppose $s = 0$, so that $F = F'$. Condition (i) says that $\overrightarrow{FP}$ and $\vec{N}$ are parallel at all points of **S**. [This prevents **S** from being part of a cone with F and F' at its vertex, a surface for which $\overrightarrow{FP}$ and $\vec{N}$ are always perpendicular. Since r is not a function of θ and ϕ on a cone, this example is easy to overlook.] Thus $\overrightarrow{FP} \times \vec{N} = \vec{0}$ for all P. Using the expressions for $\overrightarrow{FP}$ and $\vec{N}$ given in (22) and (23), and setting the components of $\overrightarrow{FP} \times \vec{N}$ equal to zero, we obtain

$$-r^2 r_\phi \sin\theta = 0,$$

$$r^2 r_\phi \cos\theta\cos\phi + r^2 r_\theta \sin\theta\sin\phi = 0, \quad \text{and}$$

$$r^2 r_\phi \cos\theta\sin\phi - r^2 r_\theta \sin\theta\cos\phi = 0.$$

For points P not on the x-axis, $r \neq 0$ and $\sin\theta \neq 0$, so we conclude from the first equation that $r_\phi = 0$. It then follows from the other two equations that $r_\theta = 0$ as well. This means that r is constant as a function of ϕ and θ, so **S** is (part of) a sphere.

Finally suppose that only F' is a point at infinity. We can assume without loss of generality that F' is determined by the x-axis. Thus we can replace $\overrightarrow{F'P}$ in condition (24) by the vector $(1, 0, 0)$ and obtain

$$0=\begin{vmatrix} r\cos\theta & r\sin\theta\cos\phi & r\sin\theta\sin\phi \\ 1 & 0 & 0 \\ X & Y & Z \end{vmatrix}$$

$$= -r\sin\theta(Z\cos\phi - Y\sin\phi) = r^2 r_\phi \sin\theta.$$

As before, we conclude that r is constant as a function of ϕ for points of **S** not on the x-axis, so **S** is (part of) a surface of revolution. This time condition (ii) says that the cross section **C** of **S** in the xy-plane is (part of) a parabola, so **S** is (part of) a paraboloid.

We have shown that the only candidates for surfaces with reflection properties are planes, spheres, paraboloids of revolution, and the ellipsoids and hyperboloids of revolution obtained by revolving ellipses and hyperbolas about the line through their foci. The verification that all these surfaces actually have reflection properties is routine. This proves:

THEOREM. *A smooth connected surface has a reflection property if and only if it is a connected subset of one of the following: a plane, a sphere, a paraboloid of revolution, or an ellipsoid or hyperboloid obtained by revolving an ellipse or hyperbola about the line through its foci.*

A version of this result for hypersurfaces in n-dimensional Euclidean space can be found in [4].

Conclusion

If we exclude the somewhat degenerate case of a straight line, which does not uniquely determine its "foci" F and F', then the smooth connected curves with reflection properties are precisely the conic sections: circles, ellipses, hyperbolas, and parabolas (or pieces of those curves). This adds another important characterization of the conic sections to a list that includes their two best-known characterizations as the nondegenerate plane sections of a right circular cone and as the nondegenerate plane curves with equations of degree two. Moreover, we have shown that the conic sections generate, by rotation about a line through their foci, the only smooth connected curved surfaces with reflection properties.

I would like to thank Bernd Wegner, who pointed out the need for a subtle change in my original definition (see [4]) of reflection property for surfaces.

REFERENCES

1. B. Bridge, *Treatise on the Construction, Properties, and Analogies of the Three Conic Sections* (from the 2nd London edition), jointly published by Durrie and Peck, New Haven and Collins, Keese, and Co., New York, 1839.
2. H. Brueggeman, *Conic Mirrors*, Focal Press, London, 1968.
3. J. Coolidge, *A History of the Conic Sections and Quadric Surfaces*, Oxford Univ. Press, London, 1947.
4. D. Drucker, Euclidean hypersurfaces with reflection properties, *Geometriae Dedicata* 33 (1990), 325–329.
5. C. H. Edwards, Jr. and D. E. Penney, *Calculus and Analytic Geometry*, 2nd edition, Prentice-Hall, Englewood Cliffs, NJ, 1986.
6. R. Ellis and D. Gulick, *Calculus with Analytic Geometry*, 4th edition, Harcourt Brace Jovanovich, San Diego, CA, 1990.
7. H. Flanders, The optical property of the conics, *Amer. Math. Monthly* 75 (1968), 399.
8. J. H. Foster and J. J. Pedersen, On the reflective property of ellipses, *Amer. Math. Monthly* 87 (1980), 294–297.
9. D. S. Greenstein, On extremal optical paths and the law of reflection, *Amer. Math. Monthly* 66 (1959), 798–800.
10. D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelsea Publishing Co., New York, 1952, translated by P. Nemenyi.
11. J. M. H. Olmsted, *Calculus With Analytic Geometry*, Vol. 1, Appleton-Century-Crofts, New York, 1966.
12. G. Salmon, *A Treatise on Conic Sections*, Longmans, Green and Co., London, 1917. Reprinted by Chelsea Publishing Co., New York, 1948.
13. C. Smith, *An Elementary Treatise on Conic Sections By the Methods of Co-ordinate Geometry*, Macmillan, London, 1918.
14. D. M. Y. Sommerville, *Analytical Conics*, G. Bell and Sons, London, 1924.
15. B. Spain, *Analytical Conics*, Pergamon Press, New York, 1957.
16. S. K. Stein, *Calculus and Analytic Geometry*, 4th edition, McGraw-Hill, New York, 1987.
17. L. Whitt, The standup conic presents: the parabola and applications, *UMAP Journal* 3 (1982), 285–313.
18. ______, The standup conic presents: the ellipse and applications, *UMAP Journal* 4 (1983), 157–183.
19. ______, The standup conic presents: the hyperbola and applications, *UMAP Journal* 5 (1984), 9–21.
20. C. Zwikker, *Advanced Plane Geometry*, North-Holland, Amsterdam, 1950.

Are These Figures Oxymora?

HAROLD L. DORWART
Salisbury, CT 06068

BRANKO GRÜNBAUM*
University of Washington
Seattle, WA 98195

Oxymoron [Greek *oxymoron*, a smart saying which at first view appears foolish, from *oxys*, sharp, and *morŏs*, dull, foolish] a figure of speech in which opposite or contradictory ideas or terms are combined.

(After Noah Webster)

The century-old topic of point and line configurations straddles the fence between projective geometry and combinatorics. In this article we shall be concerned mainly with the geometric approach and will attempt to highlight certain aspects of such configurations that we find very interesting. We hope the reader will agree. These aspects seem not to be widely known, possibly because of the confusing nature of the usual terminology, and—even more—due to the general decline in familiarity with geometric facts. Also, although there is a great amount of known material, it is scattered in many papers, a large fraction of which appeared in rather inaccessible journals. Unfortunately, there is no book that would present a reasonable account of such material. It is remarkable that in an elementary topic such as configurations, there are still many unsolved questions, and that fruitful connections to other branches of mathematics and its applications are fueling a renewed interest. The reason for the following pages is the hope that they may help awaken in our students (and in other readers) an interest in geometry. The paper may also afford them a chance to "try their wings" in independently developing a nontrivial but easily accessible topic, and to experience the fact that "elementary" questions may be hard enough to have resisted solution even to this day. Although the material of this note traditionally would appear in the context of projective geometry, the reader may consider that all the points and lines are in the ordinary Euclidean plane. Many of the references are given for the sake of historical interest and understanding of the development, and we do not expect the reader to spend much time looking for them.

Cyclically inscribed multilaterals

By way of introduction to our topic, consider FIGURE 1. It is evident that A is cutting B's hair, B is cutting C's hair,..., and F is cutting A's hair. (Such groupings of individuals probably do not exist at present, but may soon be formed if the price of haircuts continues to increase!) Now consider FIGURE 2, representing a 9_3 *configuration*, that is, a family of nine points and nine lines, with three of the points on each of

*Research supported in part by NSF grants DMS-8620181 and DMS-9008813.

the lines, and three of the lines passing through each of the points. In this configuration, three *trilaterals* have been indicated by differing texture of the lines. Here we must say a few words regarding terminology. A "*multilateral*" or "n-lateral" is a cyclic sequence of n distinct points called, say, $A_1, A_2, \ldots, A_n$, and n lines $A_1A_2, A_2A_3, \ldots, A_{n-1}A_n, A_nA_1$ determined by the indicated pairs of points. "Trilaterals" and "quadrilaterals" are n-laterals with $n = 3$ or $n = 4$, respectively. In the sequel we shall simplify the notation by omitting the letter A, and labelling the points by natural numbers. To avoid confusion with other points of the plane, we shall call the points determining a multilateral its *vertices*, and we shall use this term also for the "points" of the various configurations. The difference between trilaterals and the more familiar triangles is that triangles are formed by segments determined by pairs of points, while trilaterals are formed by complete, unbounded lines; similarly for the difference between multilaterals and polygons—we shall have occasion to consider polygons later in this article. The early writers on configurations used the word "polygons" both when they meant "multilaterals" and when they meant "polygons."

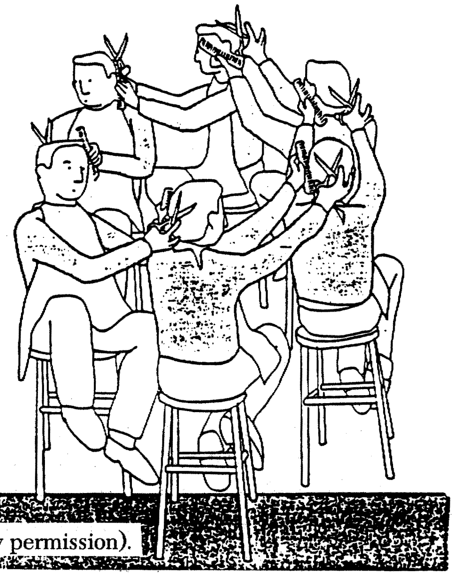


FIGURE 1

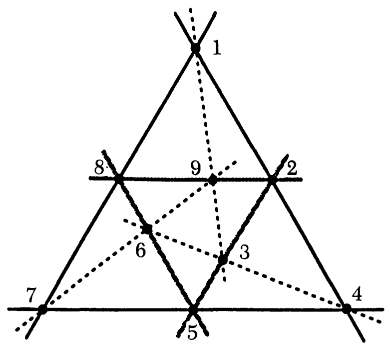


FIGURE 2

A 9_3 configuration, in which a cycle of three trilaterals, each inscribed in the preceding, has been indicated by different texture of the lines. Similar visual aids are used in the following diagrams.

Although this seems not to have bothered them too much—see Hilbert [11], translator's footnote on p. 110 of the English version—we find it preferable and more logical to use different words for different concepts that often occur in the same context. We shall say that, in FIGURE 2, the trilateral determined by the vertices labeled 2, 5, 8 is *inscribed* in the trilateral 1, 4, 7, in analogy to the usual interpretation of the word. By "inscribed" we mean that the vertices of the former lie on the lines of the latter, one vertex on each line. It is obvious that the trilateral 1, 4, 7 is inscribed in the trilateral 3, 6, 9, which, in turn, is inscribed in the trilateral 2, 5, 8. Hence the situation is completely analogous to the one shown in FIGURE 1.

Although trilaterals may appear strange at first, the use of that concept leads to some neat mathematics. Some of it goes back to the beginnings of study of configurations (see, for example, [15]). To illustrate this, fix your attention on the cycle of trilaterals of FIGURE 2 when written in the following form, in which each row corresponds to a trilateral:

$$\begin{array}{ccc} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9. \end{array}$$

This square array may be expanded to a similarly constructed rectangular array

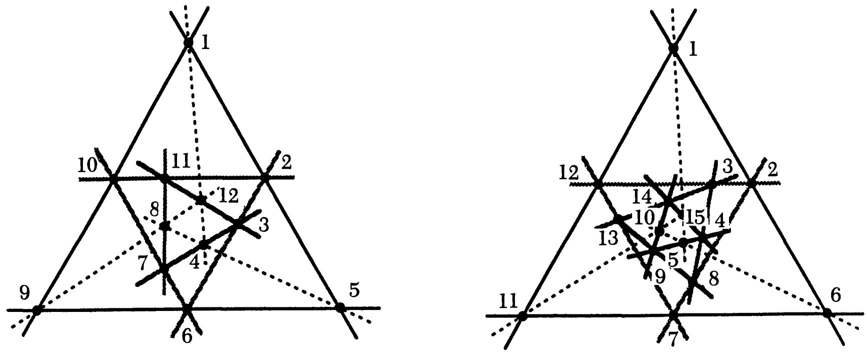
$$\begin{array}{ccc} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12. \end{array}$$

The new array can be taken as a representation of FIGURE 3(a), which is a 12_3 configuration with a cycle of four trilaterals, each inscribed in the preceding one. Further expansion leads to the cycle of five trilaterals in FIGURE 3(b), and so on. For clarity, in these diagrams and in the following ones, only those parts of the lines involved have been shown that are necessary to show how one multilateral is inscribed in another; we used shadings of various kinds to distinguish the multilaterals involved.

When the original array is expanded in the right-hand direction we have

$$\begin{array}{cccc} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12, \end{array}$$

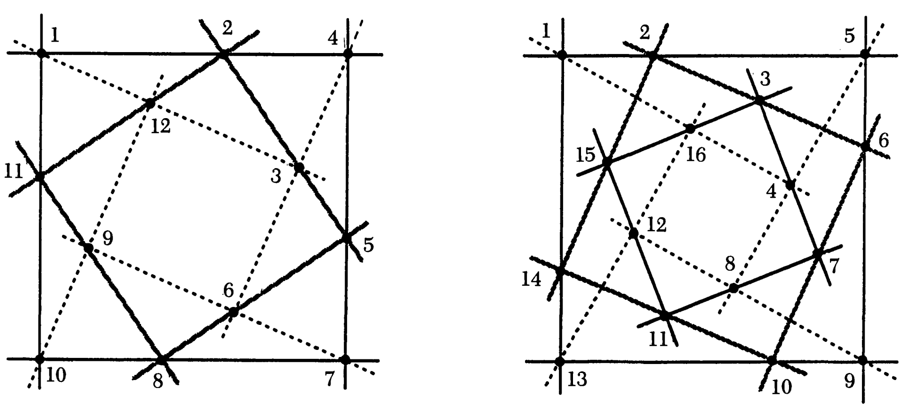
which represents FIGURE 4(a), a 12_3 configuration with a cycle of three inscribed quadrilaterals, then to FIGURES 4(b), 5, and so on. However, starting with $k = 5$, a somewhat surprising new possibility arises, namely a sequence of just two mutually inscribed "star-shaped" k-laterals, leading to configurations n_3 with $n = 2k \geq 10$; the cases $k = 5$ and 8 are illustrated in FIGURE 6. The existence of such configurations has been mentioned only relatively recently; the earliest publications we are aware of in which examples of such configurations are shown are van de Craats [22] and Zacharias [23].



(a) (b)

FIGURE 3

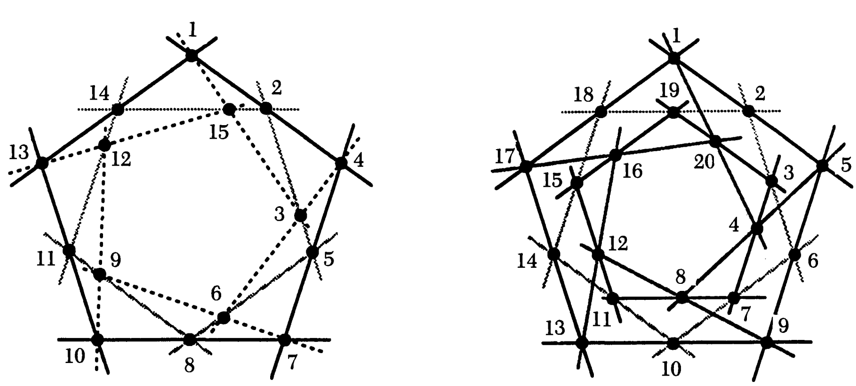
A 12_3 and a 15_3 configuration, with cycles of four or five inscribed trilaterals.



(a) (b)

FIGURE 4

A 12_3 and a 16_3 configuration, with cycles of three of four inscribed quadrilaterals.



(a) (b)

FIGURE 5

A 15_3 and a 20_3 configuration, with cycles of three or four inscribed 5-laterals.

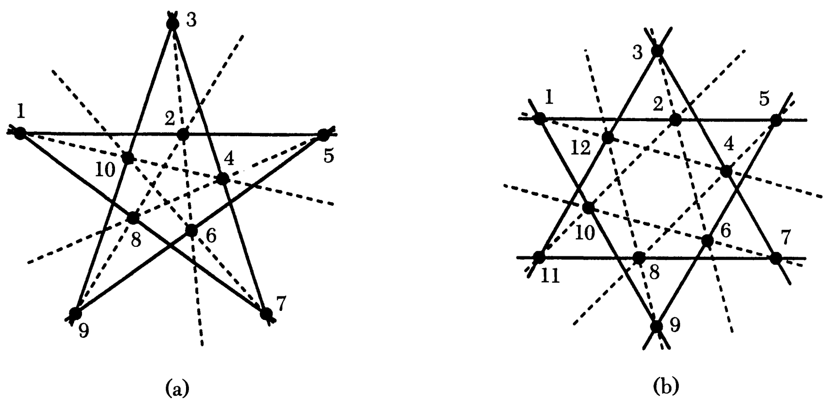


FIGURE 6
A 10_3 and a 16_3 configuration, with two mutually inscribed "star-shaped" 5-laterals or 8-laterals.

It may be noted that in constructing the configurations of these figures, the early multilaterals can be inscribed as desired, but for the final one the possibilities are severely limited by the requirement that its lines must pass through the vertices of the first multilateral. This shows that there is a nontrivial question hiding behind our drawings: Are the lines shown as straight really straight, and do they in fact pass through the vertices in the way the diagrams lead us to believe? You may enjoy experimenting a little on your own, by drawing examples of configurations in which the multilaterals are based on regular polygons (as in our diagrams) to see at what stage the freedom of choice stops. Once you reach such an experimental insight, you may try to prove its validity in general (most easily by using an argument based on the continuity of the way lines depend on the points determining them, and vice versa).

Configuration tables

As a bonus for the work that has been done in constructing the sequences of configurations $9_3, 12_3, 15_3, \ldots; 12_3, 16_3, 20_3;\ 15_3, 20_3, 25_3, \ldots,$ and so on, a second aspect of line configurations appears. By taking pairs of vertices, determined by following the natural order of the integers used to label the vertices, and considering the segments they define, we obtain a closed circuit of segments in each figure. They define a *polygon* in the configuration. Since it contains all the vertices, in analogy with graph theory we shall call it a *Hamiltonian polygon* of the configuration. For the 9_3 configuration of FIGURE 2, the Hamiltonian polygon is formed by the segments $\overline{12}$, $\overline{23}$, $\overline{34}$, $\overline{45}$, $\overline{56}$, $\overline{67}$, $\overline{78}$, $\overline{89}$, $\overline{91}$, as shown in FIGURE 7(a). Additional examples are shown in FIGURES 7(b, c), 8, and 9. Each such Hamiltonian polygon determines a *Hamiltonian multilateral* (consisting of the lines that contain the sides of the polygon); this multilateral is *self-inscribed* in the sense of our definition. (In some publications —especially older ones—the expression "self-inscribed and circumscribed polygon" is used. Besides being unwieldy, it is also unreasonable if the word "polygon" is taken in its usual meaning.) The senior of the authors of this article, thinking that a former engineering colleague might be interested in the resemblance of FIGURES 7, 8, 9 to gearwheels, showed them to him and was surprised by his comment that his immediate reaction was to wonder if the series of drawings had any connection with the fashionable areas of fractals, chaos, etc.!

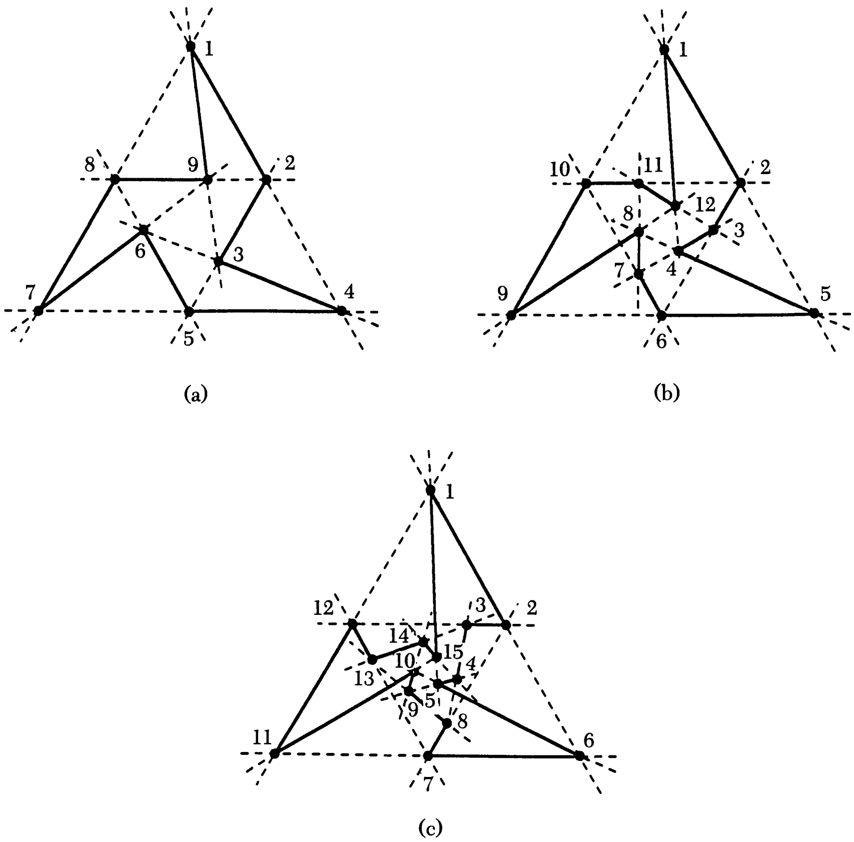


FIGURE 7
Hamiltonian polygons in the configurations of FIGURES 2 and 3.

