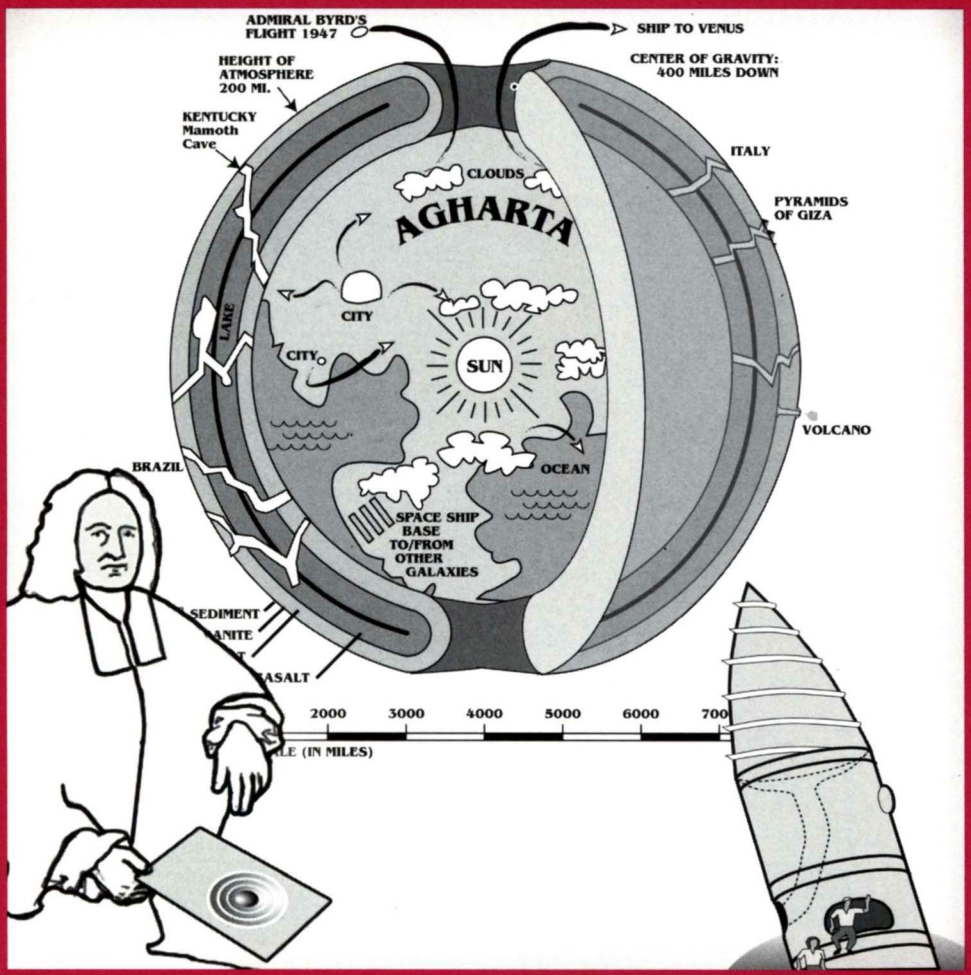
Vol. 75, No. 5, December 2002

MATHEMATICS MAGAZINE



The Hollow Earth

- The Gravity of Hades
- Beyond the Celestial Sphere: Oriented Projective Geometry and Computer Graphics

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

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

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

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

Cover image: *The Hollow Earth*, by Jason Challas. Based on faulty data he inherited from Newton, Halley believed the Earth to be 4/9 hollow and posited life on the interior surfaces. The Hollow Earth Society tells us that superior beings inhabit the inner region, which they call Agharta. In Edgar Rice Burroughs' science fiction novel *At the Earth's Core*, adventurers reach the land of Pellucidar by means of a cork-screw burrowing machine. Simoson's article, *The Gravity of Hades*, examines the physics behind such possibilities.

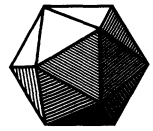
Jason Challas lectures on the gravity of computer art at Santa Clara University.

AUTHORS

Kevin G. Kirby is Associate Professor of Mathematics and Computer Science at Northern Kentucky University. He obtained his Ph.D. from Wayne State University in 1988. His research interests lie in biological and quantum computing, but years of teaching computer graphics have put him under the spell of geometry. In 1998 he won the MAA's Polya Award for a paper on the linear algebra of neural associative memory.