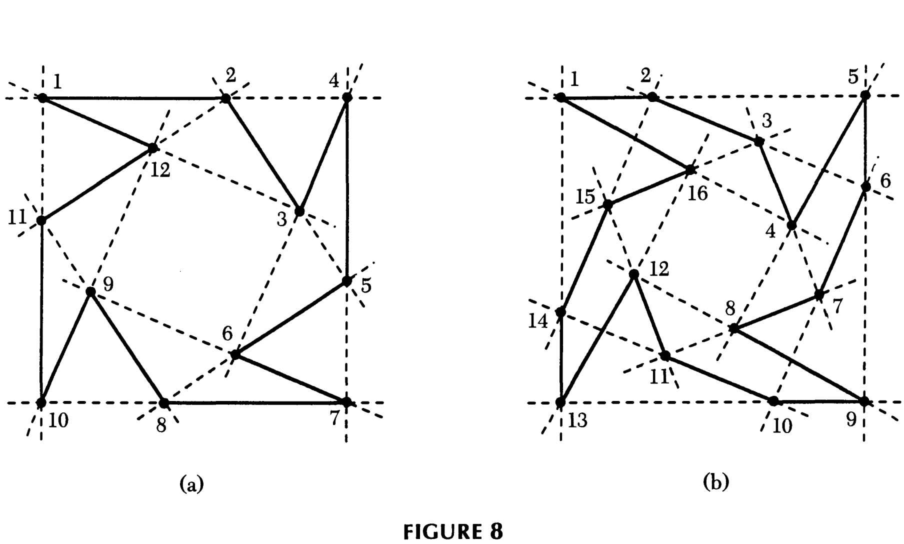


FIGURE 8
Hamiltonian polygons in the configurations of FIGURE 4.

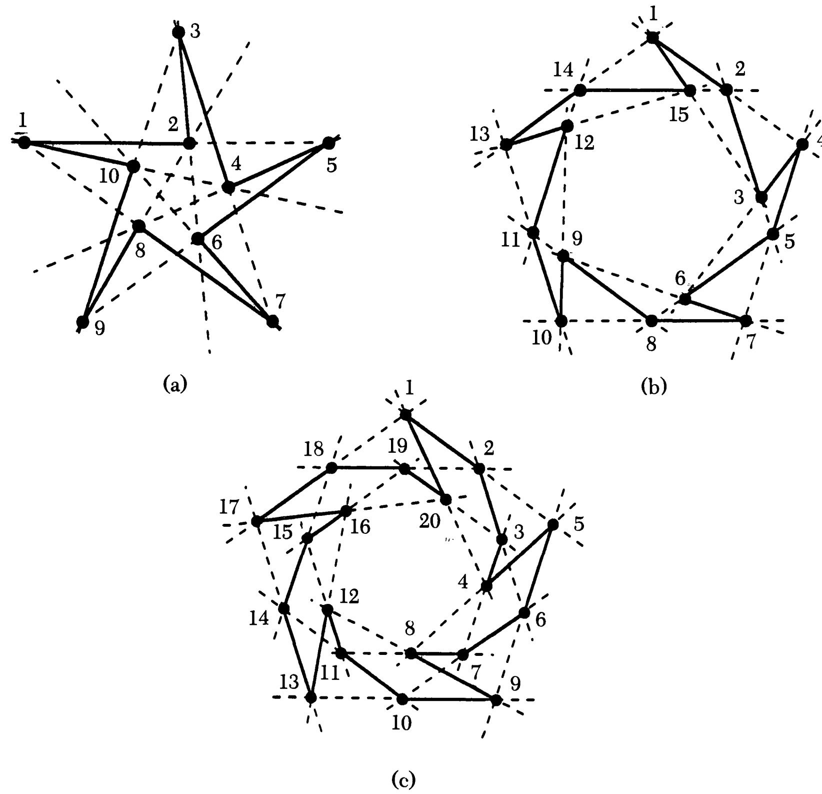


FIGURE 9
Hamiltonian polygons in the configurations of FIGURES 5 and 6(a).

Now, a few words concerning the numbering of the points in our drawings. The combinatorial aspect of configurations starts with the fact that every labeled geometric configuration—such as the one in FIGURE 2—leads to a rectangular table of integers where the labels of the three vertices on a line appear as the three entries of a column. Thus FIGURE 2 corresponds to the following table.

TABLE 1

1	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	1
4	5	6	7	8	9	1	2	3

We used the first two rows of this table to define the polygon and the multilateral described in the preceding paragraph. The third row, of course, lists in each column the third vertex of the configuration on each of the lines through the first two vertices of the same column. The table is an example of a *combinatorial* (or "strategic") 9_3 configuration. The interplay between combinatorial and geometric configurations leads to many results and to a variety of interesting unsolved problems.

Rules for the formation of various combinatorial n_3 configurations can be given easily (see, for example, [14]). In most enumerations of configurations, the first step consists in the construction of combinatorial configurations, followed by elimination of repetitions, that is, configurations that can be made to coincide by renaming the vertices and the lines. But such combinatorial configurations may or may not be realizable as geometric configurations in the Euclidean (or projective) plane, thus leading to many questions. Briefly, the known facts are as follows. To form a combinatorial n_3 configuration, n must be greater than 6. There is a unique (up to labeling) 7_3 and a unique 8_3, both of considerable importance in many parts of modern mathematics; neither of these two configurations is geometrically realizable (see, for example, [5]). There exist three distinct 9_3 combinatorial configurations that are all realizable [11], 10 distinct 10_3 combinatorial configurations all but one of which are realizable (see [4]), as well as 31 11_3 and 229 12_3 combinatorial configurations, all of which are realizable. For the 11_3 configurations, which have been independently enumerated by several authors (see [13], [2], [7]), diagrams for all cases have been provided by Daublebski [3]. However, diagrams can be misleading (see FIGURE 10 for a "fake" configuration that pretends to realize the one 10_3 configuration that is not geometrically realizable). The question of the realizability of the 11_3 configurations was finally settled only recently by Sturmfels and White [20], [21]. Daublebski [3] enumerated the combinatorial configurations 12_3 and found 228 different ones. This was long considered to be the correct number, but recently Gropp [9] found an additional configuration 12_3, which is specified in Table 2. (We are indebted to Dr. Gropp for making available to us his still unpublished results and for permission to refer to them.) The 228 configurations 12_3 found by Daublebski were shown by Sturmfels and White [20], [21] to be geometrically realizable. Gropp's "new" configuration is also geometrically realizable; a diagram for it is shown in FIGURE 11, and coordinates for the vertices (which show that the realization is no "fake") can be computed quite easily. (The coordinates underlying the diagram in FIGURE 11 were found by assuming arbitrary but convenient positions, compatible with the collinearities, for all points except 1, 5, and 7, and computing the location of point 1 for which all required collinearities take place.) Gropp [8] reports that there are 2,036 combinatorial configurations 13_3, and 21,399 combinatorial configurations 14_3. One of the 14_3 configurations is not connected; it consists of two copies of the 7_3 configuration, and is therefore not geometrically realizable. It is not known whether all 13_3 configurations and the 21,398 connected 14_3 configurations are geometrically realizable. For $n > 14$ no enumeration of the possible configurations (combinatorial or geometric) has been carried out. The statement by Steinitz [18], [19] that for $n \geq 11$ all n_3 configurations are "probably realizable" is contradicted by the example in FIGURE 12.

TABLE 2

1	2	3	4	5	6	7	8	9	10	11	12
2	4	5	8	1	7	12	10	3	9	6	11
3	6	7	12	4	1	9	2	11	5	8	10

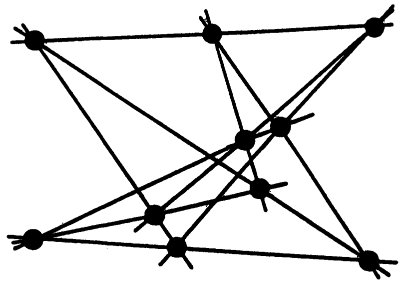


FIGURE 10

A geometric "realization" of one of the 10 combinatorial configurations 10_3—the one that, in fact, is not realizable. In this diagram lines that are only approximately straight have been used, thus leading to the illusion of realizability.

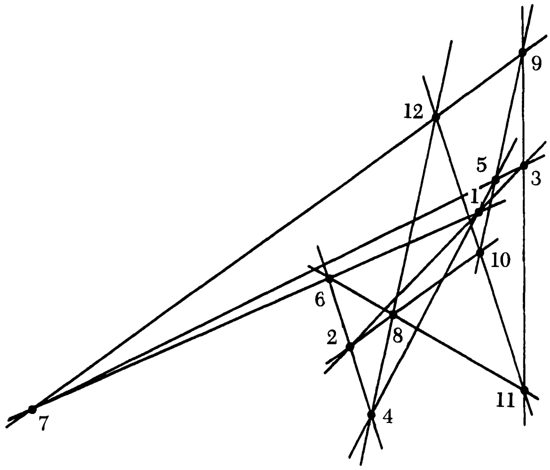


FIGURE 11
A geometric realization of the 12_3 configuration discovered by Gropp [9] and specified in Table 2.

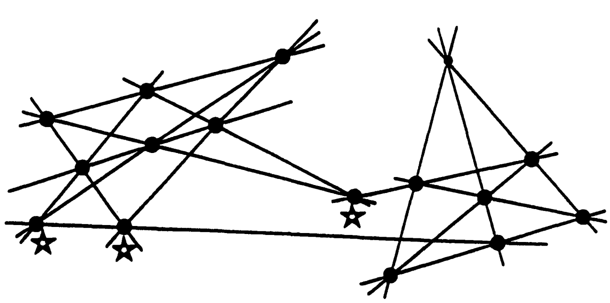


FIGURE 12
Another "fake" configuration. Although it appears to be a 16_3 configuration of points and lines, the combinatorial configuration 16_3 that seems to underlie it cannot be realized by straight lines meeting by threes as shown: The nine points and eight lines in the left half of the diagram satisfy the conditions of the famous theorem of Pappus from projective geometry, which implies that the three points marked by asterisks are necessarily collinear.

Regarding questions of realizability, it may be mentioned that some years ago the junior author of this article conjectured that every n_3 configuration that is geometrically realizable in the Euclidean plane is also geometrically realizable in the *rational plane*, that is, in the subplane of the Euclidean plane consisting of those points both of whose coordinates are rational numbers. This conjecture has been verified by Sturmfels and White [20], [21] for all the configurations 11_3, and for the 228 configurations 12_3 found by Daublebski [3]. That Gropp's configuration 12_3 is realizable in the rational plane recently was communicated to us by Bernd Sturmfels; the conjecture is still undecided for all values of n larger than 13.

Hamiltonian multilaterals

The reader has doubtlessly observed that the third row of TABLE 1 is surprisingly regular. Tables of this nature—in which each row is just a cyclic permutation of the first—are called *cyclic*; the tables for all figures shown in this note (except FIGURES 10 to 14) are either cyclic, or differ from cyclic ones by cyclic displacements of some subsets of the integers. For example, the third row, shown below, of the table for FIGURE 3(b) is of the latter kind, involving a cyclic permutation of the boxed integers.

$$6 \quad \boxed{12} \quad 8 \quad 9 \quad 10 \quad 11 \quad \boxed{2} \quad 13 \quad 14 \quad 15 \quad 1 \quad \boxed{7} \quad 3 \quad 4 \quad 5$$

Different labelings of the vertices of our configurations lead to different self-inscribed (that is, Hamiltonian) multilaterals, though not every labeling yields such a multilateral. In [6] the 36 different 9-laterals of the Pappus configuration 9_3 (that is essentially different from the 9_3 configuration in FIGURE 2) are shown, together with the 24 10-laterals for the Desargues 10_3 configuration; see also [1] and [6] for six different cycles of three mutually inscribed trilaterals in the Pappus 9_3, and for six pairs of mutually inscribed 5-laterals in the Desargues 10_3. The geometrically not realizable 7_3 and 8_3 combinatorial configurations also admit cyclic tables (see, for example, [11]). On the other hand, it should be stressed that not every geometric n_3 configuration admits a Hamiltonian multilateral. Clearly, an 18_3 consisting of two unrelated copies of 9_3 cannot have a Hamiltonian multilateral. However, even connected configurations (that is, configurations in which it is possible to reach from any vertex to any other vertex by steps along lines of the configuration) may lack Hamiltonian multilaterals. This was first established by Steinitz [17], refuting a claim made by Kantor [12] that every connected n_3 has a Hamiltonian multilateral. It is known (see Gropp [10]) that all connected configurations n_3 with $n \leq 14$ have Hamiltonian multilaterals. The smallest known example of a non-Hamiltonian connected geometric configuration is the 22_3 shown in FIGURE 13, but it may well be that there exist smaller ones.

In this context we would like to clarify and correct a statement made in an earlier paper. Given a configuration n_3 (combinatorial or geometric), it is always possible to choose for its presentation a table in which all n labels appear in each of the rows (hence each appears precisely once in each row). This is a nontrivial fact, which was established by Steinitz [16]. It is equivalent to the assertion that for any given configuration, one can find a finite family of polygons such that each vertex of the configuration is also a vertex of exactly one of the polygons. (In other words, although —as mentioned above—not every configuration admits a Hamiltonian multilateral, every configuration does admit a family of multilaterals that together act in a "Hamiltonian" way.) Due to a misinterpretation of the sources, it was stated in Page and Dorwart [14], p. 83, that a representation of this kind is not always possible for configurations n_3 with $n = 12$ or larger. By Steinitz's theorem, this statement is incorrect.

Are these figures oxymora? The answer is **both** yes and no. A "cycle of inscribed multilaterals" and "self-inscribed multilaterals" are certainly *figures* (of speech) in which terms with meanings contradictory in the colloquial sense are combined; hence each is—by definition—an oxymoron. However, it is hoped that the reader has found the FIGURES (or drawings) shown in this article to be interesting as well as clear and unambiguous; they are not oxymora.

Questions for readers

For those who have found the above material stimulating, a few final challenges:

Can you find three cyclically inscribed quadrilaterals in the 12_3 configuration of FIGURE 3(a)?

Can you decide whether the 12_3 configurations in FIGURES 3(a) and 4(a) are essentially different, or can their vertices be relabeled so that the same triplets of labels correspond to collinear sets of vertices in both configurations?

Can you find a pair of mutually inscribed 6-laterals in the 12_3 configuration of FIGURE 4(a)?

Can you give a short argument to show that the 22_3 configuration in FIGURE 13 admits no Hamiltonian multilateral?

Can you find a series of three cyclically inscribed trilaterals in the Pappus configuration 9_3 shown in FIGURE 14? Can you find a Hamiltonian 9-lateral in it?

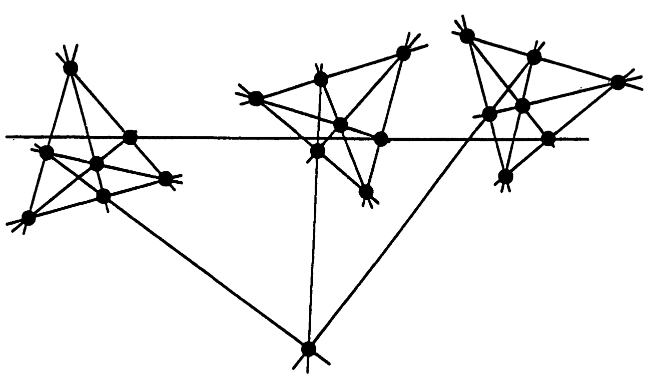


FIGURE 13

A 22_3 configuration that is connected but does not admit a Hamiltonian multilateral.

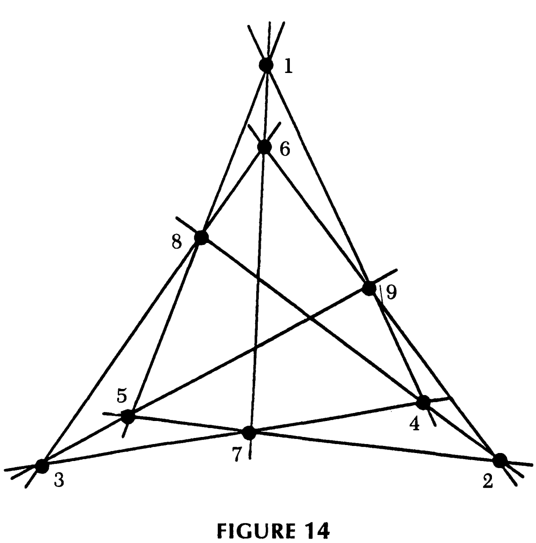


FIGURE 14

One geometric realization of the Pappus configuration 9_3.

Can you find a Hamiltonian multilateral in the 12_3 configuration of FIGURE 11 and TABLE 2?

Can you continue the series of "star-shaped" configurations shown in FIGURE 6, and find analogous $14_3, 18_3, \ldots$ configurations, consisting of pairs of mutually inscribed k-laterals? Can you find more than one way of continuing the series?

Note. The experimentation suggested at the end of Section 1, as well as many other investigations of configurations, can be carried out very conveniently using the "Geometer's Sketchpad" software for Macintosh computers. This software has just been released by Key Curriculum Press.

The authors are indebted to the anonymous referees for helpful comments.

REFERENCES

1. H. S. M. Coxeter, Self-dual configurations and regular graphs, *Bull. Amer. Math. Soc.* 56(1950), 413–455. Reprinted in *Twelve Geometric Essays*, Southern Illinois University Press, Carbondale, IL, 1968, pp. 106–149.
2. R. Daublebski von Sterneck, Die Configurationen 11_3, *Monatshefte Math. Phys.* 5(1894), 325–330.
3. R. Daublebski von Sterneck, Die Configurationen 12_3, *Monatshefte Math. Phys.* 6(1895), 223–254.
4. H. L. Dorwart, *Configurations*, WFF 'N PROOF Publishers, Ann Arbor, MI, 1968.
5. H. L. Dorwart, *The Geometry of Incidence*, Prentice-Hall, Englewood Cliffs, NJ, 1966. Reprinted as paperback by WFF 'N PROOF Publishers, Ann Arbor, MI, 1973.
6. H. L. Dorwart, Configurations: a case study in mathematical beauty, *Math. Intelligencer* 7(1985), 39–48.
7. H. Gropp, "Il metodo di Martinetti," or Configurations and Steiner systems S(2, 4, 25), *Ars Combinatoria* 24B(1987), 179–188.
8. H. Gropp, On the existence and non-existence of configurations n_k, *J. Combin. Inform. System Sci.*, 15(1990), 34–48.
9. H. Gropp, Configurations and Steiner systems S(2, 4, 25). II—Trojan configurations n_3 (to appear).
10. H. Gropp, Configurations and the Tuttle conjecture, *Ars Combinatoria* 29A(1990), 171–177.
11. D. Hilbert and S. Cohn-Vossen, *Anschauliche Geometrie*, Springer, Berlin, 1932; reprinted by Dover Publishing Co., New York, 1944. English translation: *Geometry and the Imagination*. Chelsea Publishing Co., New York, 1952.
12. S. Kantor, Die Configurationen $(3,3)_{10}$, *K. u. k. Academie der Wiss., Vienna, S.-ber. math.-nat. Cl.* 84 II (1881), 1291–1314.
13. V. Martinetti, Sulle configurazioni piane μ_3, *Annali di Matematica* 15(1887/88), 1–26.
14. W. Page and H. L. Dorwart, Numerical patterns and geometrical configurations, this MAGAZINE 57(1984), 82–92.
15. A. Schönflies, Ueber die regelmässigen Configurationen n_3, *Math. Annalen* 31(1881), 43–69.
16. E. Steinitz, Über die Construction der Configurationen n_3, Ph.D. thesis, Universität Breslau, Breslau, 1894.
17. E. Steinitz, Über die Unmöglichkeit, gewisse Configurationen n_3 in einem geschlossenen Zuge zu durchlaufen, *Monatshefte Math. Phys.* 8(1897), 293–296.
18. E. Steinitz, *Konfigurationen der projektiven Geometrie*, *Encyklop. Math. Wiss.* 3(Geometrie)(1910), 481–516.
19. E. Steinitz and E. Merlin, Configurations, *Encyclopédie des Sciences Mathématiques*, *Édition française*, Tome III, Vol. 2, 1913, pp. 144–160.
20. B. Sturmfels and N. White, *Rational Realizations of* 11_3- *and* 12_3*-configurations*, University of Minnesota Institue for Mathematics and Applications, Preprint Series #389, January 1988, Minneapolis, MN, pp. 92–123.
21. B. Sturmfels and N. White, All 11_3 and 12_3-configurations are rational, *Aequationes Mathematicae* 39(1990), 254–260.
22. J. van de Craats, On Simonis' 10_3 configuration, *Nieuw Archief voor Wiskunde* (4), 1(1983), 193–207.
23. M. Zacharias, Streifzüge im Reich der Konfigurationen: Eine Reyesche Konfiguration (15_3), Stern- und Kettenkonfigurationen, *Math. Nachrichten* 5(1951), 329–345.