Andrew Simoson has been chairman of the mathematics department at King College since 1980. During a recent calculus class after his students had responded with understanding about how gravity changes as one rises above the Earth's surface, he hazarded the reciprocal question. After a few minutes of his braver students reasoning aloud, he realized that he himself did not know how gravity changes with depth. His quest to find a good answer resulted in the romp through history presented herein.

Vol. 75, No. 5, December 2002



MATHEMATICS MAGAZINE

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

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

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

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

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

Printed in the United States of America

The Gravity of Hades

ANDREW J. SIMOSON
King College
Bristol, TN 37620
ajsimoso@king.edu

Imagine being beamed into a chamber deep within the Earth, such as depicted in the Jules Verne cavern of FIGURE 1. Is the acceleration due to gravity stronger or weaker in the chamber than at the surface of the Earth? If the reader is hesitant in answering immediately, here's how Leonard Euler answered the question in 1760 [**3**, volume I, Letter L, p. 182]:

> *We are certain ... that gravity ... acts with the greatest force at the surface of the earth, and is diminished in proportion as it removes from thence, whether by penetrating towards the centre or rising above the surface of the globe.*

Figure 1 Frontispiece for Verne's 1871 publication of *Journey to the Center of the Earth*, 100 miles below Iceland

To answer this question, we consider various models of the Earth's structure that have been proposed over the years. We show that

- for homogeneously dense planets—Euler's implicit model—gravity weakens with descent from the surface;
- for a planet possessing a homogeneous mantle that is less than 2/3 as dense as its core, a local minimum for gravity intensity exists within a sufficiently thick mantle;
- for our own Earth, having a solid inner core, a less dense liquid outer core, and an extensive even less dense mantle, gravity intensifies with descent from $-9.8\ \mathrm{m/s^2}$ at the surface to an extreme of $-10.8\ \mathrm{m/s^2}$ where the mantle meets the core.

We also give a condition involving only surface density and mean density that determines whether gravity increases or decreases with depth from the surface of a planet; and we conclude with an analysis of a body falling through a classically envisioned hole through the Earth.

Some preliminary classical mechanics

To find the gravitational acceleration induced by the Earth on a particular point P, we follow Newton and first find the gravitational acceleration induced by each point of the Earth on P, and then take the aggregate of these accelerations as the total acceleration induced by the Earth. With this idea in mind, let the mass m of a body be condensed to the single point $Q = (x, y, z)$. We wish to determine the gravitational acceleration $a(s)$ induced by the mass at Q in the direction $\mathbf{k} = \langle 0, 0, 1\rangle$ on point $P = (0, 0, s)$. See FIGURE 2.

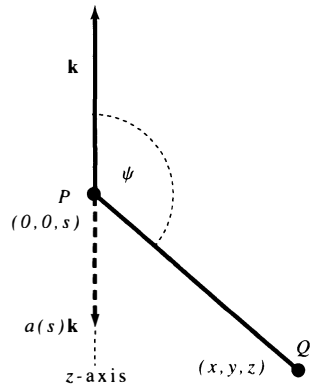


Figure 2 The pull $a(s)$ by Q on P

We start with Newton's law of gravitation, telling us that the force of gravity between two single-point masses a and b separated by r units is Gab/r^2, where $G \approx 6.67 \times 10^{-11}$ Nm^2/kg^2, the universal gravitation constant. Since the force on a mass is equal to mass times acceleration, the gravitational acceleration on the point P in the direction $\overrightarrow{PQ}$ is $Gm/(x^2 + y^2 + (z - s)^2)$. But we want only the portion of this attraction in the direction $\mathbf{k}$. Note that if ψ is the angle between $\overrightarrow{PQ}$ and $\mathbf{k}$, then

$$\cos\psi = \frac{\langle x, y, z - s\rangle \cdot \langle 0, 0, 1\rangle}{\sqrt{x^2 + y^2 + (z - s)^2}} = \frac{z - s}{\sqrt{x^2 + y^2 + (z - s)^2}}.$$

Therefore $a(s) = \cos\psi\, Gm/(x^2 + y^2 + (z - s)^2)$, which we write as

$$\boxed{a(s) = \frac{Gm(z - s)}{(x^2 + y^2 + (z - s)^2)^{\frac{3}{2}}}.} \tag{1}$$

A point mass at (x,y,z)

Ideal planet models

To apply (1) to calculate the gravitational acceleration a for a heavenly body at distance s from its center of mass, we focus on *ideal planets*, those with radius R for which each of its concentric spherical shells is of constant density. Though our planet is best described as an oblate spheroid which, when compared to the best approximating

sphere, is flattened at the poles by about 14 km and bulges at the equator by about 7 km, and though its shells have only approximately constant density, the Earth is viewed as an ideal planet for the purposes of this paper. For each ideal planet we let $a(s)$ be the outward radial gravitational acceleration induced by the planet at s units from its center. For each ideal planet, we presently show that $a(s) \leq 0$ for all $s \geq 0$. Therefore, the greatest magnitude for acceleration occurs when $a(s)$ is a minimum.

The simplest ideal planet is a single nonzero dense onion layer, or a soap bubble.

A soap bubble model Let S be a spherical shell of mass M with constant density, center $O = (0, 0, 0)$ and radius r. Then the gravitational acceleration on the point P due to the shell S is

$$a(s) = \begin{cases} 0, & \text{if } 0 \leq s < r, \\ -\dfrac{GM}{2s^2}, & \text{if } s = r, \\ -\dfrac{GM}{s^2}, & \text{if } s > r. \end{cases} \tag{2}$$

A spherical shell

The derivation of (2) is probably in every classical mechanics text, such as Fowles and Cassiday [**5**, pp. 207–209]. For completeness and because the derivation herein is a bit simpler than the ones in the mechanics texts surveyed by this author, we derive (2) by integrating (1) over the shell S. Let $\mathbf{r} = \langle r\cos\theta\sin\phi,\ r\sin\theta\sin\phi,\ r\cos\phi\rangle$ be a parameterization of S, where $0 \leq \theta \leq 2\pi$ and $0 \leq \phi \leq \pi$. Then the area element for S is $\|\frac{\partial \mathbf{r}}{\partial\theta} \times \frac{\partial \mathbf{r}}{\partial\phi}\| \, d\theta\, d\phi$, which simplifies to $r^2 \sin\phi\, d\theta\, d\phi$. Since surface area of a sphere of radius r is $4\pi r^2$, the point-mass m of (1) can be taken as $\frac{M}{4\pi r^2} r^2 \sin\phi\, d\theta\, d\phi = \frac{M}{4\pi} \sin\phi\, d\theta\, d\phi$. Therefore from (1), acceleration $a(s)$ by S on point P is

$$a(s) = \int_0^{\pi}\int_0^{2\pi} \frac{G\frac{M}{4\pi}(r\cos\phi - s)\sin\phi}{(r^2 + s^2 - 2rs\cos\phi)^{\frac{3}{2}}}\, d\theta\, d\phi = \int_0^{\pi} \frac{G\frac{M}{2}(r\cos\phi - s)\sin\phi}{(r^2 + s^2 - 2rs\cos\phi)^{\frac{3}{2}}}\, d\phi. \tag{3}$$

The change of variable $u = \cos\phi$ gives

$$a(s) = \frac{GM}{2}\int_{-1}^{1} \frac{ru - s}{(r^2 + s^2 - 2rsu)^{\frac{3}{2}}}\, du, \tag{4}$$

which, via integration by parts and simplifying, gives

$$a(s) = \frac{GM}{2s^2}\left(\frac{r - s}{\sqrt{(r - s)^2}} - \frac{r + s}{\sqrt{(r + s)^2}}\right). \tag{5}$$

If $0 < s < r$ then (5) gives 0; if $s > r$ then (5) gives $-GM/s^2$. To obtain the anomalous result for $s = r$, first simplify (4) before integrating.

Since a general ideal planet is simply a series of concentric soap bubbles, when s exceeds R, (2) is the familiar principle that when dealing with forces exerted by heavenly bodies, one can treat their masses as being concentrated at their centers. Furthermore, when $0 \leq s < R$, (2) is a well-known result of electrostatics, which has features in common with gravitation; that is, for a hollow, metal sphere carrying a surface charge, the electrical field within the sphere is 0. Similarly, for any void within

a charged metal object, the field is zero as well [**4**, chapter 5], which is *not* the case for the gravitational field within an arbitrary void of a planet.