NOTES

More on the Four-Numbers Game

DANIEL ULLMAN
The George Washington University
Washington, DC 20052

Place arbitrary integers on the four corners of a square. Then place on the midpoint of each side of the square the absolute value of the difference of the numbers associated with the adjacent corners. Connect the midpoints of the sides of the square to form a new square with integers on its corners. Now repeat the process. FIGURE 1 shows an example starting with the numbers 1, 2, 4, and 7.

Much has been written about this process ([1], [3], [4], [5], [6], [7])—the so-called four-numbers game—and its generalizations [8]. The earliest published reference seems to be in [2], where it is attributed to E. Ducci of Italy. It is a simple exercise to show that upon iteration the procedure eventually produces a square of zeroes. What is surprising is how fast this convergence actually happens in practice. Ask someone to pick four numbers at random, play the game, and you will probably find it takes eight or fewer iterations to converge to the zero square. This is despite the fact that the convergence time is unbounded—a slightly harder exercise. It is the purpose of this note to calculate the distribution of convergence times with respect to the natural probability measure on labeled squares and thereby to explain the surprising speed of convergence.

Let us generalize the process slightly by permitting real numbers on the corners of the squares. For convenience, we formulate the problem as follows. Let T: $R^4 \to R^4$ be defined by $T(a,b,c,d) = (|a-b|, |b-c|, |c-d|, |d-a|)$. For $\vec{v} \in R^4$, let the convergence time of the four-numbers game starting at $\vec{v}$ be $n(\vec{v}) = \min\{m\colon m > 0$ and $T^m(\vec{v}) = (0,0,0,0)\}$. We wish to calculate the probability that $n(\vec{v}) = k$ for small natural numbers k. Since we have no way of making sense of choosing an integer or a real number "at random," let us assume that the numbers are chosen according to a uniform distribution on $[0, N]$ for some very large N. (In fact, the distribution of the function n is easily seen to be independent of the choice of N.)

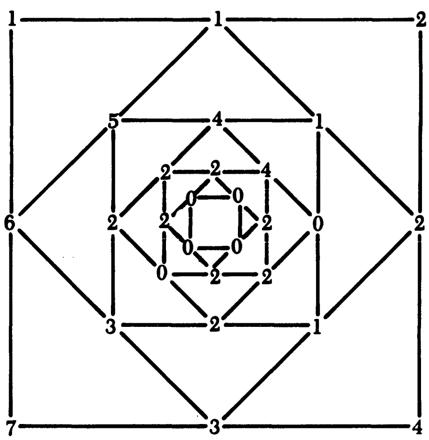


FIGURE 1
An example of the four-numbers game.

Write $\vec{e}$ for (1, 1, 1, 1). Let us say that two vectors $\vec{v}$ and $\vec{w}$ in R^4 are equivalent and write $\vec{v} \sim \vec{w}$ if $\vec{v}$ can be transformed into $\vec{w}$ by a series of moves of the following types:

1) Multiply $\vec{v}$ by a nonzero scalar.
2) Add to $\vec{v}$ a multiple of $\vec{e}$.
3) Cyclically permute the entries of $\vec{v}$.
4) Reverse the entries of $\vec{v}$ (i.e., turn (a, b, c, d) into (d, c, b, a)).

It is easy to see that $\sim$ is an equivalence relation. One equivalence class is $\{\vec{v}: \vec{v} = (a, a, a, a)$ for some real number $a\}$. A little thought will soon convince one that in every other equivalence class is a unique representative $\vec{v} = (a, b, c, d)$ with

1) $a = 0$,
2) $\max\{c, d\} = 1$,
3) $b \leqslant d$, and
4) $b + c + d \leqslant 2$.

(If a vector $\vec{v}$ satisfies only the first three properties here, then $\vec{e} - \vec{v}$, cyclically permuted or reversed, will satisfy all four.) Here is an example of this reduction: $(4, 2, 1, 7) \sim (7, 1, 2, 4) \sim (6, 0, 1, 3) \sim (0, 1, 3, 6) \sim (0, 1/6, 1/2, 1)$. By the canonical form of $\vec{v}$, we shall mean the vector satisfying these conditions that is equivalent to $\vec{v}$.

Now $T(r\vec{v} + s\vec{e}) = rT(\vec{v})$, and hence $n(r\vec{v} + s\vec{e}) = n(r\vec{v}) = n(\vec{v})$. This, along with the fact that $n(\vec{v})$ does not change if $\vec{v}$ is cyclically permuted or reversed, implies that n is constant on equivalence classes. We choose to calculate $n(\vec{v})$ only for $\vec{v}$ in canonical form.

It is not hard to see that there are basically only four kinds of canonical forms. The canonical form of a vector $\vec{v}$ is either of

type 0)	$(0, 0, 0, 0)$, or $(0, 0, x, 1)$ with $0 \leqslant x \leqslant 1$, or $(0, x, 0, 1)$ with $0 \leqslant x \leqslant 1$, or $(0, x, x, 1)$ with $0 \leqslant x \leqslant 1/2$;
type 1)	$(0, y, x, 1)$, with $0 < x < y < 1 - x$;
type 2)	$(0, x, 1, y)$, with $0 < x < y < 1 - x$; or
type 3)	$(0, x, y, 1)$, with $0 < x < y < 1 - x$.

If $\vec{v}$ is chosen at random in the sense described above, then the event that $\vec{v}$ is type 0 is of probability zero, the other three kinds of canonical forms are equally likely, and the resulting x and y end up distributed uniformly on the triangle $0 < x < y < 1 - x$. (One way to see this is to note that the event that all the components of $\vec{v}$ are between m and M forms a tesseract—i.e., a four-dimensional cube—in (a, b, c, d) space. The event that the smallest entry of $\vec{v}$ is m and the largest is M is the union of 12 of the 24 two-dimensional faces of this tesseract. Four of these faces correspond to type 2 vectors, while eight others can be divided in half by a diagonal, on one side of which lie type 1 vectors and on the other side of which lie type 3 vectors. On all faces, the two remaining coordinates are uniformly distributed on the square, and it follows that, for all three types, the (x, y) is uniformly distributed over the triangle $0 < x < y < 1 - x$.)

We would like to calculate $n(\vec{v})$ for vectors $\vec{v}$ of types 1, 2, and 3. Vectors of type 1 are easy to dismiss. (This was observed in [**5**], and [**6**]; we repeat the short calculation for the reader's convenience.) If $\vec{v} = (0, y, x, 1)$, with $0 < x < y < 1 - x$, then $T(\vec{v}) = (y, y - x, 1 - x, 1)$, $T^2(\vec{v}) = (x, 1 - y, x, 1 - y)$, $T^3(\vec{v}) = (1 - x - y, 1 - x - y, 1 - x - y, 1 - x - y)$, and $T^4(\vec{v}) = (0, 0, 0, 0)$. Hence all type 1 vectors $\vec{v}$ have $n(\vec{v}) = 4$.

If $\vec{v}$ is of type 2, then $T(\vec{v}) = (x, 1-x, 1-y, y)$ and

$$T^2(\vec{v}) = \begin{cases} (1-2x, y-x, 2y-1, y-x), & \text{if } y \geqslant 1/2; \\ (1-2x, y-x, 1-2y, y-x), & \text{if } y < 1/2. \end{cases}$$

In the first case, $T^3(\vec{v}) = (1-x-y)\vec{e}$, $T^4(\vec{v}) = \vec{0}$, and so $n(\vec{v}) = 4$. In the second case, $T^3(\vec{v}) = (1-x-y, |1+x-3y|, |1+x-3y|, 1-x-y)$, $T^4(\vec{v}) = (c, 0, c, 0)$ where $c = 2y - 2x$ if $1 + x > 3y$ and $c = 2 - 4y$ otherwise. (Note that $c > 0$ since $x < y$ and $y < 1/2$.) It follows that $T^5(\vec{v}) = c\vec{e}$, $T^6(\vec{v}) = \vec{0}$, and so $n(\vec{v}) = 6$. The two cases are equally likely.

It is the class of vectors of type 3 that exhibits unbounded convergence time. The results of an analysis akin to the one in the previous paragraph are reported in FIGURE 2 and the theorem that follows. FIGURE 2 is a drawing of the triangle $0 < x < y < 1 - x$ showing the value of $n(x, y) := n(0, x, y, 1)$. (A similar picture appears in [3].) Notice that n is constant on polygonal regions that are bounded by lines $ax + by = c$ where a, b, and c are, in fact, integers. No claim is made about $n(x, y)$ if (x, y) is on one of these lines, but the probability of this occurring is easily seen to be zero.

To formalize all this, we capture the essential features of FIGURE 2 in the following theorem, whose proof is sketched.

THEOREM. *Let* $P_1 = (0,1)$, $P_2 = (0,0)$, $P_3 = (1/2, 1/2)$, $P_4 = (1/3, 2/3)$, $P_5 = (0, 1/2)$, $P_6 = (0, 1/3)$, *and for* $n \geqslant 7$, *let* P_n *be the point of intersection of* $\overline{P_{n-1}P_{n-4}}$ *and* $\overline{P_{n-3}P_{n-5}}$. *Define* $A_1 = \triangle P_1P_4P_5$, $A_{2n} = \triangle P_{n+1}P_{n+2}P_{n+6}$ *for* $n \geqslant 1$, *and* $A_{2n+1} = \triangle P_{n+1}P_{n+5}P_{n+6}$ *for* $n \geqslant 1$. *Then the triangles* $A_1, A_2, A_3, \ldots$ *have pairwise disjoint interior, the closure of whose union is the entire closed region* $\triangle P_1P_2P_3$. *Moreover, if* (x, y) *is in the interior of triangle* A_1 *or* A_2, *then* $n(x, y) = 5$; *if* (x, y) *is in the interior of triangle* A_3, *then* $n(x, y) = 6$; *if* (x, y) *is in the interior of triangle* A_{4m+j} *where* $m \geqslant 1$ *and* $j = 0$, 1, *or* 3, *then* $n(x, y) = 6 + m$; *and if* (x, y) *is in the interior of triangle* A_{4m+2} *where* $m \geqslant 1$, *then* $n(x, y) = 5 + m$.

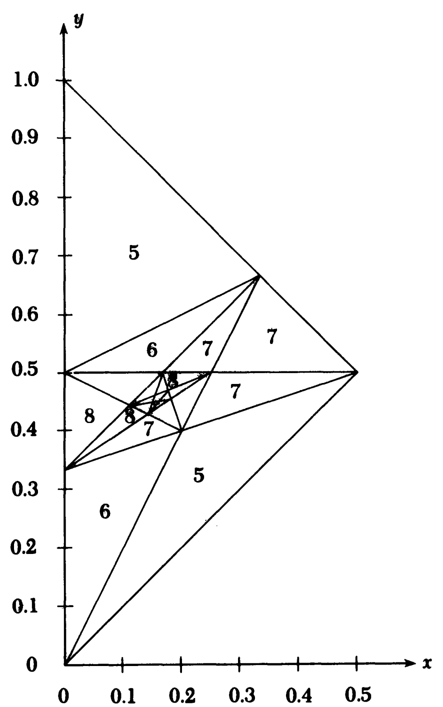


FIGURE 2
The function $n(x, y)$ for type 3 vectors.

Sketch of Proof. The first claim is made clear by FIGURE 2. The proof of the claim made in the last sentence is by induction on m. One first shows that $n(x,y)$ is as claimed when (x,y) is in A_k for $1 \leqslant k \leqslant 8$, say. This is done in the same way as the analysis for type 1 and type 2 vectors above. This verification is a bit tedious but straightforward.

For the inductive step, define a map S from the entire triangle $\triangle P_1P_2P_3$ into itself by

$$S(x;y) = \left(\frac{1-x-y}{3-x-y}, \frac{2-2y}{3-x-y}\right).$$

If $S(x,y)=(w,z)$, then

$$\begin{aligned} T(0,w,z,1) &= \left(\frac{1-x-y}{3-x-y}, \frac{1+x-y}{3-x-y}, \frac{1-x+y}{3-x-y}, 1\right) \\ &\sim (1-x-y, 1+x-y, 1-x+y, 3-x-y) \\ &\sim (0, 2x, 2y, 2) \\ &\sim (0, x, y, 1). \end{aligned}$$

If one defines points $Q_1, Q_2, Q_3, \dots$ by $Q_k = P_k$ for $1 \leqslant k \leqslant 3$ and $Q_k = S(Q_{k-2})$ for $k > 3$, then the fact that S preserves collinearity—if (x,y) lies on the line $ay + by + c = 0$, then $S(x,y)$ lies on the line $(2a+b+c)x + (-a+b)y + (-b-c) = 0$ —together with the recursive definition of P_n implies that $P_n = Q_n$ for all $n \geqslant 1$. Hence $P_n = S(P_{n-2})$. The fact that S preserves lines is now used again to imply that $A_n = S(A_{n-4})$. The inductive step now follows easily.

FIGURE 2 can be interpreted as a picture of the dynamics associated with the transformation S. If (x,y) is in a triangle marked with a k, then $S(x,y)$ will be in a triangle marked $k+1$. The orbits $(x,y), S(x,y), S^2(x,y), \dots$ spiral in toward the unique fixed point (a,b) of S. Setting $S(a,b)=(a,b)$ tells us that a is the real root of the cubic polynomial $x^3 - 5x^2 + 7x - 1 = 0$ and that $b = 3a - a^2$. (See [1], p. 27 for related equations.) The decimal expansions of these two constants are $a = .160713\dots$ and $b = .456311\dots$ and $(0,a,b,1)$ is the unique vector in canonical form that does not converge to the zero vector upon iteration of T.

We can now tabulate the probability that $n(\vec{v}) = k$ for small k by adding areas (properly scaled) of triangles in FIGURE 2 and taking into account vectors of types 1 and 2. We present these data in TABLE 1. The large values for $k=4$ and $k=6$ result from the influence of type 1 and type 2 vectors. As the table shows, the four-numbers game converges in eight or fewer steps more than 99% of the time.

TABLE 1. The probability that the four-numbers game converges in k steps.

k	$\text{Prob}\{n(\vec{v})=k\}$
0	0
1	0
2	0
3	0
4	50.0000%
5	17.7778%
6	22.9630%
7	6.0053%
8	2.3680%

k	$\text{Prob}\{n(\vec{v})=k\}$
9	0.6182%
10	0.1886%
11	0.0559%
12	0.0163%
13	0.0049%
14	0.0014%
15	0.0004%
16	0.0001%
⋮	⋮

We can also calculate the asymptotic rate of decrease of the numbers in TABLE 1. If triangle A_n has corners (x, y), (x', y') and (x'', y''), then its area is given by the well-known formula

$$\frac{1}{2}\begin{vmatrix} 1 & 1 & 1 \\ x & x' & x'' \\ y & y' & y'' \end{vmatrix}.$$

Hence the ratio of the area of A_n to the area of A_{n+4} is given by

$$\begin{vmatrix} 1 & 1 & 1 \\ x & x' & x'' \\ y & y' & y'' \end{vmatrix} \Bigg/ \begin{vmatrix} 1 & 1 & 1 \\ \dfrac{1-x-y}{3-x-y} & \dfrac{1-x'-y'}{3-x'-y'} & \dfrac{1-x''-y''}{3-x''-y''} \\ \dfrac{2-2y}{3-x-y} & \dfrac{2-2y'}{3-x'-y'} & \dfrac{2-2y''}{3-x''-y''} \end{vmatrix}$$

$$= (3-x-y)(3-x'-y')(3-x''-y'')/4.$$

Since the corners of A_n approach (a, b) as n increases, the limit of this ratio is $(3-a-b)^3/4 = 3.382976\ldots$. It is easy to see that the ratio of successive entries in TABLE 1 approaches this same limit.

Thanks are due to Jim Propp and a referee for useful comments. I would also like to acknowledge conversations with David Pitts, Hassan Sedaghat, and Jeff Shallit.

REFERENCES

1. E. R. Berlekamp, The design of slowly shrinking labelled squares, *Math. of Comp.* 29 (1975), 25–27.
2. C. Ciamberlini and A. Marengoni, Sa una interessante curiosità numerica, *Period. Mat. Ser.* 4 (1937), 25–30.
3. M. Dumont and J. Meeus, The four-numbers game, *J. Rec. Math.* 13 (1981), 89–96.
4. R. Honsberger, *Ingenuity in Mathematics*, Random House, New York, 1970.
5. Z. Magyar, A recursion on quadruples, *Amer. Math. Month.* 91 (1984), 360–362.
6. L. Meyers, Ducci's four-number problem: a short bibliography, *Crux Math.* 8 (1982), 262–266.
7. W. A. Webb, The length of the four-number game, *Fib. Quart.* 20 (1982), 33–35.
8. F. B. Wong, Ducci processes, *Fib. Quart.* 20 (1982), 97–105.

A Note on Solid Angles

A fundamental mathematical fact is that the length c of a chord of a circle of radius 1 that subtends at the center of the circle an angle of γ radians is a transcendental function of γ. It is therefore quite surprising that the three-dimensional analog is false. Let Δ be the area of a cap of a sphere of radius 1 whose base is a disc of area D. Then

$$D = \Delta[1 - \Delta/4\pi].$$

Observe that Δ measures a solid angle subtended by the disc at the center of the sphere.

—OVED SHISHA
THE UNIVERSITY OF RHODE ISLAND
KINGSTON, RI 02881

We can also calculate the asymptotic rate of decrease of the numbers in TABLE 1. If triangle A_n has corners (x,y), (x',y') and (x'',y''), then its area is given by the well-known formula

$$\frac{1}{2}\begin{vmatrix} 1 & 1 & 1 \\ x & x' & x'' \\ y & y' & y'' \end{vmatrix}.$$

Hence the ratio of the area of A_n to the area of A_{n+4} is given by

$$\begin{vmatrix} 1 & 1 & 1 \\ x & x' & x'' \\ y & y' & y'' \end{vmatrix} \Bigg/ \begin{vmatrix} 1 & 1 & 1 \\ \dfrac{1-x-y}{3-x-y} & \dfrac{1-x'-y'}{3-x'-y'} & \dfrac{1-x''-y''}{3-x''-y''} \\ \dfrac{2-2y}{3-x-y} & \dfrac{2-2y'}{3-x'-y'} & \dfrac{2-2y''}{3-x''-y''} \end{vmatrix}$$

$$=(3-x-y)(3-x'-y')(3-x''-y'')/4.$$

Since the corners of A_n approach (a,b) as n increases, the limit of this ratio is $(3-a-b)^3/4 = 3.382976\ldots$. It is easy to see that the ratio of successive entries in TABLE 1 approaches this same limit.

Thanks are due to Jim Propp and a referee for useful comments. I would also like to acknowledge conversations with David Pitts, Hassan Sedaghat, and Jeff Shallit.

REFERENCES

1. E. R. Berlekamp, The design of slowly shrinking labelled squares, *Math. of Comp.* 29 (1975), 25–27.
2. C. Ciamberlini and A. Marengoni, Sa una interessante curiosità numerica, *Period. Mat. Ser.* 4 (1937), 25–30.
3. M. Dumont and J. Meeus, The four-numbers game, *J. Rec. Math.* 13 (1981), 89–96.
4. R. Honsberger, *Ingenuity in Mathematics*, Random House, New York, 1970.
5. Z. Magyar, A recursion on quadruples, *Amer. Math. Month.* 91 (1984), 360–362.
6. L. Meyers, Ducci's four-number problem: a short bibliography, *Crux Math.* 8 (1982), 262–266.
7. W. A. Webb, The length of the four-number game, *Fib. Quart.* 20 (1982), 33–35.
8. F. B. Wong, Ducci processes, *Fib. Quart.* 20 (1982), 97–105.

A Note on Solid Angles

A fundamental mathematical fact is that the length c of a chord of a circle of radius 1 that subtends at the center of the circle an angle of γ radians is a transcendental function of γ. It is therefore quite surprising that the three-dimensional analog is false. Let Δ be the area of a cap of a sphere of radius 1 whose base is a disc of area D. Then

$$D = \Delta[1 - \Delta/4\pi].$$

Observe that Δ measures a solid angle subtended by the disc at the center of the sphere.

—OVED SHISHA
THE UNIVERSITY OF RHODE ISLAND
KINGSTON, RI 02881

A Replication Property for Magic Squares

DANIEL B. SHAPIRO
Ohio State University
Columbus, OH 43210

1. Introduction A curious property of a 3×3 magic square is exhibited when the entries are added together in 2×2 blocks. For these additions we consider the rows and columns of the square to "wrap around." As an example let us consider the square array A below, which is "magic" since each row and column (and diagonal) adds up to the same value, 15.

$$A = \begin{pmatrix} 2 & 9 & 4 \\ 7 & 5 & 3 \\ 6 & 1 & 8 \end{pmatrix}.$$

A new square A' is derived from the original one by adding the entries in 2×2 blocks, with wrap-around. For example, the $(1,1)$ entry of A' is 23, which is the sum of the entries in $\begin{pmatrix} 2 & 9 \\ 7 & 5 \end{pmatrix}$. The $(1,2)$ entry is 21, coming from the block $\begin{pmatrix} 9 & 4 \\ 5 & 3 \end{pmatrix}$ and the $(1,3)$ entry is 16, coming from the block $\begin{pmatrix} 4 & 2 \\ 3 & 7 \end{pmatrix}$. Then the full derived square is

$$A' = \begin{pmatrix} 23 & 21 & 16 \\ 19 & 17 & 24 \\ 18 & 22 & 20 \end{pmatrix}.$$

It is interesting that this pattern involves nine consecutive integers and has constant row and column sums. In fact, Dr. Matrix, the well-known numerologist, knew this property when he was seven, but never published it (see [**2**], p. 296). Furthermore, subtracting 15 from every entry of A', we get the matrix

$$\begin{pmatrix} 8 & 6 & 1 \\ 4 & 2 & 9 \\ 3 & 7 & 5 \end{pmatrix},$$

which is just a shifted version of the original A. It is obtained from A by cycling the rows two steps up and cycling the columns two steps to the left.

Generally if A is any $m \times n$ matrix, define A' to be the matrix derived from A by the process described above. This $m \times n$ matrix A' will be called the *derived matrix*. We say that A has the *replication property* if its derived matrix A' equals a constant plus a shifted version of A.

Which matrices possess this "replication property"?

We will answer this question completely using methods of linear algebra and properties of complex roots of unity. Our analysis illustrates the power of translating such "recreational" problems into the language of linear algebra.

If A has the replication property then the larger matrix $B = \begin{pmatrix} A & A \\ A & A \end{pmatrix}$ also has the property. We define such "blow-ups" more generally.

Definition 1.1 If A is an $m \times n$ matrix, a $j \times k$-*blow-up* of A is defined to be a $jm \times kn$ matrix B obtained from a $j \times k$ array of copies of A.

For example any constant matrix is a blow-up of a 1×1 matrix. It is easy to see that a blow-up of A has the replication property if and only if A does. We will prove that any array with the replication property is a linear combination of blow-ups of certain magic squares of sizes 1×1, 3×3, and 12×12.

In this presentation, a matrix is called a "magic square" if its row sums and column sums are all equal (without any restrictions on the entries or on the diagonal sums). The following Lemma is left as an elementary exercise for the reader. It is also a consequence of the theory developed in the next section.

LEMMA 1.1. *Any 3×3 magic square has the replication property.*

2. Matrix formulation Let A be an $m \times n$ matrix with entries in $\mathbb{C}$, the field of complex numbers. We will express the derived matrix A' in terms of matrix multiplication. Let P be the following $m \times m$ permutation matrix:

$$P = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Recall that a "permutation matrix" is the matrix of a linear transformation that permutes the basis vectors. Equivalently, every entry of the matrix is 0 or 1, and there is exactly one 1 in each row and in each column.

Let Q be the $n \times n$ permutation matrix formed in the same way. The product matrix AQ is obtained from A by cycling the columns one step to the left, with the first column becoming the last. Similarly P^*A is obtained from A by cycling the rows one step up, with the top row becoming the bottom. (The asterisk here denotes the conjugate transpose.) Finally we see that P^*AQ is obtained from A by cycling both the columns and the rows. Therefore, the derived matrix is exactly

$$A' = A + AQ + P^*A + P^*AQ = (I + P^*)A(I + Q).$$

Here we write I for the identity matrix of the appropriate size. Similarly let J be the matrix whose entries are all equal to 1. The dimensions of J can always be inferred from the context.

By definition, a matrix A has the replication property if the derived matrix A' equals a constant matrix cJ plus a shifted version of A. The "shifted version" of A is obtained by cycling the columns of A some number of times to the left and cycling the rows some number of times up. That is, the shifted version of A is $P^{*r}AQ^s$ for some integers r, s. This notation allows a more precise definition of the replication property.

Definition 2.1. The $m \times n$ matrix A has the replication property if and only if there exist integers r, s and a complex number c satisfying

$$(I + P^*)A(I + Q) = cJ + P^{*r}AQ^s.$$

For example, in the 3×3 case done above we have $(I + P^*)A(I + P) = 15J + P^{*2}AP^2$. The key to handling this condition in general is to employ the eigenvectors of P and Q. For any mth root of unity α in $\mathbb{C}$, define the vector $v_\alpha \in \mathbb{C}^m$ by setting

$$v_\alpha = \begin{pmatrix} 1 \\ \alpha \\ \alpha^2 \\ \vdots \\ \alpha^{m-1} \end{pmatrix}.$$

Since $\alpha^m = 1$ it follows that $Pv_\alpha = \alpha^{m-1}v_\alpha = \bar{\alpha}v_\alpha$, so that v_α is an $\bar{\alpha}$-eigenvector of P. There are m distinct choices for α so these eigenvectors $\{v_\alpha\}$ form a basis of $\mathbb{C}^m$. Moreover this basis is an orthogonal one: The "inner products" are

$$v_{\alpha_1}^* v_{\alpha_2} = \begin{cases} m & \text{if } \alpha_1 = \alpha_2, \\ 0 & \text{if } \alpha_1 \neq \alpha_2. \end{cases}$$

If β is an nth root of unity in $\mathbb{C}$, define the vector $w_\beta \in \mathbb{C}^n$ similarly. Then $Qw_\beta = \bar{\beta}w_\beta$ and $\{w_\beta\}$ forms an orthogonal basis of $\mathbb{C}^n$.

Now suppose A is an $m \times n$ matrix having the replication property. For roots of unity α, β as above, let us multiply the equation in Definition 2.1 on the left by v_α^* and on the right by w_β. To simplify the result note that $v_\alpha^* P^* = (Pv_\alpha)^* = \alpha v_\alpha^*$, and that if $\alpha, \beta \neq 1$ then $v_\alpha^* J = 0$ and $Jw_\beta = 0$. Therefore,

$$\text{if } \alpha \neq 1 \text{ or } \beta \neq 1 \text{ then } (1+\alpha)(1+\bar{\beta})v_\alpha^* A w_\beta = \alpha^r \bar{\beta}^s v_\alpha^* A w_\beta. \qquad (*)$$

To exploit this condition we consider that quantity $v_\alpha^* A w_\beta$. Since the set $\{v_\alpha\}$ forms a basis of $\mathbb{C}^m$ and $\{w_\beta\}$ forms a basis of $\mathbb{C}^n$, the set of matrices $\{v_\alpha w_\beta^*\}$ forms a basis of all $m \times n$ matrices. (The proof of this claim is an interesting exercise in linear algebra.) Thus our $m \times n$ matrix A can be uniquely expressed as

$$A = \sum_{\alpha,\beta} x_{\alpha\beta} v_\alpha w_\beta^*,$$

for suitable scalar coefficients $x_{\alpha\beta}$. Using the inner product properties we find that

$$mn \cdot x_{\alpha\beta} = v_\alpha^* A w_\beta.$$

Therefore, if A has the replication property and if $x_{\alpha\beta} \neq 0$, then $(1+\alpha)(1+\bar{\beta}) = \alpha^r \bar{\beta}^s$.

PROPOSITION 2.2. *If α, β are roots of unity satisfying $(1+\alpha)(1+\bar{\beta}) = \alpha^r\bar{\beta}^s$ for some integers r, s then either* (1) *α and β are primitive* 3*rd roots of unity, or* (2) *α and β are primitive* 12*th roots of unity with $\beta = \alpha^5$ or α^{-5}.*

This Proposition is deduced from a theorem of M. Newman [3] in the Appendix at the end of this article.

PROPOSITION 2.3. *Let A be an $m \times n$ matrix with coefficients $x_{\alpha\beta}$ defined as above where α is an mth root of unity and β is an nth root of unity. If A has the replication property then for each α, β where $x_{\alpha\beta} \neq 0$, one of the following four cases holds. Furthermore, $x_{\alpha\beta} \neq 0$ in case* (3) *and $x_{\sigma\tau} \neq 0$ in case* (4) *cannot both occur.*

(1) $\alpha = \beta = 1$.

(2) *α, β are primitive* 3rd *roots of unity.*

(3) *α is a primitive* 12th *root of unity and $\beta = \alpha^5$.*

(4) *α is a primitive* 12th *root of unity and $\beta = \alpha^{-5}$.*

Conversely, if the coefficients $x_{\alpha\beta}$ satisfy these conditions then A has the replication property.

Proof. Suppose A has the replication property and $x_{\alpha\beta} \neq 0$. If $\alpha = \beta = 1$ we have (1). Otherwise from equation $(*)$ above we find $(1+\alpha)(1+\bar{\beta}) = \alpha^r\bar{\beta}^s$ and Proposition 2.2 implies that α, β must satisfy one of the cases (2), (3), or (4). Finally suppose $x_{\alpha\beta} \neq 0$ for α, β in case (3) and $x_{\sigma\tau} \neq 0$ for some σ, τ in case (4). Then $(1+\alpha)(1+\bar{\beta}) = \alpha^r\bar{\beta}^s$ and $(1+\sigma)(1+\bar{\tau}) = \sigma^r\bar{\tau}^s$. We know that $\alpha^6 = -1$ and $\alpha^4 - \alpha^2 + 1 = 0$, so that $\bar{\beta} = \alpha^{-5} = -\alpha$ and $(1+\alpha)(1+\bar{\beta}) = 1 - \alpha^2 = -\alpha^4 = \alpha^{10}$. Since $\alpha^r\bar{\beta}^s = \alpha^{r-5s}$ we conclude $r - 5s \equiv 10 \pmod{12}$. A similar calculation for σ and τ shows that $r + 5s \equiv 3 \pmod{12}$. These two congruences are incompatible.

Conversely one can check that there exist values r, s of which $(1+\alpha)(1+\bar{\beta}) = \alpha^r\bar{\beta}^s$ for all the roots of unity α, β in cases (1), (2), and (3). Therefore, if A is a linear combination of the matrices $v_\alpha w_\beta^*$ for indices α, β in these cases, then A has the replication property. A similar argument works for the cases (1), (2), and (4). More explicit calculations for these casés are given below.

Proposition 2.3 allows us to list all the matrices with the replication property. Case (1) corresponds just to the constant matrices. If the $m \times n$ matrix A is a (nonzero) combination of matrices from case (2) then it is just a blow-up of a 3×3 matrix. To see this, note that α and β are primitive 3rd roots of unity so that m and n must be multiples of 3. Also v_α and w_β are blow-ups of the vectors $u_\alpha = \begin{pmatrix} 1 \\ \alpha \\ \alpha^2 \end{pmatrix}$ and $u_\beta = \begin{pmatrix} 1 \\ \beta \\ \beta^2 \end{pmatrix}$ so that $v_\alpha w_\beta^*$ is a blow up of the 3×3 matrix $u_\alpha u_\beta^*$.

To analyze the 3×3 case let $X = u_\rho u_\rho^*$ and $Y = u_\rho u_{\rho^2}^*$ where $\rho = e^{2\pi i/3}$. Then

$$X = \begin{pmatrix} 1 & \rho^2 & \rho \\ \rho & 1 & \rho^2 \\ \rho^2 & \rho & 1 \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & \rho^2 & 1 \\ \rho^2 & 1 & \rho \end{pmatrix}.$$

The conjugates of X and Y are $\bar{X} = u_{\rho^2}u_{\rho^2}^*$ and $\bar{Y} = u_{\rho^2}u_\rho^*$, respectively, so all four possibilities for α, β are covered. The four-dimensional space spanned here consists of matrices with the replication property, because X, Y, $\bar{X}$, and $\bar{Y}$ have the replication property with "shifts" $r = s = 2$. The matrices B_i, C_i listed below form another basis of this space. Specifically, $X - \bar{X} = (\rho^2 - \rho)B_0$ and $X + \bar{X} = B_0 + 2B_1$, while C_0, C_1 are related similarly to Y.

$$B_0 = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix},$$

$$C_0 = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \quad C_1 = \begin{pmatrix} 0 & 1 & -1 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

It is easy to check that these matrices, together with the constant matrix J, span the space of all 3×3 magic squares. Thus A is a combination of the matrices $v_\alpha w_\beta^*$ for α, β in cases (1) and (2) if and only if A is a blow-up of a 3×3 magic square.

Continuing the analysis, suppose the $m \times n$ matrix A is a (nonzero) combination of matrices from case (3). Copying the previous argument, let u_α be the vector in 12 dimensions generated by the 12th root of unity α. We see that A must be a blow-up of a 12×12 matrix that is a combination of the matrices $u_\alpha u_\beta^*$. A matrix D is said to be a "signed-back-circulant" matrix if each row of D is obtained from the previous row by shifting cyclically one step to the left and changing sign.

LEMMA 2.4. *Let D be a 12×12 matrix that is a linear combination of the matrices $u_\alpha u_\beta^*$ where α, β are as in case* (3). *Then D is a signed-back-circulant matrix of the following type for some scalars a, b, c, d*:

$$D = D(a,b,c,d)$$
$$= \begin{pmatrix} a & b & a+c & b+d & c & d & -a & -b & -a-c & -b-d & -c & -d \\ -b & -a-c & -b-d & -c & -d & a & b & a+c & b+d & c & d & -a \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Proof. Let σ be a primitive 12th root of unity. Then $\sigma^4 - \sigma^2 + 1 = 0$, and $\sigma^6 = -1$ so that $\bar{\sigma}^5 = -\sigma$. Let $B = u_\sigma u_{\sigma^5}^*$. Then

$$B = \begin{pmatrix} 1 & -\sigma & \sigma^2 & -\sigma^3 & \sigma^4 & -\sigma^5 & -1 & \sigma & -\sigma^2 & \sigma^3 & -\sigma^4 & \sigma^5 \\ \sigma & -\sigma^2 & \sigma^3 & -\sigma^4 & \sigma^5 & 1 & -\sigma & \sigma^2 & -\sigma^3 & \sigma^4 & -\sigma^5 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Each row is obtained from the one above by multiplying by σ. It is also a signed-back-circulant matrix.

Now σ can be replaced by the other primitive 12th roots of unity, namely by $\sigma^7 = -\sigma$, $\sigma^{11} = \sigma^{-1} = \bar{\sigma}$ and $\sigma^5 = -\bar{\sigma}$. The corresponding matrices are $B^{\pm}$, $\overline{B}$ and $\overline{B}^{\pm}$, where $B^{\pm}$ is obtained from B by a checkerboard sign change. These four matrices span a four-dimensional space consisting of matrices with the replication property.

A simpler basis for this space is $\{D_1, D_2, D_3, D_4\}$ where these D_i are signed-back-circulant matrices, D_1 has first row $(1,0,1,0,0,0,-1,0,-1,0,0,0)$, and D_2, D_3, and D_4 are obtained by successively shifting columns one step to the right. Here for instance, D_3 is a scalar multiple of $B + B^{\pm} - \overline{B} - \overline{B}^{\pm}$, while D_1 is a scalar multiple of $B + B^{\pm} + \overline{B} + \overline{B}^{\pm} + 4D_3$. A typical member of this space of matrices is $aD_1 + bD_2 + cD_3 + dD_4$, which is the matrix indicated in Lemma 2.4.

Note that the matrix $D(a,b,c,d)$ is a magic square (the row and column sums are all zero). The cases (1), (2), and (3) of Proposition 2.3 can be combined (since the values of r and s are compatible) to form a 9-parameter family of 12×12 magic squares satisfying the replication property. We leave to the reader the job of writing down some of these squares explicitly and noticing how they are built from 3×3 blocks.

In the case (4) the matrix turns out to be a "signed-circulant." That is, each row is obtained from the previous row by shifting cyclically one step to the *right* and changing sign.

LEMMA 2.5. *Let E be a 12×12 matrix that is a linear combination of the matrices $u_\alpha u_\beta^*$ where α, β are as in case* (4). *Then E is a signed-circulant matrix of the following type for some scalars a, b, c, d*:

$$E = E(a,b,c,d)$$
$$= \begin{pmatrix} a & b & a+c & b+d & c & d & -a & -b & -a-c & -b-d & -c & -d \\ d & -a & -b & -a-c & -b-d & -c & -d & a & b & a+c & b+d & c \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Proof. The argument is the same as in Lemma 2.4, except that we start from the matrix $C = u_\sigma u_{\sigma^{-5}}^*$ in place of B.

As in the previous situation, the cases (1), (2), and (4) can be combined to form a different 9-parameter family of magic squares satisfying the replication property. The following theorem is a summary of the preceding discussion.

THEOREM 2.6. *If A is any $m \times n$ matrix with the replication property, then one of the following holds*:

(1) *A is a blow-up of a 1×1 square,*

(2) *A is a blow-up of a 3×3 magic square,*

(3) *A is a blow-up of a 12×12 magic square in one of the 9-parameter families listed above.*

3. Generalizations One variation of the problem is to introduce a scale factor. For example let

$$M = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \\ 5 & 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 & 1 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix} \quad \text{and} \quad N = \begin{pmatrix} 1 & 2 & 5 & 7 & 6 & 3 \\ 3 & 1 & 2 & 5 & 7 & 6 \\ 6 & 3 & 1 & 2 & 5 & 7 \\ 7 & 6 & 3 & 1 & 2 & 5 \\ 5 & 7 & 6 & 3 & 1 & 2 \\ 2 & 5 & 7 & 6 & 3 & 1 \end{pmatrix}.$$

The derived matrix of M decreased by 15 is a shifted version of $-M$. The derived matrix of N decreased by 4 equals $3N$.

Generally if the derived matrix of A equals a constant plus a shifted version of λA we say that A has the replication property with scale factor λ. If a non-constant matrix A has this property for some rational λ then one can show that $\lambda = 1, -1, 2, -2, 3$, or -3. As for the case $\lambda = 1$, all such matrices A can be listed. Theorem 2 of Newman's paper [3] is relevant here.

Another variation of our problem is to generalize the definition of the "derived matrix." The case analyzed above uses the 2×2 pattern $\begin{pmatrix} * & * \\ * & * \end{pmatrix}$. For any pattern $\mathscr{P}$ and any matrix A we can define the "$\mathscr{P}$-derived matrix" $A(\mathscr{P})$ and consider the "$\mathscr{P}$-replication property." For some small patterns complete answers can be found.

For example let $\mathscr{L} = (* \quad *)$ and $A = \begin{pmatrix} 1 & 5 & 6 \\ 7 & 3 & 2 \end{pmatrix}$. Then the $\mathscr{L}$-derived matrix is $A(\mathscr{L}) = \begin{pmatrix} 6 & 11 & 7 \\ 10 & 5 & 9 \end{pmatrix}$. This A does satisfy the $\mathscr{L}$-replication property, since $A(\mathscr{L}) - 4J$ is a shift of the original matrix A. Similarly if $\mathscr{T} = \begin{pmatrix} * & * \\ * & \end{pmatrix}$, then the matrix

$$A = \begin{pmatrix} 5 & 2 & 1 & 4 \\ 4 & 5 & 2 & 1 \\ 1 & 4 & 5 & 2 \\ 2 & 1 & 4 & 5 \end{pmatrix}$$

satisfies the $\mathscr{T}$-replication property.

Appendix. Proof of Proposition 2.2 The Proposition 2.2 is a somewhat technical result proved using Galois theory. Much of the work was done by M. Newman [3].

THEOREM. *Suppose x and y are rational numbers with $0 \leqslant x \leqslant y \leqslant \frac{1}{2}$ satisfying the equation $sin(\pi x)sin(\pi y) = \frac{1}{4}$. Then (x, y) is one of the pairs $(\frac{1}{6}, \frac{1}{6})$, $(\frac{1}{10}, \frac{3}{10})$, $(\frac{1}{12}, \frac{5}{12})$.*

This beautiful result is Theorem 1 of Newman's article [3]. Newman's argument uses the inequality $|\sin(x)| \leqslant |x|$ and some results from the Galois theory of cyclotomic fields. We will not reproduce the proof here.

Proof of Proposition 2.2. We are given roots of unity α, β such that $(1+\alpha)(1+\bar{\beta})$ is itself a root of unity. Choose u, v with $\alpha = -u^2$ and $\bar{\beta} = -v^2$. Then u and v are roots of unity and we may assume they both lie in the first quadrant. (To see this, note that we may replace u by $-u$ and, allowing α to be replaced by α^{-1}, we may also replace u by u^{-1}. A combination of such replacements will put u in the first quadrant. We may normalize v similarly.)