As a simple example showing that neither of these two properties necessarily holds if a heavenly body is not an ideal planet, let S be a hollow, right circular cylinder with mass M, height 2, radius 1, center at the origin, and central axis aligned with the z-axis. Since the surface area of S is 6π, assume that S has uniform density of $M/6\pi$. To calculate the gravitational acceleration induced by S at the point $(0, 0, s)$ where $s \geq 0$, use (1) and integrate over the top, the bottom, and the sides of S, giving $a(s)$ as

$$\frac{GM}{6\pi}\left(\int_0^1\int_0^{2\pi}\frac{(1-s)r}{(r^2+(1-s)^2)^{\frac{3}{2}}}\,d\theta\,dr-\int_0^1\int_0^{2\pi}\frac{(1+s)r}{(r^2+(1+s)^2)^{\frac{3}{2}}}\,d\theta\,dr\right.$$
$$\left.+\int_{-1}^1\int_0^{2\pi}\frac{(z-s)}{(1+(z-s)^2)^{\frac{3}{2}}}\,d\theta\,dz\right),$$

which simplifies to

$$a(s)=\begin{cases}\dfrac{GM}{3}\left(\dfrac{s+2}{\sqrt{1+(1+s)^2}}+\dfrac{s-2}{\sqrt{1+(1-s)^2}}\right), & \text{if } 0\leq s<1,\\[2ex] \dfrac{GM(3-2\sqrt{5})}{15}, & \text{if } s=1,\\[2ex] \dfrac{GM}{3}\left(\dfrac{s+2}{\sqrt{1+(1+s)^2}}+\dfrac{s-2}{\sqrt{1+(1-s)^2}}-2\right), & \text{if } s>1.\end{cases}$$

Observe that for $0 < s < 1$, $a(s) > 0$, which means that if a particle is on the z-axis between 0 and 1, then it is attracted towards the top of the cylinder. To see that this formula conforms asymptotically to an inverse square law, observe that $\lim_{s\to\infty} s^2 a(s)$ is a constant.

In general, when calculating the gravitational acceleration induced by a heavenly body at a certain point, any nonspherical symmetry of the body promulgates nightmarish integrals. But fortunately for astrophysics, planets and stars are more or less ideal planets; furthermore, many of the central problems of celestial mechanics involve forces of gravitation at many radii from the heavenly bodies in question, so that the inverse square law can be used with confidence and relief.

To find the cumulative gravitational acceleration induced by the point-masses of a nontrivial ideal planet, it is convenient to sum them layer by collective spherical layer. With this idea in mind, we integrate (2) appropriately to find $a(s)$.

The general density model Let S be a ball of radius R, mass M, and center O whose density at s units from O is given by $\delta(s)$. That is, S is spherically symmetric with respect to its density. Then

$$a(s)=\begin{cases}-\dfrac{4\pi G}{s^2}\displaystyle\int_0^s \rho^2\delta(\rho)\,d\rho, & \text{if } 0\leq s\leq R,\\[2ex] -\dfrac{GM}{s^2}, & \text{if } s\geq R.\end{cases} \tag{6}$$

An ideal planet with general density

To derive (6) from (2) note that the mass of the shell at radius ρ is $4\pi\rho^2\delta(\rho)$. Therefore,

$$a(s) = \begin{cases} -\displaystyle\int_0^s \frac{4G\pi\rho^2\delta(\rho)}{s^2}\,d\rho, & \text{if } 0 \le s \le R, \\ -\displaystyle\int_0^R \frac{4G\pi\rho^2\delta(\rho)}{s^2}\,d\rho & \text{if } s \ge R, \end{cases}$$

which simplifies to (6), since $\int_0^R 4\pi\rho^2\delta(\rho)\,d\rho = M$. Furthermore, note that the two formulas of (6) agree at $s = R$.

Constant acceleration: an old implicit model

When reading of mythical heroes venturing off into the underworld, such as Orpheus seeking to reclaim his lost love or Hercules throttling Cerberus, the watchdog of Hades with three heads, one forms the impression that no matter how deep the heroes descend into the Earth, gravity remains constant. To catch a bit of the flavor of these tales, here's a snippet of Ovid's *Metamorphoses* [**11**, Book X], describing Orpheus and his wife's ascent from Hades:

> *Now thro' the noiseless throng their way they bend,*
> *And both with pain the rugged road ascend;*
> *Dark was the path, and difficult, and steep,*
> *And thick with vapours from the smoaky deep.*

The drawings in FIGURE 3, based on two old Grecian urns, show these two heroes in Hades; Orpheus is serenading the lords of the underworld and Hercules is about to chain Cerberus, depicted with only two heads. Turner [**15**] enumerates other literary characters who frequented the underworld.

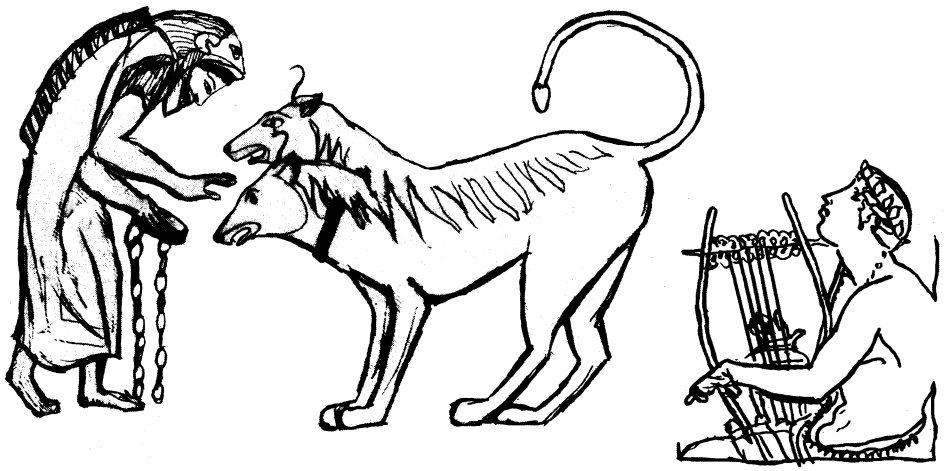


Figure 3 Hercules and Orpheus in Hades (renditions by Jason Challas)

Such an impression is especially vivid when reading Dante (in *The Divine Comedy* written about 1300 C.E.), who journeys with his ghost guide Virgil from the Earth's surface, down through 24 levels of Hell, down through the very center of the Earth. FIGURE 4 [**2**, p. 84] shows one artist's rendition of Dante's model of the Earth. When Dante scrambles over a broken bridge deep within Hell it is with the same effort he would have used along a road to Rome. Arrows and other projectiles lofted in Hell

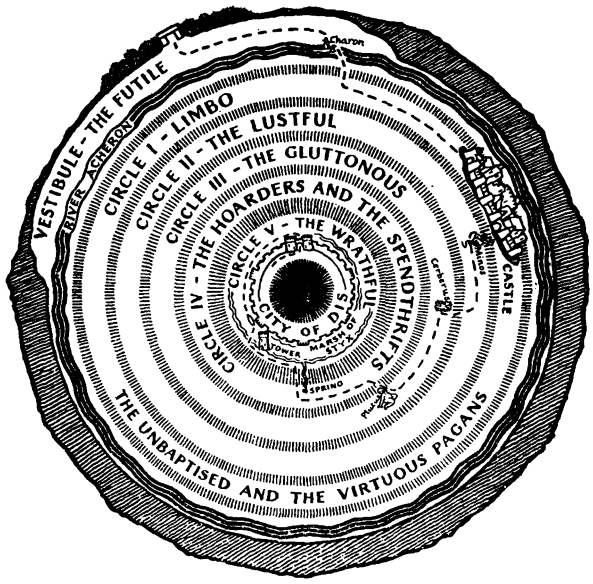


Figure 4 Levels in Dante's Earth

appear to follow the same trajectories as at the surface of the Earth. When he clambers about the hip of a somewhat immobilized, gargantuan, and hair-covered Satan at the literal center of the Earth on his way to the antipodes, Dante changes directions, maneuvering with a degree of exertion akin to that of reversing one's bodily position while clinging to a root-covered, vertical rockface in the Apennines.

> *And when we had come to where the huge thigh-bone*
> *Rides in its socket at the haunch's swell,*
> *My guide, with labour and great exertion,*
>
> *Turned head to where his feet had been, and fell*
> *To hoisting himself up upon the hair,*
> *So that I thought us mounting back to Hell.*
>
> Canto xxxiv, lines 76–81 [**2**, p. 287]

Up and down reverse dramatically. Dante's Satan, more precisely, Satan's crotch, is thus a veritable singularity. After climbing on a bit more, Dante takes a rest, looks back, and sees Lucifer's legs sticking up.