It follows from the hypothesis that $(u-u^{-1})(v-v^{-1})$ is a real root of unity, which is negative since u and v are in the first quadrant. Hence $(u-u^{-1})(v-v^{-1}) = -1$. Expressing $u = e^{i\pi x}$ and $v = e^{i\pi y}$ we conclude that $\sin(\pi x)\sin(\pi y) = \frac{1}{4}$. The numbers x and y here are rational numbers between 0 and $\frac{1}{2}$. Interchanging α, β if necessary we may assume $0 \leqslant x \leqslant y \leqslant \frac{1}{2}$, and Newman's Theorem tells us all the possibilities for (x, y).

If $x = y = \frac{1}{6}$ then α, β are primitive 3rd roots of unity with $\bar{\beta} = \alpha$. Then $\alpha^2 + \alpha + 1 = 0$ and $(1+\alpha)(1+\bar{\beta}) = (1+\alpha)^2 = \alpha$, and the equation in Proposition 2.2 is satisfied for suitable r, s. Recalling that α and β could have been replaced by their inverses, we find that α and β can be either of the primitive 3rd roots of unity.

If $x = \frac{1}{10}$ and $y = \frac{3}{10}$ then α, β are primitive 5th roots of unity with $\bar{\beta} = \alpha^3$. Since $\alpha^4 + \alpha^3 + \alpha^2 + \alpha + 1 = 0$ we find that $(1+\alpha)(1+\bar{\beta}) = (1+\alpha)(1+\alpha^3) = -\alpha^2$. However $-\alpha^2$ cannot equal $\alpha^r \bar{\beta}^s = \alpha^{r+3s}$ since -1 is not an integral power of α. Hence this case is eliminated.

If $x = \frac{1}{12}$ and $y = \frac{5}{12}$, then α, β are primitive 12th roots of unity with $\bar{\beta} = \alpha^5$. The calculation for $(1+\alpha)(1+\bar{\beta})$ was done in the proof of Proposition 2.3. Since α and β could have been replaced by their inverses, we find that α is any of the four primitive 12th roots of unity and $\beta = \alpha^5$ or α^{-5}.

The Proposition 2.2 can also be deduced from the more advanced work of Conway and Jones **[1]**.

REFERENCES

1. J. H. Conway and A. J. Jones, Trigonometric diophantine equations (On vanishing sums of roots of unity), *Acta Arith.* 30 (1976), 229–240.
2. M. Gardner, *Penrose Tiles to Trapdoor Ciphers*, W. H. Freeman, New York, 1989.
3. M. Newman, Some results on roots of unity, with an application to a diophantine problem, *Aeq. Math.* 2 (1969), 163–166.

For Every Answer There Are Two Questions

A. M. FINK
Iowa State University
Ames, IA 50011

JUAN A. GATICA[1]
University of Iowa
Iowa City, IA 52242

Duality is a concept that dominates modern optimization theory. The most familiar one is that associated with linear programming. Here we look at a different one. Most students are introduced to optimization in a first course in calculus. We propose to show that the idea and use of duality can be introduced in such a course.

In a general way, duality asserts the existence of two solutions of extrema problems that are intimately related. Ideally, the two problems should use the same data; one problem should have a solution if and only if the other one does; and the solution of one should give information about the solution of the other.

Consider by way of example two problems that are in every calculus text.

I. Find the rectangle of maximum area when the perimeter is fixed.

II. Find the rectangle of minimum perimeter when the area is fixed.

An alert calculus student will observe that the extremal rectangle in both problems is a square. Is this an accident?

One way to begin is to observe that "Every time you solve an optimization problem you have proved an inequality." This is hardly a deep statement, but it is a useful principle. For example, if one has solved Problem I and discovered that the answer is a square then one can formulate a useful inequality. For let A be the area of any rectangle of perimeter P. Then $A \leqslant A_1$ where A_1 is the area of the square of perimeter P. But $A_1 = (P/4)^2$ so that

$$A < P^2/16 \text{ unless the rectangle is a square, when equality holds.} \tag{1}$$

When (1) is rewritten as

$$4\sqrt{A} < P \text{ unless the rectangle is a square, when equality holds,} \tag{2}$$

then we also have a solution of problem II. That is, problems I and II have the same answer because the inequality (1) solves both of them simultaneously. They may be called dual problems. The procedure shows the value of solving problem 1 with P as a parameter.

This idea is introduced in Niven [1] with the statement that it (duality) is true for geometric problems and will be used (without a proof being given). See also Fink [2]. One might ask what the principle is and how widely it is applicable.

The simple problem above serves as a special case of the general principle. Let X be a set and let f and g be real-valued functions defined on X. Suppose that there is a subset $Y \subset X$ such that

i) $f|Y$ and $g|Y$ are onto $\mathbb{R}^+ = (0, \infty)$
ii) for $y_1, y_2 \in Y$, $f(y_1) < f(y_2)$ if and only if $g(y_1) < g(y_2)$.

[1]This research was partially supported by AFSOR87-1040.

We then consider the two statements (which may be true or false):

A) For every $c>0$, $\max_{x \in X} f(x)$ under the constraint $g(x)=c$ is attained at a $y_0 \in Y$ (with $g(y_0)=c$).
B) For every $d>0$, $\min_{x \in X} g(x)$ under the constraint $f(x)=d$ is attained at a $y_0 \in Y$ (with $f(y_0)=d$).

If $X=\{\text{rectangles}\}$, $Y=\{\text{squares}\}$, f is the area function and g is the perimeter function then we have the above situation and both A and B are true.

THEOREM. (*Duality Principle.*) *With X, Y, f, g as above, satisfying i) and ii), then A holds if and only if B holds.*

As a preliminary to the proof of the theorem we observe that $g \circ (f|Y)^{-1}: \mathbb{R}^+ \to \mathbb{R}^+$ (and also $f \circ (g|Y)^{-1}: \mathbb{R}^+ \to \mathbb{R}^+$) is a strictly increasing function. To see this, consider $g \circ (f|Y)^{-1}$ and suppose that $c \in \mathbb{R}^+$ and $y_1, y_2 \in Y$, with $f(y_1)=f(y_2)=c$. Then we must have that $g(y_1)=g(y_2)$ since otherwise we would contradict ii); thus $g \circ (f|Y)^{-1}$ is well defined. For ease of notation we will write $g \circ f^{-1}$ for $g \circ (f|Y)^{-1}$.

If $c_1, c_2 \in \mathbb{R}^+$, $c_1<c_2$ and $y_1 \in f^{-1}(c_1), y_2 \in f^{-1}(c_2)$ then, since $f(y_1)<f(y_2)$ we must have $g(y_1)<g(y_2)$ by ii) so that $g \circ f^{-1}(c_1)<g \circ f^{-1}(c_2)$, and $g \circ f^{-1}$ is increasing.

On the other hand if we assume that $g \circ f^{-1}: \mathbb{R}^+ \to \mathbb{R}^+$ is well defined and strictly increasing, then ii) must hold. For if $y_1, y_2 \in Y$ with $f(y_1)<f(y_2)$, then it must be the case that $g \circ f^{-1}(f(y_1))<g \circ f^{-1}(f(y_2))$ and hence, since the function is well defined, we must have $g(y_1)<g(y_2)$.

Conversely, if $g(y_1)<g(y_2)$ then it cannot be the case that $f(y_1) \geqslant f(y_2)$, since if it were then $g \circ f^{-1}(f(y_2)) \geqslant g \circ f^{-1}(f(y_2))$ and $g(y_1) \geqslant g(y_2)$, a contradiction.

The conclusion we have reached is that ii) is equivalent to

ii)′ $g \circ f^{-1}$ is well defined and increasing on $\mathbb{R}^+$.

By a similar argument one can show that ii) is also equivalent to

ii)″ $f \circ (g|Y)^{-1}$ is well defined and increasing on $\mathbb{R}^+$.

Proof of the Theorem. Suppose that A holds. We use the principle that we should write the inequality that is proved. It is: For all $x \in X$ such that $g(x)=c, f(x) \leqslant f(y_0)$ for some $y_0 \in Y$ such that $g(y_0)=c$. Thus

$$f(x) \leqslant f \circ g^{-1}(c) \quad \text{when } g(x)=c. \tag{3}$$

Now let d be given and consider the problem of minimizing $g(x)$ when $f(x)=d$. If $x \in X$ with $f(x)=d$, set $c=g(x)$ and we have (3) and $d \leqslant f(g^{-1}(c))$. Since $g \circ f^{-1}$ is increasing we have

$$g \circ f^{-1}(d) \leqslant c.$$

But there is a $y_1 \in Y$ such that $f(y_1)=d$, and with this substitution we have

$$g(y_1) \leqslant c=g(x) \text{ for all } x \text{ such that } f(x)=d, \tag{4}$$

which proves that B holds. The reverse implication is proved in the same way.

The key to the argument is the inequality (3) that corresponds to (1) for the special problem.

It should be noted that this theorem gives a strong version of duality. At the risk of being redundant we can say this in several ways; (a) the "answers" are the same in both cases, namely a $y_0 \in Y$; (b) the solution to both problems can be read off a single inequality (3); and (c) the existence of an extremum in Y for one problem implies the existence of an extremum in Y for the other problem.

The last statement (c) is a strong form of duality. It gives the existence of an extremum without invoking any topology. For example, a student in calculus would solve problem I by writing

$$A(x)=xy=x(P/2-x), \qquad 0\leqslant x\leqslant P/2$$

and find the solution by using the existence of the maximum on a closed interval. But for problem II, eliminating a variable leads to

$$P(x)=2x+\frac{2A}{x}, \qquad 0<x<\infty,$$

and one needs a more sophisticated argument (to a first-year student) to prove the existence of a minimum. The duality principle gives the existence directly.

It should be noted that if $f|Y$ and $g|Y$ are one-to-one then we have uniqueness in both statements.

By way of giving more examples we choose some that might be of interest.

EXAMPLE 1. Minimize x^2+y^2 subject to $x^2y=1$ and $x>0$, $y>0$. Since the constraint set is not compact we choose instead to maximize x^2y subject to $x^2+y^2=d$ and $x\geqslant 0$, $y\geqslant 0$. Calculus shows that the maximizing point of $y(d-y^2)$, $0\leqslant y\leqslant\sqrt{d}$ is at $y^2=d/3$. Here we have $X=R^2_+$, $Y=\{(x,y)\colon x=\sqrt{2}\,y\}$, $f(x,y)=x^2y$ and $g(x,y)=x^2+y^2$. Then $f|_Y=2y^3$, $g|_Y=3y^2$ and $(f\circ g|_Y)^{-1}(c)=2(c/3)^{3/2}$ and we can apply the duality principle. The original minimization problem has a solution with the minimizing point satisfying $x=\sqrt{2}\,y$. It follows that $y=2^{-1/3}$ and $x=2^{1/6}$ and the minimum value is $3(2)^{-2/3}$.

As an alternate way of thinking about the problem, we may let $D=x^2y$ and write the solution of the maximizing problem in Example 1 as

$$D<2\left(\frac{d}{3}\right)^{3/2} \quad \text{unless } x=\sqrt{2}\,y. \tag{5}$$

Rewriting this as

$$3\left(\frac{D}{2}\right)^{2/3}<d \quad \text{unless } x=\sqrt{2}\,y, \tag{6}$$

we solve the original problem by taking $d=1$ in (6). This way of thinking about the problem highlights the inequality (3) instead of the theorem.

EXAMPLE 2. Let $a>0$ be given, $X=\{\text{rectangles}\}$ and f the area function and g be the function $2al+2w$ with l being the "length" and w the "width". We can interpret g as a cost function when two kinds of fence are used for fencing a rectangle.

Consider the minimum cost problem with the area given. Again the constraint $xy=A$ is noncompact so we choose to solve instead the problem of maximizing the area subject to the cost being a constant C. One finds that the maximizing point lies in

$$Y=\{\text{rectangles whose width is } a \text{ times the length}\}$$

so that $f|_Y=al^2$ and $g|_Y=4al$. The inequality (3) is

$$A\leqslant C^2/16a$$

from which one can solve the original problem. We have obtained a rule of thumb for the prospective fence buyer: In order to minimize cost of fencing a rectangle with given area, spend half of your money on each kind of fence.

EXAMPLE 3. Let $X=$ {right circular cylinders} and let V and S be the volume and surface functions. The easier of the two problems to solve is to maximize volume when the surface area is given. One finds the answer lies in $Y=$ {square cylinders} = {right circular cylinders with diameter equal to the height}. Then $V|_Y=2\pi r^3$ and $S|_Y=6\pi r^2$ for r the radius and $V\circ S^{-1}(c)=c^{3/2}/3\sqrt{6\pi}$. Thus the problem of minimizing the surface area with volume fixed is also solved by a square cylinder. The inequality (3) in this case is

$$V\leqslant\frac{1}{3\sqrt{6\pi}}S^{3/2}.$$

EXAMPLE 4. Let $X=R^n_+$, $f(x)=[\prod_{i=1}^n x_i]^{1/n}$, and $g(x)=(1/n)\sum_{i=1}^n x_i$. Again the problem of maximizing f subject to $g=c$, $x_i\geqslant 0$ has a solution in $Y=\{x|x_1=x_2=\cdots=x_n\}=$ the "diagonal in X". In this case $f|Y=g|Y=f\circ g^{-1}=$ identity and (3) is the arithmetic-geometric mean inequality.

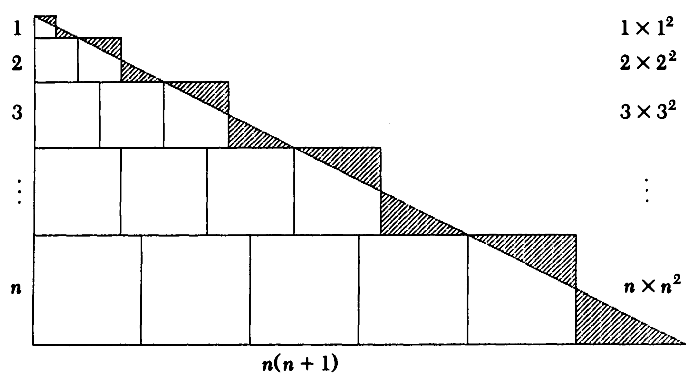
REFERENCES

1. I. Niven, *Maxima and Minima without Calculus*, MAA, Washington, DC, 1981.
2. A. M. Fink, Max-Min without calculus, *The Math Log*, XV 3 (1971).

Proofs without Words

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

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



—GEORG SCHRAGE
UNIVERSITY OF DORTMUND
4600 DORTMUND, GERMANY

the prospective fence buyer: In order to minimize cost of fencing a rectangle with given area, spend half of your money on each kind of fence.

EXAMPLE 3. Let $X = \{\text{right circular cylinders}\}$ and let V and S be the volume and surface functions. The easier of the two problems to solve is to maximize volume when the surface area is given. One finds the answer lies in $Y = \{\text{square cylinders}\} = \{\text{right circular cylinders with diameter equal to the height}\}$. Then $V|_Y = 2\pi r^3$ and $S|_Y = 6\pi r^2$ for r the radius and $V \circ S^{-1}(c) = c^{3/2}/3\sqrt{6\pi}$. Thus the problem of minimizing the surface area with volume fixed is also solved by a square cylinder. The inequality (3) in this case is

$$V \leqslant \frac{1}{3\sqrt{6\pi}} S^{3/2}.$$

EXAMPLE 4. Let $X = R_+^n$, $f(x) = [\prod_{i=1}^n x_i]^{1/n}$, and $g(x) = (1/n)\sum_{i=1}^n x_i$. Again the problem of maximizing f subject to $g = c$, $x_i \geqslant 0$ has a solution in $Y = \{x | x_1 = x_2 = \cdots = x_n\}$ = the "diagonal in X". In this case $f|Y = g|Y = f \circ g^{-1}$ = identity and (3) is the arithmetic-geometric mean inequality.

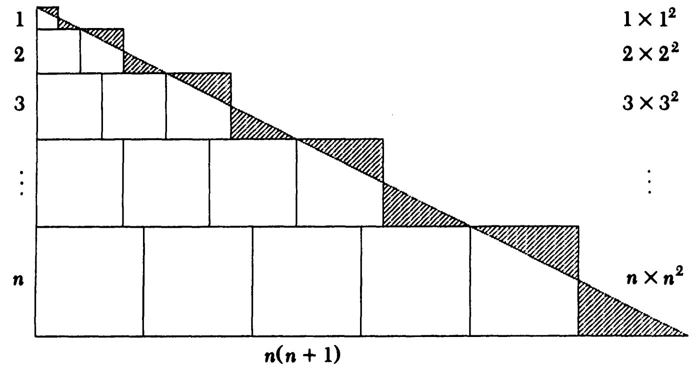
REFERENCES

1. I. Niven, *Maxima and Minima without Calculus*, MAA, Washington, DC, 1981.
2. A. M. Fink, Max-Min without calculus, *The Math Log*, XV 3 (1971).

Proofs without Words

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

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



—GEORG SCHRAGE
UNIVERSITY OF DORTMUND
4600 DORTMUND, GERMANY

When is the Product of Two Derivatives a Derivative?

MICHAEL W. BOTSKO
Saint Vincent College
Latrobe, PA 15650

Not always. Certainly the set of all derivatives on a closed interval $[a, b]$ is a linear space that contains the continuous functions, but this set is not closed under multiplication. In fact, the square of a derivative is not always a derivative as we will see in the following example. The main purpose of this note, however, is to give a very basic proof of a theorem by R. J. Fleissner that provides sufficient conditions for the product of two derivatives to be a derivative.

Instead of actually giving an example of a derivative whose square is not a derivative we will proceed indirectly. Let us assume that the square of a derivative always is a derivative. To contradict this statement define f and g on [0, 1] such that

$$f(x)=\begin{cases}\sin(1/x) & \text{if } x>0\\ 0 & \text{if } x=0\end{cases}, \quad\text{and}\quad g(x)=\begin{cases}\cos(1/x) & \text{if } x>0\\ 0 & \text{if } x=0.\end{cases}$$

Both f and g are derivatives on [0, 1]. To see that f is a derivative define F and h so that

$$F(x)=\begin{cases}x^2\cos(1/x) & \text{if } x>0\\ 0 & \text{if } x=0\end{cases}, \quad\text{and}\quad h(x)=\begin{cases}2x\cos(1/x) & \text{if } x>0\\ 0 & \text{if } x=0.\end{cases}$$

Clearly $F'(x)=h(x)+f(x)$ on [0, 1]. Since h is continuous, it is a derivative and thus f is a derivative. Similarly g is a derivative on [0, 1]. Now by our assumption f^2+g^2 is a derivative on [0, 1]. But this is impossible since f^2+g^2 has the value of 1 if $x>0$ and the value of 0 if $x=0$ and hence does not satisfy the intermediate value property. Thus the product of two derivatives need not be a derivative.

The question now arises as to when the product of two derivatives is a derivative. This problem was studied as far back as 1911 when W. H. Young **[6]** gave the following tentative answer:

THEOREM 1. *If f has a bounded derivative on $[a, b]$ (and hence is a derivative) and g is a derivative on $[a, b]$, then fg is a derivative on $[a, b]$.*

More recently J. Foran [3] improved Young's result by establishing the following theorem. See [4, p. 104] for a review of the definition of *absolute continuity*.

THEOREM 2. *If f is absolutely continuous on $[a, b]$ and g is a derivative on $[a, b]$, then fg is a derivative on $[a, b]$.*

Finally in 1975, R. J. Fleissner **[2]** established the following result. Recall that f is of *bounded variation* on $[a, b]$ if there exists a positive number M such that $\Sigma|f(x_k)-f(x_{k-1})|\leqslant M$ for all partitions $P=\{x_0, x_1, \dots, x_n\}$ of $[a, b]$.

THEOREM 3. *If f is continuous and of bounded variation on $[a, b]$ and g is a derivative, then fg is a derivative on $[a, b]$.*

Since a function with a bounded derivative is absolutely continuous and since an absolutely continuous function is continuous and of bounded variation, Theorem 3 implies Theorem 2 that, in turn, implies Theorem 1. Also it is of interest to note that neither f nor g of our example is of bounded variation.

Fleissner's proof involves the Denjoy integral [5, pp. 241–246] and is not easily accessible to undergraduates. To remedy that situation, we present a simple proof that is based on the concept of the Riemann-Stieltjes integral and hence within reach of students in an Advanced Calculus course.

Our proof of Theorem 3. By Theorem 8-15 in [1, p. 169] any continuous function of bounded variation can be written as the difference of two continuous increasing functions. Thus we may assume without loss of generality that f is continuous and increasing. Let G denote a primitive of g. We propose to show that

$$T(x) = G(x)f(x) - \int_a^x G\,df$$

is a primitive of fg. The integral on the right is, of course, the Riemann-Stieltjes integral that exists because G is continuous and f is increasing. We apply the mean-value theorem for Riemann-Stieltjes integrals to obtain

$$\begin{aligned} T(x+h) - T(x) &= G(x+h)f(x+h) - G(x)f(x) - \int_x^{x+h} G\,df \\ &= G(x+h)f(x+h) - G(x)f(x) - G(y)[f(x+h) - f(x)] \end{aligned}$$

where y is between x and $x+h$. Some elementary manipulations yield for the above expression

$$f(x+h)[G(x+h) - G(x)] - [f(x+h) - f(x)][G(y) - G(x)]$$

and hence

$$\begin{aligned} (1/h)[T(x+h) - T(x)] &= f(x+h)[G(x+h) - G(x)]/h - [f(x+h) - f(x)] \\ &\quad \times \{[G(y) - G(x)]/(y-x)\}\{(y-x)/h\}. \end{aligned}$$

We now take the limit as $h \to 0$ and obtain

$$\begin{aligned} T'(x) = f(x)G'(x) - \lim_{h \to 0} &[f(x+h) - f(x)] \\ &\times \{[G(y) - G(x)]/(y-x)\}\{(y-x)/h\}. \end{aligned}$$

Since $G'(x) = g(x)$, f is continuous at x, $|(y-x)/h| \leqslant 1$, and

$$\lim_{h \to 0} [G(y) - G(x)]/(y-x) = G'(x),$$

we obtain the desired result

$$T'(x) = f(x)g(x).$$

REFERENCES

1. T. M. Apostol, *Mathematical Analysis*, Addison-Wesley, Reading, MA, 1960.
2. R. J. Fleissner, On the product of derivatives, *Fund. Math.* 88 (1975) 173–178.
3. J. Foran, On the product of derivatives, *Fund. Math.* 80 (1973), 293–294.
4. H. L. Royden, *Real Analysis*, 2nd edition, Macmillan Publishing Co., New York, 1968.
5. S. Saks, *Theory of the Integral*, 2nd edition, Monografie Matematyczne, Warszawa-Lwow, 1937.
6. W. H. Young, A note on the property of being a differential coefficient, *Proc. London Math. Soc.* 9 (1911), 360–368.

Data Exchange and Permutation Length

LAWRENCE FIALKOW
SUNY, College at New Paltz
New Paltz, NY 12561

HECTOR SALAS
University of Puerto Rico
Mayaguez, P.R. 00708

A familiar theme in algorithms texts [**1**, **2**, **3**, **6**, **7**, **8**] is the estimation of "worst-case" runtimes for various sort routines. A more sophisticated issue, typically involving nontrivial counting arguments, is the calculation of "average" runtimes for these algorithms. For the well-known "exchange" sort (or "bubble" sort), the texts develop the $O(n^2)$ worst-case runtime by considering the worst-case behavior during each successive "pass" of the routine over appropriate sublists of the input data list. D. Knuth [**6**] also develops the $O(n^2)$ average runtime, but this cannot be based on a pass-by-pass average analysis. Perhaps for this reason, the texts neglect the average number of data exchanges occurring during a *single pass* of the sort. Our aim is to show that, nevertheless, this quantity has an interesting combinatorial interpretation as a measure of the average "length" of a permutation.

Let $x_1, \dots, x_n$ denote an "input list" of n distinct real numbers. Exchange-sort transforms this list into a list sorted in increasing order via the following program: During the first pass of the sort, as j runs from 1 to $n-1$, we compare x_j to x_{j+1} and "exchange" these data items if $x_j > x_{j+1}$; thus the largest list value "bubbles" up to the n-th list position. (For example, 5, 2, 4, 6, 3, 1 is transformed into 2, 4, 5, 3, 1, 6.) During a subsequent pass $i, i > 1$, we perform the same operation on the sublist $x_1, \dots, x_{n-i+1}$; the program terminates after completing a pass in which no exchange occurs, or after completing $n-1$ passes.

D. Knuth [**6**] has shown that the average runtime of exchange-sort for "random" input data may be expressed in terms of combinatorial properties of permutations. Indeed, any list x input to the sort algorithm may be viewed as a permutation σ_x of $\{1, \dots, n\}$, and in the sequel we seek to relate the processing of x by exchange-sort to algebraic properties of σ_x.

Let S_n be the group of permutations of order n and let 1_n denote the identity element of S_n. Each permutation $\sigma \in S_n$ is an injective mapping of the set $\Lambda_n \equiv \{1, \dots, n\}$ onto itself, and we denote the action of σ by $(\sigma(1), \dots, \sigma(n))$. A pair of indices (i, j) is an *inversion* of the permutation $x \equiv (x_1, \dots, x_n)$ if $i < j$ and $x_i > x_j$. We define $i(x)$ to be the number of inversions of x.

Recall that an element $\tau \in S_n$ is a *transposition* if there exist distinct indices i and j such that $\tau(i) = j$, $\tau(j) = i$ and $\tau(k) = k$ whenever $k \neq i, j$; we denote τ by $\langle i, j \rangle$. A fundamental property of permutations is that each permutation can be decomposed as a (noncommutative) product of several transpositions. (We explain this decomposition in more detail in the third section.) We define the *length* of a permutation $x \in S_n$ ($x \neq 1_n$) to be the smallest positive integer m such that x can be factored as a product of m transpositions; if $x = 1_n$, we set $l(x) = 0$. Viewing x as an input sort list, $l(x)$ thus represents the minimum number of successive (possibly nonadjacent) data exchanges required to transform the list into increasing order. There are two additional parameters that may be defined as a result of applying the exchange-sort algorithm to x:

$t(x)$: the total number of exchanges applied by the sort;
$e(x)$: the total number of exchanges during the first pass of the sort.
To illustrate these parameters we consider input lists x, which are permutations of $\{1,2,3\}$:

x	$i(x)$	$t(x)$	$e(x)$	$l(x)$	minimal transpositional decomposition
(123)	0	0	0	0	
(132)	1	1	1	1	$\langle 2,3\rangle$
(213)	1	1	1	1	$\langle 1,2\rangle$
(231)	2	2	1	2	$\langle 1,3\rangle\langle 1,2\rangle$
(312)	2	2	2	2	$\langle 1,2\rangle\langle 1,3\rangle$
(321)	3	3	2	1	$\langle 1,3\rangle$
Totals	$I_3=9$	$T_3=9$	$E_3=7$	$L_3=7$	
Averages	$i_3=9/6$	$t_3=9/6$	$e_3=7/6$	$l_3=7/6$	

Knuth [6] shows that, in general, for any input permutation, $i(x)=t(x)$, and the average number of inversions (equivalently, the average number of total exchanges) for an input of length n is $n(n-1)/4$.

In this paper, we examine the values $e(x)$ and $l(x)$. As we have seen in the above example, these values may differ for an individual x; however, we will show that the average number of exchanges for a list of length n during the *first* pass of exchange-sort is equal to the average transpositional length of a permutation of order n, and that this common average value is $n-(1+1/2+\cdots+1/n)$.

Acknowledgment. In presenting the results of this paper, we do not attest to their originality. Owing to the vastness of the literature on permutations, it would not be surprising if our results appear elsewhere, perhaps in a different context. On the other hand, we have not been able to locate any references for these results, and in any case our motivation is to display to a general audience the attractive interplay between computer science and mathematics that may arise from even the simplest of considerations. In this regard, we thank the referees for numerous helpful suggestions.

Average data exchanges For an input list x of size n, $e(x)$ denotes the number of data exchanges performed by exchange-sort during the first pass. In this section we derive a formula for the expected value of $e(x)$. Since exchanges depend only on the relative order of the list elements, and not on particular numerical values, we regard an input list as a member of S_n. Since card $(S_n)=n!$, our aim is thus to compute

$$e_n \equiv (1/n!) \sum_{\sigma\in S_n} e(\sigma),$$

the average number of exchanges during the first pass for input lists of size n. As a notational convenience, we also define e_0 to be 0.

In analyzing the first-pass behavior of exchange-sort, it is particularly helpful to focus on the location of n within the input list $\sigma\in S_n$. Indeed, if $\sigma(i)=n$, let $\gamma_\sigma=(\sigma(1),\dots,\sigma(i-1))$. During the first pass, $e(\gamma_\sigma)$ exchanges occur as the routine scans γ_σ, then the resulting value in position $i-1$ is *not* exchanged with n in position i, and finally $n-i$ exchanges occur as the value n is pushed to the rightmost list position; thus $e(\sigma)=e(\gamma_\sigma)+n-i$. This observation is the basis for the following recursion formula for e_n.

THEOREM 1. *For* $n \geqslant 1$, $e_n = (n-1)/2 + (1/n)\Sigma_{i=1}^{n} e_{i-1}$.

Proof. Let $E_n = \Sigma\{e(\sigma)\colon \sigma \in S_n\}$, so that $e_n = 1/n!\,E_n$. For $1 \leqslant i \leqslant n$, let $S_{i,n} = \{\sigma \in S_n\colon \sigma(i) = n\}$; thus card $(S_{i,n}) =$ card $(S_{n-1}) = (n-1)!$. The above analysis shows that for each $\sigma \in S_{i,n}$, $e(\sigma) = e(\gamma_\sigma) + n - i$, and thus

$$E_{i,n} \equiv \sum\{e(\sigma) : \sigma \in S_{i,n}\} = \sum\{e(\gamma_\sigma) : \sigma \in S_{i,n}\} + (n-1)!(n-i). \qquad (1)$$

To analyze the first term of expression (1), let S_{i-1}^{n-1} denote the set of all ordered lists consisting of $i-1$ distinct elements chosen from the set Λ_{n-1}. For each $\tau \in S_{i-1}^{n-1}$, let $D_\tau = \{\sigma \in S_{i,n} : \gamma_\sigma = \tau\}$; for $\sigma \in D_\tau$, the first i elements are the ordered elements of τ followed by n, while the remaining $n-i$ elements can be any permutation of the elements of $\Lambda_n \backslash \tau \cup \{n\}$. Thus card $(D_\tau) = (n-i)!$, and $D_\tau \cap D_\rho = \varnothing$ for $\tau \neq \rho$. It follows that

$$\begin{aligned}\sum\{e(\gamma_\sigma) : \sigma \in S_{i,n}\} &= \sum\{(n-i)!\,e(\sigma) : \sigma \in S_{i-1}^{n-1}\}\\ &= (n-i)!\sum\{e(\sigma) : \sigma \in S_{i-1}^{n-1}\}. \qquad (2)\end{aligned}$$

Recall that exchanges depend only on the relative ordering of list elements, not on particular values. Thus, in the sum $\Sigma\{e(\sigma) : \sigma \in S_{i-1}^{n-1}\}$, the terms for the σ's that are permutations of a *fixed* subset of size $i-1$ chosen from Λ_{n-1} contribute exactly E_{i-1}. (Indeed, we may assume the fixed subset is Λ_{i-1}.) Since there are $\binom{n-1}{i-1}$ distinct subsets of Λ_{n-1} of size $i-1$, we have

$$\sum\{e(\sigma) : \sigma \in S_{i-1}^{n-1}\} = \binom{n-1}{i-1} E_{i-1}. \qquad (3)$$

Combining (1), (2), and (3),

$$E_{i,n} = (n-i)!\binom{n-1}{i-1} E_{i-1} + (n-1)!(n-i).$$

Since $E_n = \Sigma\{E_{i,n} : i = 1,\ldots,n\}$, then

$$\begin{aligned}e_n &= (1/n!)\,E_n = (1/n!)\sum\{E_{i,n} : i = 1,\ldots,n\}\\ &= (1/n!)\sum\left\{(n-i)!\binom{n-1}{i-1}E_{i-1} + (n-1)!(n-i) : i = 1,\ldots,n\right\}\\ &= (1/n!)\sum\{(n-1)!/(i-1)!\,E_{i-1} : i = 1,\ldots,n\}\\ &\quad + (1/n)\sum\{(n-i) : i = 1,\ldots,n\}\\ &= (1/n)\sum\{e_{i-1} : i = 1,\ldots,n\} + (n-1)/2.\end{aligned}$$

We next simplify the recursion formula as follows.

COROLLARY 2. *For* $n \geqslant 1$, $e_n = e_{n-1} + (n-1)/n$.

Proof.

$$\begin{aligned}e_n &= (1/n)\sum\{e_{i-1} : i = 1,\ldots,n\} + (n-1)/2\\ &= ((n-1)/n)(1/(n-1))\left[\sum\{e_{i-1} : i = 1,\ldots,n-1\} + e_{n-1}\right] + (n-1)/2\\ &= ((n-1)/n)\left[e_{n-1} - (n-2)/2 + e_{n-1}/(n-1)\right] + (n-1)/2\\ &= e_{n-1} + (n-1)/n.\end{aligned}$$

We may now obtain a closed formula for e_n:

COROLLARY 3. *For* $n \geqslant 2$, $e_n = n - (1 + 1/2 + \cdots + 1/n)$.

Proof. By successive application of Corollary 2, $e_n = (n-1)/n + (n-2)/(n-1) + \cdots + 1/2$, and the result follows.

Average permutation length Recall that $x \in S_n$ is an *m-cycle* $(2 \leqslant m \leqslant n)$ if there exists an ordered subset $A \equiv \{a_1, \dots, a_m\} \subset \wedge_n$ so that $x(a_i) = a_{i+1}$ $(1 \leqslant i \leqslant m-1)$, $x(a_m) = a_1$ and $x(j) = j$ if $j \in \wedge_n \backslash A$. We refer to this cycle as $X_A \equiv \langle a_1, \dots, a_m \rangle$. If $A = \{j\} \subset \wedge_n$, the *1-cycle* $X_A \equiv \langle j \rangle$ is defined to be a copy of the identity permutation. Cycles X_A and X_B are *disjoint* if A and B are disjoint subsets of $\wedge_n$. It is well known that each permutation may be factored as a (commutative) product of disjoint cycles [5]. By including appropriate 1-cycles, we can thus express each $x \in S_n$ as $x = X_{A_1} \dots X_{A_p}$, where $\wedge_n$ is the *disjoint union* of the A_i's. This decomposition of x into disjoint cycles is unique up to the order of the factors; we refer to this decomposition as the *canonical factorization* and we let $c(x) = p$ denote the number of cyclic factors (including the required trivial factors).

Further observe that for $m > 1$ the m-cycle X_A admits a noncommutative factorization as the product of $m-1$ 2-cycles (transpositions), namely $\langle a_1, \dots, a_m \rangle = \langle a_1, a_m \rangle \dots \langle a_1, a_3 \rangle \langle a_1, a_2 \rangle$. Thus each $x \in S_n$ may be decomposed as $x = t_1 \dots t_k$, where each t_i is a transposition, and if $x \neq 1_n$, we define the *length* of x, $l(x)$, to be the smallest integer k for which x admits such a factorization; if $x = 1_n$, we define $l(x) = 0$. Note that, in general, $l(x) \neq e(x)$. If $x = (4,2,1,3) \in S_4$, then $e(x) = 3$ and $l(x) = 2$ (since $x = \langle 1,3 \rangle \langle 1,4 \rangle$), while if $y = (2,3,4,1) \in S_4$, then $e(y) = 1$ and $l(y) = 3$ (since $y = \langle 2,1 \rangle \langle 2,4 \rangle \langle 2,3 \rangle$). Our main result shows that the *average* behavior of l is identical to that of e.

Let $l_n = 1/n! \Sigma_{x \in S_n} l(x)$, the average length of a permutation of order n.

THEOREM 4. *For* $n \geqslant 2$, $l_n = e_n = n - (1 + 1/2 + \cdots + 1/n)$.

We begin by relating permutation length to the canonical factorization.

THEOREM 5. *For* $x \in S_n$, $l(x) = n - c(x)$.

Proof. Each m-cycle $(m \geqslant 1)$ may be written as a product of $m-1$ transpositions, as noted above. Thus for $x \in S_n$, the canonical factorization of x as $c_1 \cdots c_p$, where c_i is an m_i-cycle $(1 \leqslant i \leqslant p)$, implies a factorization of x as a product of $(m_1 - 1) + \cdots + (m_p - 1)$ transpositions, whence $l(x) \leqslant m_1 + \cdots + m_p - p = n - c(x)$.

We prove the reverse inequality by induction on $l(x) \geqslant 0$. If $l(x) = 0$, x is the identity, so the canonical factorization is $x = \Pi\{\langle i \rangle : i = 1, \dots, n\}$, whence $c(x) = n$. If $l(x) = 1$, x is a transposition, with canonical factorization of the form $x = \langle a, b \rangle \Pi\{\langle j \rangle : j \neq a, b\}$, so $c(x) = n - 1$.

Let $k > 1$ and assume that $l(y) = n - c(y)$ for all $y \in S_n$ with $l(y) < k$. Suppose $l(x) = k$ and write x as a product of k transpositions, $x = \langle i_{2k-1}, i_{2k} \rangle \dots \langle i_1, i_2 \rangle$, $1 \leqslant i_1, \dots, i_{2k} \leqslant n$, $i_{2j-1} \neq i_{2j}$ $(1 \leqslant j \leqslant k)$. Let $y = \langle i_{2k-3}, i_{2k-2} \rangle \dots \langle i_1, i_2 \rangle$ (if $k = 2$, y is a transposition). Clearly $l(y) \leqslant k - 1$, and since $x = \langle i_{2k-1}, i_{2k} \rangle y$ and $l(x) = k$, it follows that $l(y) = k - 1$. Our inductive hypothesis thus implies that $c(y) = n - k + 1$.

Let the canonical factorization of y be $y = c_1 \dots c_{n-k+1}$; we claim that i_{2k-1} and i_{2k} belong to different cycles of this decomposition. For otherwise, we may assume i_{2k-1} and i_{2k} are in c_1 and we may assume c_1 is of the form $c_1 = \langle t, u, \dots, r, s, v \dots, w \rangle$, where $t = i_{2k}$ and $s = i_{2k-1}$. (In general, some of u, r, v, w may be absent.) A direct calculation shows that $c_1 = \langle t, u, \dots, r, s, v, \dots, w \rangle = \langle s, t \rangle \langle t, u, \dots, r \rangle \langle s, v, \dots, w \rangle$. Thus

$$x = \langle s,t\rangle y = \langle s,t\rangle c_1 \dots c_{n-k+1}$$
$$= \langle t,u,\dots,r\rangle\langle s,v,\dots,w\rangle c_2 \dots c_{n-k+1}$$

(since $\langle s,t\rangle\langle s,t\rangle$ is the identity). It follows that $c(x) = n-k+2$, whence $l(x) \leqslant n-(n-k+2) = k-2 = l(x)-2$, a contradiction.

Thus we may assume i_{2k-1} is in c_1, i_{2k} is in c_2 and that c_1 and c_2 are of the form

$$c_1 = \langle s,v,\dots,w\rangle \quad (s = i_{2k-1}) \quad \text{and}$$
$$c_2 = \langle t,u,\dots,r\rangle \quad (t = i_{2k}).$$

Now $x = \langle s,t\rangle c_1 c_2 \dots c_{n-k+1}$, and a calculation shows that $\langle s,t\rangle c_1 c_2$ is the cycle

$$c_0 \equiv \langle t,u,\dots,r,s,v,\dots,w\rangle.$$

Since $x = c_0 c_3 \dots c_{n-k+1}$, then $c(x) = n-k = n-l(x)$.

In the next result, we denote a permutation $p \equiv (p_1,\dots,p_n)$ by

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}.$$

LEMMA 6. *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & \cdots & i & \cdots & n \\ j_1 & j_2 & j_3 & \cdots & 1 & \cdots & j_n \end{pmatrix} \in S_n, \qquad i \neq 1, j_1 \neq 1,$$

and let

$$\tilde{A} = \begin{pmatrix} 2 & 3 & \cdots & i & \cdots & n \\ j_2 & j_3 & \cdots & j_1 & \cdots & j_n \end{pmatrix} \in S_{n-1}.$$

Then $l(A) = l(\tilde{A}) + 1$.

Proof. In the canonical factorization of A, consider the (nontrivial) cycle containing i:

$$\begin{pmatrix} 1 & j_1 & \cdots & b & i \\ j_1 & a & \cdots & i & 1 \end{pmatrix}.$$

In the canonical factorization of $\tilde{A}$, the cycle that contains i is thus

$$\begin{pmatrix} j_1 & \cdots & b & i \\ a & \cdots & i & j_1 \end{pmatrix}.$$

The other factors in the canonical decompositions of A and $\tilde{A}$ match identically, and thus $c(A) = c(\tilde{A})$. Theorem 5 thus implies

$$l(\tilde{A}) = (n-1) - c(\tilde{A}) = n-1-c(A)$$
$$= (n-1) - (n-l(A)) = l(A) - 1.$$

Remark. If

$$A = \begin{pmatrix} 1 & 2 & \cdots & n \\ 1 & j_2 & \cdots & j_n \end{pmatrix} \quad \text{and} \quad \tilde{A} = \begin{pmatrix} 2 & \cdots & n \\ j_2 & \cdots & j_n \end{pmatrix},$$

then A has 1 more cycle than $\tilde{A}$ and thus $l(A) = l(\tilde{A})$.

For $n > 1$, let $L_n = \Sigma_{x \in S_n} l(x)$, and set $L_1 = 0$.

LEMMA 7. *If* $n > 1$, $L_n = nL_{n-1} + (n-1)(n-1)!$.

Proof. We have

$$L_n = \sum_{j=1}^{n} \sum \{l(A): A \in S_n, A(1) = j\}.$$

Now, by the remark, $\Sigma\{l(A): A \in S_n, A(1) = 1\} = \Sigma\{l(B): B \in S_{n-1}\} = L_{n-1}$, and Lemma 6 implies that for $j \neq 1$,

$$\begin{aligned}\sum\{l(A): A \in S_n, A(1) = j\} &= \sum\{l(B) + 1: B \in S_{n-1}\}\\ &= \sum\{l(B): B \in S_{n-1}\} + (n-1)! = L_{n-1} + (n-1)!.\end{aligned}$$

Thus $L_n = L_{n-1} + (n-1)(L_{n-1} + (n-1)!) = nL_{n-1} + (n-1)(n-1)!$.

We are now ready for the "algebraic" analogue of Corollary 2.

COROLLARY 8. *If* $n > 1$, $l_n = l_{n-1} + (n-1)/n$.

Proof. Divide by $n!$ in the equality of Lemma 7.

Proof of Theorem 4. In view of Corollaries 2 and 8, l_n satisfies the same recursion relation as e_n. Since $l_0 = e_0 = 0$ and $l_1 = e_1 = 0$, we must have $l_n = e_n$ for all n. This equality and Corollary 3 prove Theorem 4.

REFERENCES

1. A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley Publishing Co., Reading, MA, 1976.
2. ______, *Data Structures and Algorithms*, Addison-Wesley Publishing Co., Reading, MA, 1983.
3. S. Baase, *Computer Algorithms, Introduction to Design and Analysis*, 2nd edition, Addison-Wesley Publishing Co., Reading, MA, 1988.
4. A. K. Dewdney, *The Turin Omnibus, 61 Excursions in Computer Science*, Computer Science Press, Inc., Rockville, MD, 1989.
5. I. N. Herstein, *Topics in Algebra*, Blaisdell, New York, 1964.
6. D. Knuth, *The Art of Computer Programming*, Vol. 3, Sorting and Searching, Addison-Wesley Publishing Co., Reading, MA, 1973.
7. R. Sedgewick, *Algorithms*, Addison-Wesley Publishing Co., Reading, MA, 1983.
8. N. Wirth, *Algorithms and Data Structures*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1986.

PROBLEMS

LOREN C. LARSON, *editor*
St. Olaf College

GEORGE GILBERT, *associate editor*
Texas Christian University

Proposals

To be considered for publication, solutions should be received by November 1, 1992.

1398. *Proposed by Florin S. Pîrvănescu, Slatina, Romania.*

Let $f, g\colon \mathbf{R} \to \mathbf{R} - \{0, 1\}$ be functions such that

$$f(x+1) = \frac{g(x)}{f(x)} \quad \text{and} \quad g(x+1) = \frac{g(x)-1}{f(x)-1}$$

for every $x \in \mathbf{R}$. Prove that f and g are periodic.

1399. *Proposed by Ioan Sadoveanu, Ellensburg, Washington.*

Let a, b, c be positive numbers such that $a + b + c = 1$. Let x_0, y_0, z_0 be positive numbers and for each $n \geq 0$, let

$$x_{n+1} = ax_n + by_n + cz_n, \qquad y_{n+1} = x_n^a y_n^b z_n^c, \qquad z_{n+1} = \left(\frac{a}{x_n} + \frac{b}{y_n} + \frac{c}{z_n}\right)^{-1}.$$

Show that the sequences (x_n), (y_n), and (z_n) each converge, and that they converge to the same limit.

1400. *Proposed by Frances Barry (student) and Desmond MacHale, University College, Cork, Ireland.*

Let S be a semigroup. Suppose that n is a fixed positive integer such that $xy = y^n x^n$ for all $x, y \in S$. Prove that S is commutative.

ASSISTANT EDITORS: CLIFTON CORZAT, BRUCE HANSON, RICHARD KLEBER, KAY SMITH, *and* THEODORE VESSEY, *St. Olaf College and* MARK KRUSEMEYER, *Carleton College. We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals should be accompanied by solutions, if at all possible, and by any other information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. An asterisk (*) next to a problem number indicates that neither the proposer nor the editors supplied a solution.*

Solutions should be written in a style appropriate for Mathematics Magazine. Each solution should begin on a separate sheet containing the solver's name and full address.

Solutions and new proposals should be mailed in duplicate to Loren Larson, Department of Mathematics, St. Olaf College, 1520 St. Olaf Ave., Northfield, MN 55057-1098 or mailed electronically via fax: (507) 663-3549 or e-mail: larson@stolaf.edu.

1401. *Proposed by Edward T. H. Wang, Wilfrid Laurier University, and Wan-Di Wei, University of Waterloo, Waterloo, Canada.*

Determine the number of permutations π of $1, 2, \ldots, 2n$ with the property that $|\pi(i+1) - \pi(i)| = n$ for some i, $1 \le i \le 2n-1$. Express your answer in the form $\Sigma_{k=1}^{n} A_k$. (Problem 6 of the 1989 International Mathematical Olympiad asks for a proof that there are more permutations with the above property than those without the property.)

1402. *Proposed by Florin S. Pîrvănescu, Slatina, Romania.*

Let ABC be a given triangle, and M, N, and P be arbitrary points in the interiors of the line segments BC, CA, and AB respectively. Let lines AM, BN, and CP intersect the circumcircle of ABC in points Q, R, and S respectively. Prove that

$$\frac{AM}{MQ} + \frac{BN}{NR} + \frac{CP}{PS} \ge 9.$$

Quickies

Answers to the Quickies are on page 200.

Q791. *Proposed by Barry Cipra, Northfield, Minnesota.*

Suppose you have n coins and your opponent has $n+1$. You each toss all your coins and count the number of heads. You lose if you have fewer heads, otherwise you win (i.e., you win all ties). Assuming the coins are fair, is this game fair?

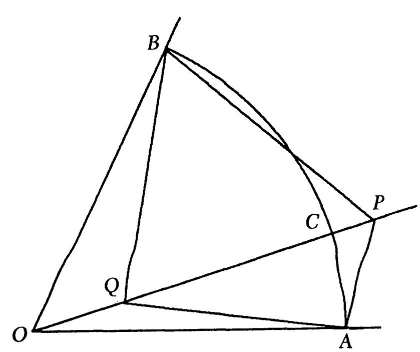
Q792. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Canada.*

Determine all positive integer triples (x, y, z) satisfying the Diophantine equation

$$x^4 + y^4 + z^4 = 2y^2z^2 + 2z^2x^2 + 2x^2y^2 - 3.$$

Q793. *Proposed by Ismor Fischer, Naval Postgraduate School, Monterey, California.*

Let AOB be a given sector of a circle, and C an arbitrary point on the arc AB. Let parallel lines through A and B intersect line OC at points P and Q respectively (see figure). Show that the area of the five-sided polygon $OAQPB$ is constant, independent of C and the parallel lines AP and BQ.



Solutions

Ratio of $\varphi(n)$ to n **June 1991**

1373. *Proposed by William P. Wardlaw, U.S. Naval Academy, Annapolis, Maryland.*

Find the set S' of all accumulation points of the set $S = \{\varphi(n)/n: n \in \mathbf{N}\}$, where φ is the Euler phi function and $\mathbf{N}$ is the set of positive integers.

Solution by John Harrington (student), University of Idaho, Moscow, Idaho.

S' is the closed interval $[0,1]$. To see this, we will make use of the following well-known results from number theory:

$$\frac{\varphi(n)}{n} = \prod_{p|n}(1-1/p), \quad \text{and} \tag{1}$$

$$\lim_{k\to\infty}\prod_{j=1}^{k}(1-1/p_j) = 0, \tag{2}$$

where p_j denotes the jth prime.

By (1), S' is contained in $[0,1]$. Let $\alpha(n) = \varphi(n)/n$. $\alpha(n)$ is a multiplicative function, and by (2), $\alpha(p_1p_2\dots p_{\underline{n}}) = \prod_{j=1}^{k}(1-1/p_j)$ is arbitrarily small for large k. This implies that $0 \in S'$. Since $\alpha(p_k) = 1 - 1/p_k$ is arbitrarily close to 1 for large k, we see that $1 \in S'$. Let $c \in (0,1)$, and let $\varepsilon > 0$ be such that $(c-\varepsilon, c) \subset (0,1)$. Choose m so large that $1/p_m < \varepsilon$ and $c < 1 - 1/p_m = \alpha(p_m)$. Let $a_n = \alpha(p_mp_{m+1}\cdots p_n)$. The set of terms in the monotone-decreasing sequence $(a_n)_{n=m}^{\infty}$ that are greater than or equal to c is finite (by (2)) and nonempty (since it contains a_m). Let a_i be the smallest member of this set. Then

$$a_i - a_{i+1} = a_i/p_{i+1} < 1/p_{i+1} < 1/p_m < \varepsilon.$$

Thus,

$$c - \varepsilon \le a_i - \varepsilon < a_{i+1} < c,$$

and this completes the proof.

Also solved by Brian D. Beasley, W. E. Briggs, David Callan, Li Chun Che (student, Hong Kong), Con Amore Problem Group (Denmark), Preston Dinkins, Mordechai Falkowitz (Israel), Kevin Ford (student), Thomas Jager, David W. Koster, Peter W. Lindstrom, Jean-Marie Monier (France), Daniel Neuenschwander (Switzerland), F. C. Rembis, John S. Sumner, and the proposer.

Callan *notes that the result is mentioned in Sierpinski's Elementary Theory of Numbers, and it is the preamble to Problem* 6070, *American Mathematical Monthly*, January 1976; October 1977.

Tessellations of n-space **June 1991**

1374. *Proposed by David Moews and Michael Reid, students, University of California, Berkeley.*

Let H be a unit n-dimensional hypercube, and A be any set of hyperfaces of H. Let H_A be the figure created by adjoining unit hypercubes at each hyperface in A. Show that, regardless of A, H_A tessellates n-space.

Solution by the proposers.

We will show that H_A tessellates n-space by translations only.

Note that the unit hypercube tessellates by translation only, with copies of the hypercube in bijection with $\mathbf{Z}^n$. After a permutation of coordinates, we may assume that A contains both hyperfaces in each of the first r directions, one of the hyperfaces in each of the next s directions (which we may assume to be the hyperface in the positive direction), and neither hyperface in the final t directions, with, of course, $r+s+t=n$.

Now the function f defined on $\mathbf{Z}^n$ by

$$f(x_1, x_2, \dots, x_n) = x_1 + 2x_2 + 3x_3 + \cdots + nx_n$$

can be considered as a function on the hypercubes in the tessellation of n-space. On H_A, this function takes the values $-r,\ -r+1,\ -r+2, \dots, 0, \dots, r+s$, a block of $m = 2r+s+1$ consecutive integers, with no repetition. Let G be the kernel of the composite map $\mathbf{Z}^n \to \mathbf{Z} \to \mathbf{Z}/m\mathbf{Z}$, where the left arrow is f, and the right arrow is projection. Then the hypercubes in H_A form a complete set of coset representatives for G in $\mathbf{Z}^n$. Therefore, if we translate H_A by elements of G, we tile n-space.

Also solved by Mark S. Anderson.

A family of product congruences **June 1991**

1375. *Proposed by Lorraine L. Foster, California State University, Northridge, California.*

Prove that for each integer $k \geq 3$ there exist positive integers $n_1, n_2, \dots, n_k$ such that $\prod_{i \neq j} n_i \equiv 1 \pmod{n_j}$, for $j = 1, 2, \dots, k$. (Note: Problem 1339, February 1990; solution, February 1991, treats the case $k = 3$.)

Solution by Centre Problem Solving Group, Centre College, Danville, Kentucky.

For a given k, set $x_1 = 2$, $x_2 = 3$, $x_j = (\prod_{i<j} x_i) + 1$ if $2 < j < k$, and $x_k = (\prod_{i<k} x_i) - 1$. Clearly, $\prod_{i<k} x_i \equiv 1 \pmod{x_k}$. Also,

$$\prod_{i \neq j} x_i = \left(\prod_{i<j} x_i\right)\left(\prod_{j<i<k} x_i\right) x_k \equiv (-1)(1)(-1) \equiv 1 \pmod{x_j}.$$

Observation by Michael Reid, University of California, Berkeley.

For each k, there are only finitely many solutions to the system of congruences

$$n_1 \dots n_{i-1} n_{i+1} \dots n_k \equiv 1 \pmod{n_i}, \quad 1 < n_1 \leq n_2 \leq \cdots \leq n_k.$$

We begin the proof by observing that for integers $1 < n_1 \leq n_2 \leq \cdots \leq n_k$,

$$n_1 \dots n_{i-1} n_{i+1} \dots n_k \equiv 1 \pmod{n_i}, \quad i = 1, 2, \dots, k$$

if and only if

$$n_2 n_3 \dots n_k + n_1 n_3 \dots n_k + n_1 n_2 n_4 \dots n_k + \cdots + n_1 n_2 \dots n_{k-1} \equiv 1 \pmod{n_1 n_2 \dots n_k}$$

and this is true if and only if

$$\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k} - \frac{1}{n_1 n_2 \dots n_k}$$

is a positive integer. We will prove there are only finitely many solutions to this in integers greater than 1.

Since

$$0 < \frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k} - \frac{1}{n_1 n_2 \ldots n_k} < \frac{k}{2},$$

it suffices to show that for fixed positive integers a, b, and k, the equation

$$\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k} - \frac{1}{b n_1 n_2 \ldots n_k} = \frac{a}{b} \qquad (*)$$

has only finitely many integral solutions greater than 1.

We prove this claim by induction on k. For $k = 1$,

$$\frac{1}{n_1} - \frac{1}{b n_1} = \frac{a}{b}$$

is equivalent to $n_1 = (b-1)/a$, so there is at most one solution. Assume the result for $k-1$. A solution to $(*)$ satisfies

$$\frac{a}{b} = \frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k} - \frac{1}{b n_1 n_2 \ldots n_k} < \frac{k}{n_1}$$

or $n_1 < kb/a$. Thus there are only finitely many possibilities for n_1. The equation $(*)$ may be rewritten as

$$\frac{1}{n_2} + \cdots + \frac{1}{n_k} - \frac{1}{(b n_1) n_2 \ldots n_k} = \frac{a n_1 - b}{b n_1}.$$

Since $a n_1 - b > 0$, the inductive hypothesis implies there are only finitely many $1 < n_2 \leq \cdots \leq n_k$ satisfying $(*)$ for a given n_1, and this completes the induction.

Also solved by David Callan, Con Amore Problem Group (Denmark), Mordechai Falkowitz (Israel), Kevin Ford (student), Russell Jay Hendel, Thomas Jager, Kevin W. Koster, Heinz-Jürgen Seiffert (Germany), John S. Sumner, and the proposer.

Hugh Edgar, *San Jose State University, points out that something needs to be done to avoid the trivial solution* $n_1 = n_2 = \cdots = n_k = 1$.

Reid *asks the following question. Suppose that* $n_1, n_2, \ldots n_k$ *is a (finite) sequence of pairwise relatively prime positive integers. Can we necessarily extend this to a sequence* $n_1, n_2, \ldots, n_l$ *for some* $l \geq k$ *such that* $\prod_{i \neq j} n_i \equiv 1 \ (\mathrm{mod}\, n_j)$, *for each* $j = 1, 2, \ldots, l$?

An application of Fermat's Little Theorem **June 1991**

1376. *Proposed by Eric Canning, student, and Marion B. Smith, California State University, Bakersfield, California.*

If p is a prime and n an integer such that $1 < n \leq p$, then $\varphi\left(\sum_{k=0}^{p-1} n^k\right) \equiv 0 \pmod{p}$, where φ is the Euler phi function.

Solution by Rich Bauer, Seattle, Washington.

We have $\sum_{k=0}^{p-1} n^k = (n^p - 1)/(n-1)$. Suppose that q is a prime dividing $(n^p - 1)/(n-1)$. Then $n^p \equiv 1 \pmod{q}$. If $n \equiv 1 \pmod{q}$, then $\sum_{k=0}^{p-1} n^k \equiv p \pmod{q}$ or $q = p$, contradicting $1 < n \leq p$. By Fermat's Little Theorem, we know that $n^{q-1} \equiv 1 \pmod{q}$. Therefore p divides $q - 1$.

For any integer $N = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$, with p_i prime and $a_i \geq 1$, we know that $\varphi(N) = p_1^{a_1 - 1}(p_1 - 1) \cdots p_r^{a_r - 1}(p_r - 1)$. With $N = ((n^p - 1)/(n-1))$ and $q = p_1$, say, it is clear that $q - 1$ divides $\varphi(N)$. But from above, p divides $q - 1$, so p divides $\varphi(N)$ and we are done.

Also solved by Seung-Jin Bang (Korea), Madelaine Bates, David Callan, Con Amore Problem Group (Denmark), David Doster, Jesse Deutsch, Hugh Edgar, F. J. Flanigan, Kevin Ford (student), Arne Fransén (Sweden), Nick S. Hekster (The Netherlands), Thomas Jager, David W. Koster, H. K. Krishnapriyan, Peter W. Lindstrom, Daniel Neuenschwander (Switzerland), Allan Pedersen (Denmark), F. C. Rembis, Sinai Robins and Matthew Isom (student), Heinz-Jürgen Seiffert (Germany), Daniel B. Shapiro, Lawrence Somer, John S. Sumner, Michael Vowe (Switzerland), Paul J. Zwier, and the proposers.

Vowe *notes that this problem is a special case of Problem* 4068 *in the Amer. Math. Monthly*, 1941, 476–477.

A triangle invariant **June 1991**

1377. *Proposed by Hüseyin Demir, Middle East Technical University, Ankara, Turkey.*

Let DEF be a variable triangle inscribed in triangle ABC, and let U, X, V, Y, W, Z be the midpoints of the line segments BD, DC, CE, EA, AF, and FB, respectively. Show that the expression

$$|UVW| + |XYZ| - \tfrac{1}{2}|DEF|$$

for areas is constant.

Solution by Hans Kappus, Mathematisches Institut der Universität, Basel, Switzerland.

Denote the expression in question by S. We show that $S = (3/4)$ Area ABC.

Since $S/$Area ABC remains unchanged under affine transformations we may choose the affine coordinate system so that $A = (0,0)$, $B = (1,0)$, and $C = (0,1)$. Now let

$$D = (1-r, r), \qquad E = (0, s), \qquad F = (t, 0); \qquad 0 \le r, s, t, \le 1.$$

Then we have

$$U = (1 - r/2, r/2), \qquad V = (0, (1+s)/2), \qquad W = (t/2, 0),$$
$$X = ((1-r)/2, (1+r)/2), \qquad Y = (0, s/2), \qquad Z = ((1+t)/2, 0).$$

Using these coordinates the following areas may be calculated in a straightforward manner:

$$\text{Area } UVW = \tfrac{1}{8}(2 - r + 2s - t - rs + rt - st)$$
$$\text{Area } XYZ = \tfrac{1}{8}(1 + r + t - rs + rt - st)$$
$$\text{Area } DEF = \tfrac{1}{2}(s - rs + rt - st).$$

From this it follows that $S = 3/8 = (3/4)$ Area ABC.

Also solved by Beno Arbel (Israel), H. Guggenheimer, Francis M. Henderson, John G. Heuver (Canada), Thomas Jager, Václav Konečny, Helen M. Marston, Ralph Merrill, José Heber Nieto (Venezuela), Chrysostom G. Petalas (Greece), F. C. Rembis, Robert L. Young, Paul J. Zwier, an unsigned solution, and the proposer.

Several people mentioned that the problem is incorrect as stated. The intention in the problem was that D, E, F should be on line segments BC, CA, and AB respectively. A corrected version of this problem appears as 1371 *in April* 1991, *and several solutions are given in the April* 1992 *issue. Somehow the uncorrected version did not get lifted from the file of accepted proposals, so it inadvertently reappeared. Apologies.*

Answers

Solutions to the Quickies on page 195.

A791. For each combination of coin tosses C, let C' be the combination produced by reversing every coin. It is clear that this association yields an involution on the set of all combinations of coin tosses. But it is also easy to see that it reverses the win-loss outcome of any combination. Therefore the number of winning combinations equals the number of losing combinations, so the game is fair.

A792. The equation is equivalent to

$$(x+y+z)(y+z-x)(z+x-y)(x+y-z)=3.$$

Hence, $x+y+z=3$ and then $x=y=z=1$.

A more interesting problem is to find integers w such that

$$(x+y+z)(y+z-x)(z+x-y)(x+y-z)=3w^4$$

has solutions other than $x=y=z=w$. Geometrically, this problem is equivalent to finding integer triangles having the same area as an equilateral triangle of side w.

A793. It suffices to show that the desired area is equal to that of $\triangle AOB$. So consider the special case in which the parallel lines through A and B coincide with the side AB, and let M denote the intersection of OC and AB. We need only demonstrate that Area $\triangle AQM =$ Area $\triangle BPM$. But this is clear; $APBQ$ is a trapezoid, and therefore Area $\triangle APQ =$ Area $\triangle APB$ since each has the same base and height. Subtracting Area $\triangle APM$ from each gives the result.

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.

PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies. Quarterly, 112 pp each, 6" × 9". Brian J. Winkel, editor, Dept. of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN 47803. \$34/yr in US, \$40/yr outside.

PRIMUS, celebrating its first birthday, is a refereed journal that provides a much-needed "forum for the exchange of ideas in mathematics education at the college level." Some of the articles in its first year have treated advice for undergraduates on how to write mathematics, using software in teaching differential equations, having students write word problem poems, implementing Taylor polynomials on a spreadsheet, introducing writing and speaking in mathematics courses, and lots more ideas on how to reform and revitalize the instruction in your courses, from intermediate algebra through calculus, statistics, geometry, and differential equations. Read *PRIMUS* for ideas, and share your own ideas by writing for it.

Maddox, John, The endless search for primality, *Nature* 356 (26 March 1992) 283. Gladwell, Malcolm, Mathematical nirvana has 227,832 digits and is divisible by 1, *Washington Post* (28 March 1992) A10.

After 19 hours of searching, a Cray-2 supercomputer at AEA Technology's Harwell Laboratories in Britain has found that $2^{756,839} - 1$ is prime. That number thus becomes the 32^{nd} known Mersenne prime (a prime of the form $2^p - 1$, with p a prime) and the largest known prime. It had been four years since discovery of the previous Mersenne prime. The computer program used, which was designed by David Slowinski (Cray Research), is based on the Lucas-Lehmer test. By the theorem in Euclid's *Elements* Book XIII, the new Mersenne prime also yields the 32^{nd} known perfect number; it is unknown whether there are infinitely many. "But is it more than stamp-collecting?" asks the *Nature* article. When asked on National Public Radio whether this discovery would help computers run faster or produce other tangible results in the lives of consumers, Slowinski replied that the same techniques useful in doing this computation efficiently are indeed the stuff of which such advances are made. (The *Nature* article falsely asserts that if P is the product of the first n primes, then $P + 1$ and $P - 1$ are both primes.)

Cipra, Barry, You can't hear the shape of a drum, *Science* 255 (27 March 1992) 1642–1643.

In 1966, Mark Kac asked "Can you hear the shape of a drum?" (*American Mathematical Monthly* 73 (4, Part II) 1–23). In one dimension, the answer is yes (the length of a violin string is determined by the lowest frequency produced by plucking it); in dimensions greater than two, the answer is no (there are "hyperdrums" of different shapes that make the same "hypersounds"). What about dimension 2? For a vibrating membrane, are the size and shape of the bounding curve uniquely determined by the eigenvalues of the wave motions of the membrane? Kac thought not, but "I may well be wrong and I am not prepared to bet large sums either way." Perhaps he should have trusted his intuition, because Carolyn Gordon and David Webb (Washington University) and Scott Wolpert (University of Maryland) have exhibited counterexamples at last.

Cohen, Marcus, Edward D. Gaughan, Arthur Knoebel, Douglas S. Kurtz, and David Pengelley, *Student Research Projects in Calculus*, MAA, 1991; ix + 216 pp, $21 (P).

Do students learn mathematics better through individual and group research projects? Here are 103 imaginative projects that are carefully written and well-thought-out, appeal to students' imagination and creativity, and teach concepts. The projects, which focus almost entirely on pure mathematics, come with guidelines and advice on productive use. Some projects are delightfully fantastical: Students will enjoy the takeoff on the show *Star Trek*, and faculty will find "The All-Purpose Calculus Project" irresistibly funny.

Preston, Richard, Profiles: The mountains of pi, *The New Yorker* (2 March 1992) 36–67.

Another contribution to the common myth of mathematicians as crazy and ill-kempt weirdos and Don Quixotes, this is nevertheless an enthralling portrait of Gregory and David Chudnovsky. These two Soviet emigré brothers have designed and built a supercomputer in their apartment from mail-order parts and used it to generate—and check for patterns in—two billion digits of pi. So far, no patterns. Says David, "We need a trillion digits."

Bass, Thomas, Road to ruin, *Discover* 13 (5) (May 1992) 56–61.

This is a popular exposition of Braess's paradox, which states that adding capacity or a new route to a congested network may make the situation worse for every user. This article has little additional detail beyond Gina Kolata's article "What if they closed 42d Street and nobody noticed?" (*New York Times* (25 December 1990) 38), reviewed in this column in vol. 64 (3) (June 1991) 207.

Stewart, Ian, Mathematical recreations: Concentration: A winning strategy, *Scientific American* 265 (4) (October 1991) 126–128.

Concentration began as a card game and found popularity as a TV game show of the same name. Players take turns turning over cards, trying to turn over matched pairs. Memory and chance play a role, but so does strategy; and this article reveals a simple optimal strategy due to Uri Zwick and Michael S. Paterson (University of Warwick). (It is not a "winning" strategy, as it does not guarantee a win.)

Cipra, Barry, What goes around comes around, *Science* 255 (6 March 1992) 1212–1213.

"[I]f Arnold Schwarzenegger crushes a basketball, how many unbroken rubber bands can Magic Johnson wrap around it?" In mathematical terms: "How many closed geodesics are there for any Riemannian metric on a sphere?" The answer—infinitely many—is the joint work of John Franks (Northwestern University) and Victor Bangert (University of Freiburg); their collaboration represents the introduction of methods from dynamical systems into differential geometry, an idea that may prove even more fruitful in the future.

Stewart, Ian, Justifying the means, *Nature* 354 (21 November 1991) 185–186.

Iteration of the function $f(a,b) = (\frac{1}{2}(a+b), \sqrt{ab})$ produces a fixed point, with both coordinates converging to the same number, the *arithmetic-geometric mean* (AGM). Gauss devised this mean, and Jonathan and Peter Borwein made it the central concept of their book *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity* (Wiley, 1987). Stewart, basing this article on a paper by Shaun Bullett (*Topology* 30 (1991) 171–190), considers this process from the viewpoint of iterated systems, with its emphasis on the orbits and fixed points of the map and the fractal nature of the dynamics involved. Bullett has also established a "curious connection" between the AGM and the octahedron; other quadratic maps correspond to other regular solids.

Bolt, Brian, *Mathematics Meets Technology*, Cambridge University Press, 1991; x + 203 pp, (P).

Discusses the kinematics of a wide variety of mechanisms—pulleys, sprockets, chains, gears, cranes, linkages, winches, screws, cams, ratchets, levers—"as seen through the eyes of a mathematician." Hundreds of well-drawn figures illustrate mechanical devices of all kinds, together with exercises on the mathematics (geometry, trigonometry, and algebra) involved. Utterly fascinating!

Stewart, Ian, Mathematical Recreations: All paths lead away from Rome, *Scientific American* 266 (4) (April 1992) 150–152.

In a circular arena, how can a gladiator avoid a lion, if both can run at the same speed? Not by running in a circle concentric with the arena, but by following a *squiral*: "a spiral made up of successive line segments, each at right angles to the radius." The problem can be generalized by adding more lions ("n lions ... always catch [the gladiator] in an n-dimensional ball, but $n-1$ lions cannot if the warrior adopts the right tactics") and then adding obstacles to the arena (some open problems here). The results are too late for the clients who would have had the greatest stake in the results of the research ... .

Lesurf, Jim, A spy's guide to chaos, *New Scientist* (1 February 1992) 29–33.

Adding random noise to an encoded message can weaken cryptanalytic attacks that are based on statistical analysis. Deterministic pseudorandom number generators provide one way for sender and receiver to generate the same stream of deterministic noise. The author of this article and Malcolm Robertson (both at University of St. Andrews) have "discovered that simple nonlinear devices in circuits ... can produce random noise signals equivalent to generating more than a billion apparently random numbers every second." The devices involve semichaotic oscillators.

Peterson, Ivars, Bringing random walkers into new territory, *Science News* 141 (8 February 1992) 84–85. Shlesinger, Michael F., New paths for random walkers, *Nature* 355 (30 January 1992) 396–397. Larralde, H., et al., *Nature* 355 (30 January 1992) 423–426.

What happens if you generalize random walk to allow for multiple walkers? Well, you would have a model for spread of populations. What happens? Over time, the boundary of the sites visited passes through two qualitative transitions, eventually becoming rougher and fractal in shape. And a simple question produces rich results.

Selvin, Paul, Profile of a field: Mathematics: Heroism is still the norm, *Science* 255 (13 March 1992) 1382–1383.

In a special section devoted to women in science, this article reminds readers in all disciplines of the small numbers of women in top graduate university mathematics departments and of the explicit sexism that most encounter.

Kenschaft, Patricia Clark and Sandra Keith (eds.), *Winning Women into Mathematics*, MAA, 1991; ix + 78 pp, $12 (P).

Here is a comprehensive resource on the issues of women's participation in mathematics, from a list of cultural patterns that are causing women to be underrepresented in mathematics, through a history of women in mathematics and in the MAA, to overviews of model programs. Included are the scripts of the "Micro-Inequities" skits, whose presentations have proven so popular at national meetings of the MAA.

NEWS AND LETTERS

TWENTIETH ANNUAL U.S.A. MATHEMATICAL OLYMPIAD PROBLEMS AND SOLUTIONS

1. In triangle ABC, angle A is twice angle B, angle C is obtuse, and the three side lengths a, b, c are integers. Determine, with proof, the minimum possible perimeter.

Solution. Let a, b, c be the lengths of the sides opposite angles A, B, C, respectively. By the law of sines,

$$\frac{a}{b} = \frac{\sin 2B}{\sin B} = 2\cos B,$$

$$\frac{c}{b} = \frac{\sin(\pi - 3B)}{\sin B} = \frac{\sin 3B}{\sin B} = 4\cos^2 B - 1.$$

Hence $c/b = (a/b)^2 - 1$, from which

$$a^2 = b(b+c). \qquad (1)$$

Since we are looking for a triangle of smallest perimeter, we may assume that a, b and c have no common prime factor; otherwise a smaller example would exist. So in view of (1), b and $b+c$ have no common prime divisors. Since their product is a perfect square, b and $b+c$ must be perfect squares. Thus, for some relatively prime integers m and n, we have $b = m^2$, $b + c = n^2$, $a = mn$ and $n/m = a/b = 2\cos B$. The angle $C = \pi - 3B$ is obtuse, so $0 < B < \pi/6$ which implies $\sqrt{3}/2 < \cos B < 1$ and thus $\sqrt{3} < n/m < 2$. It is easy to see that this inequality has no integer solution with $m = 1, 2$ or 3. Hence $m \geq 4$, $n \geq 7$ and

$$a + b + c = mn + n^2 \geq 4 \cdot 7 + 7^2 = 77.$$

In fact, the pair $(m, n) = (4, 7)$ generates a triangle with $(a, b, c) = (28, 16, 33)$ and this triangle meets all the necessary geometric conditions, so 77 is the minimum possible perimeter.

2. For any nonempty set S of numbers, let $\sigma(S)$ and $\pi(S)$ denote the sum and product, respectively, of the elements of S. Prove that

$$\sum \frac{\sigma(S)}{\pi(S)} = (n^2 + 2n) - \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)(n+1),$$

where "Σ" denotes a sum involving all nonempty subsets S of $\{1, 2, 3, \ldots, n\}$.

Solution. It is convenient to extend the definitions of σ and π by setting $\sigma(S) = 0$ and $\pi(S) = 1$ in case S is empty. With this convention, the sum $\sum \sigma(S)/\pi(S)$ over subsets can be extended to include the empty set without changing the result. Consider the easier problem of computing $\sum_{S \subseteq A} 1/\pi(S)$, where A is a finite set of positive numbers. We have

$$\sum_{S \subseteq A} \frac{1}{\pi(S)} = \prod_{a \in A} \left(1 + \frac{1}{a}\right) \qquad (2)$$

in view of the one-to-one correspondence between the terms on the left and those obtained by expanding the product on the right. Let $[n] = \{1, 2, \ldots, n\}$. As a special case of (2) we have

$$\sum_{S \subseteq [n]} \frac{1}{\pi(S)} = \prod_{k=1}^{n} \left(1 + \frac{1}{k}\right) = n + 1. \qquad (3)$$

For $k = 1, 2, 3, \ldots$ let

$$A_k = \sum_{S \subseteq [k]} \frac{\sigma(S)}{\pi(S)}.$$

The terms which contribute to A_k but not to A_{k-1} are of the form $\sigma(S)/\pi(S)$ where $S = S' \cup \{k\}$ and $S' \subseteq [k-1]$. Thus

$$\frac{\sigma(S)}{\pi(S)} = \frac{\sigma(S') + k}{k\,\pi(S')},$$

and

$$A_k - A_{k-1} = \sum_{S' \subseteq [n-1]} \frac{\sigma(S') + k}{k\,\pi(S')} = \frac{1}{k} A_{k-1} + k,$$

in view of (3). The recurrence formula may be written

$$\frac{A_k}{k+1} - \frac{A_{k-1}}{k} = \frac{k}{k+1} = 1 - \frac{1}{k+1},$$

and this holds for all $k \geq 1$ if we set $A_0 = 0$. Summing from $k = 1$ to n, the left hand side telescopes and we find

$$\frac{A_n}{n+1} = n - \left(\frac{1}{2} + \cdots + \frac{1}{n} + \frac{1}{n+1}\right),$$

which then yields the required result.

3. Show that, for any fixed integer $n \geq 1$, the sequence

$$2,\ 2^2,\ 2^{2^2},\ 2^{2^{2^2}}, \ldots (\operatorname{mod} n)$$

is eventually constant.

[The tower of exponents is defined by $a_1 = 2$, $a_{i+1} = 2^{a_i}$. Also $a_i \pmod n$ means the remainder which results from dividing a_i by n.]

Solution. Our proof is by induction on n. The case $n = 1$ is clear. Now take $n > 1$ and suppose that the result is true for all positive integers less than n. We distinguish two cases.

Case 1: n is even. Write $n = 2^k q$ where $k \geq 1$ and q is odd. By induction, the sequence $a_1, a_2, a_3, \ldots$ is eventually constant modulo q. Clearly $a_i \equiv 0 \pmod{2^k}$ for all large i. Since 2^k and q are relatively prime, $2^k | (a_{i+1} - a_i)$ and $q | (a_{i+1} - a_i)$ imply that $n | (a_{i+1} - a_i)$. Thus the sequence $a_1, a_2, a_3, \ldots$ is eventually constant modulo n.

Case 2: n is odd. In this case, there exists a positive integer $r < n$ such that

$$2^r \equiv 1 \pmod n. \qquad (4)$$

Indeed, Euler's extension of Fermat's theorem yields (4) with $r = \phi(n)$ where ϕ is the Euler function. By induction, the sequence $a_1, a_2, a_3, \ldots$ is eventually constant modulo r. But in view of (4), $a_i \equiv c \pmod r$ implies

$$a_{i+1} = 2^{a_i} = 2^{m_i r + c} \equiv 2^c \pmod n.$$

Thus the sequence $a_1, a_2, a_3, \ldots$ is eventually constant modulo n.

4. Let $a = \dfrac{m^{m+1} + n^{n+1}}{m^m + n^n}$, where m and n are positive integers. Prove that $a^m + a^n \geq m^m + n^n$.

[You may wish to analyze the ratio $\dfrac{a^N - N^N}{a - N}$ for real $a \geq 0$ and integer $N \geq 1$.]

Solution. If N is a positive integer and $a \neq N$, then of course

$$\frac{a^N - N^N}{a - N} = a^{N-1} + a^{N-2}N + \cdots + N^{N-1}.$$

If $0 \leq a < N$, then this is the sum of N terms each less than or equal to N^{N-1}, and consequently

$$\frac{a^N - N^N}{a - N} \leq N^N.$$

Similarly, if $a > N$, then

$$\frac{a^N - N^N}{a - N} \geq N^N.$$

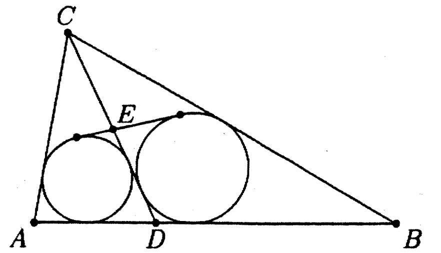
Since $a - N$ is negative in the first case and positive in the second, we obtain

$$a^N - N^N \geq (a - N)N^N \qquad (5)$$

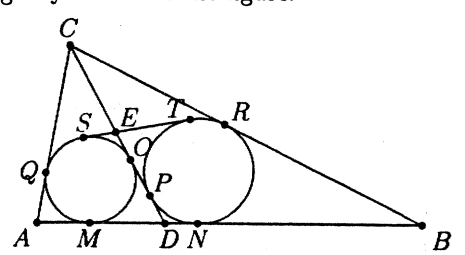
is both cases. Certainly this result is also true for $a = N$; thus the inequality (5) holds for all $a \geq 0$. We now use (5) to establish the desired result by observing that $(a^m + a^n) - (m^m + n^n)$

$$\begin{aligned} &= (a^m - m^m) + (a^n - n^n) \\ &\geq (a - m)m^m + (a - n)n^n \\ &= a(m^m + n^n) - (m^{m+1} + n^{n+1}) \\ &= 0. \end{aligned}$$

5. Let D be an arbitrary point on side AB of a given triangle ABC, and let E be the interior point where CD intersects the external common tangent to the incircles of triangles ACD and BCD. As D assumes all positions between A and B, prove that the point E traces the arc of a circle.



Solution. Experimentation with straightedge and compass suggests that point E traces the arc of a circle with center C. We prove that this is in fact the case by showing that the length of segment CE is constant for all positions of point D. This we accomplish by systematically matching pairs of equal tangents in the figure, working from CE back toward the fixed lengths of triangle ABC.
Let M through T be the points of tangency shown in the figure.



By equal tangents,

$$CE = CO - EO = CQ - SE,$$

and

$$CE = CP - EP = CR - ET.$$

Adding these and repeatedly using equal tangents, we obtain

$$\begin{aligned} 2 \cdot CE &= CQ + CR - (SE + ET) \\ &= (CA - QA) + (CB - RB) - ST \\ &= (CA - AM) + (CB - NB) - MN \\ &= CA + CB - (AM + MN + NB) \\ &= CA + CB - AB. \end{aligned}$$

Therefore

$$CE = \tfrac{1}{2}(CA + CB - AB),$$

and the problem is solved.

THIRTYSECOND ANNUAL INTERNATIONAL MATHEMATICAL OLYMPIAD

PROBLEMS

1. Given a triangle ABC, let I be the center of its inscribed circle. The internal bisectors of the angles A, B, C meet the opposite sides in A', B', C' respectively. Prove that

$$\frac{1}{4} < \frac{AI \cdot BI \cdot CI}{AA' \cdot BB' \cdot CC'} \leq \frac{8}{27}.$$

2. Let $n > 6$ be an integer and $a_1, a_2, \ldots, a_k$ be all the natural numbers less than n and relatively prime to n. If

$$a_2 - a_1 = a_3 - a_2 = \cdots = a_k - a_{k-1} > 0,$$

prove that n must be either a prime number or a power of 2.

3. Let $S = \{1, 2, 3, \ldots, 280\}$. Find the smallest integer n such that each n-element subset of S contains five numbers which are pairwise relatively prime.

4. Suppose G is a connected graph with k edges. Prove that it is possible to label the edges $1, 2, \ldots, k$ in such a way that at each vertex which belongs to two or more edges, the greatest common divisor of the integers labeling those edges is equal to 1.

[A *graph* consists of a set of points, called *vertices*, together with a set of *edges* joining certain pairs of distinct vertices. Each pair of vertices u, v belongs to at most one edge. The graph G is *connected* if for each pair of distinct vertices x, y there is some sequence of vertices $x = v_0, v_1, v_2, \ldots, v_m = y$ such that each pair v_i, v_{i+1} $(0 \leq i < m)$ is joined by an edge of G.]

5. Let ABC be a triangle and P an interior point of ABC. Show that at least one of the angles $\angle PAB$, $\angle PBC$, $\angle PCA$ is less than or equal to 30°.

6. An infinite sequence $x_0, x_1, x_2, \ldots$ of real numbers is said to be *bounded* if there is a constant C such that $|x_i| \leq C$ for every $i \geq 0$.

Given any real number $a > 1$, construct a bounded infinite sequence $x_0, x_1, x_2, \ldots$ such that

$$|x_i - x_j||i - j|^a \geq 1$$

for every pair of distinct nonnegative integers i, j.

DERIVE® a compact card.

DERIVE®, A Mathematical Assistant is now available for palmtops through 486-based PCs.

DERIVE is a registered trademark of Soft Warehouse, Inc.

The DERIVE® program solves both symbolic and numeric problems, and it plots beautifully too.

- Symbolic math from algebra through calculus.
- Plots in both 2-D and 3-D.
- Simple, letter-driven menu interface.
- Solves equations exactly.
- Understands vectors and matrices.
- Split or overlay algebra and plot windows.
- Displays accepted math notation.
- Performs arithmetic to thousands of digits.
- Simplifies, factors and expands expressions.
- Does exponential, logarithmic, trigonometric, hyperbolic and probability functions.
- Taylor and Fourier series approximations.
- Permits recursive and iterative programming.
- Can generate Fortran, Pascal and Basic statements.

System requirements

PC version: MS-DOS 2.1 or later, only 512Kb RAM and one 3.5" or 5.25" disk drive. Suggested retail price is $250.

ROM-card version: Hewlett-Packard 95LX Palmtop computer. Suggested retail price is $289.

Contact Soft Warehouse for a list of dealers. Or, ask at your local computer store, software store or HP calculator dealer. Dealer inquires are welcome.

2000 Years of Mathematical Knowledge on a Disk

Soft Warehouse Inc.
HONOLULU • HAWAII

Soft Warehouse, Inc. • 3660 Waialae Avenue
Suite 304 • Honolulu, HI, USA 96816-3236
Phone: (808) 734-5801 • Fax: (808) 735-1105

CONTENTS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, D.C. 20036