> *And if I stood dumbfounded and aghast,*
> *Let those thick-witted gentry judge and say,*
> *Who do not see what point it was I'd passed.*
>
> Canto xxxiv, lines 91–93 [**2**, p. 287]

Dante may have borrowed a similar image from the prolific, first-century writer, Plutarch, who mentions this same singularity, although less graphically so, conjecturing

> *... if a man should so coalesce with the earth that its centre is at his navel, the same person at the same time has his head up and his feet up too.* [**12**, p. 67]

Such literary evidence suggests that the *implicit* density model of the Earth used in these imaginings is that density function $\delta(s)$ which generates a constant gravitational acceleration. For $0 < s < R$, such a density function satisfies

$$-\frac{4\pi G}{s^2}\int_0^s \rho^2\delta(\rho)\,d\rho = C,$$

where C is some constant. Multiplying the above equation by s^2 and then differentiating, and simplifying yields $\delta(s) = k/s$, where k is some constant. That is, the ancient model of the Earth appears to be one wherein density and distance from the Earth's center vary inversely! Note that even in such a model, where the density at the center is unbounded, the mass of the Earth is yet finite.

A common-notion constant density model

Insisting that gravity remain constant on descent into the Earth requires a singularity in its density. Much simpler and more natural is to take this density as constant. In this author's informal survey of colleagues both in mathematics and physics, their almost universal and emphatic answer, "*Gravity decreases with descent!*," subsumes this model. Perhaps one of the reasons for such ardent responses from physicists and mathematicians is that almost all of us in our undergraduate differential equation days solved an exercise very much like this one, (which we solve in general at the end of this paper):

> *Inside the earth, the force of gravity is proportional to the distance from the center. If a hole is drilled through the earth from pole to pole, and a rock is dropped in the hole, with what velocity will it reach the center?* [**13**, p. 24]

A tempting reason to guess that gravity wanes with depth is the following, incorrect, intuitive argument: *"As we go down, the portion of the Earth above will exert an upward force on us, hence lessening the downward force on us exerted by the remainder of the Earth."* But as (2) makes abundantly clear, at s units from the center, only the mass within s units of the center determines $a(s)$.

Taking a ball S of uniform density δ, mass M, and radius R, gives by way of (6),

$$a(s) = \begin{cases} -\dfrac{4\pi G\delta s}{3}, & \text{if } 0 \le s \le R, \\ -\dfrac{GM}{s^2} & \text{if } s \ge R. \end{cases} \tag{7}$$

A planet of constant density

That is, gravity and distance from the center are directly proportional up to the surface of the Earth. See FIGURE 5a.

The reader may enjoy contrasting (7) with the acceleration functions of nonideal planets of constant density. For example, let S be a right circular cylinder of radius 1, height 2, mass M, center at the origin, and central axis aligned with the z-axis. Integrating (1) over S using cylindrical coordinates gives the gravitational acceleration $a(s)$ induced by S at the point $(0, 0, s)$ for $s \ge 0$ as

$$a(s) = MG\left(\sqrt{1+(1+s)^2} - \sqrt{1+(1-s)^2} + |1-s| - |1+s|\right),$$

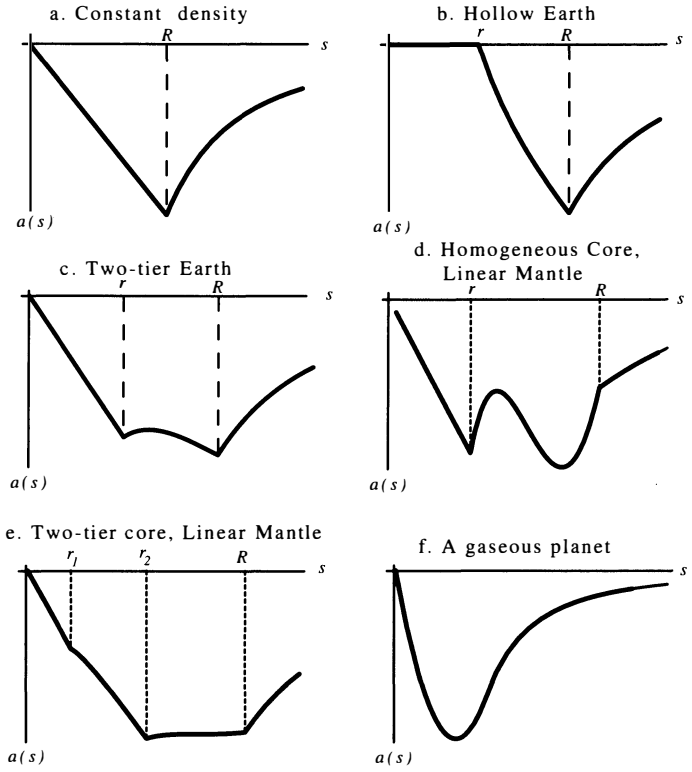


Figure 5 Simple earth models (R = radius of Earth)

whose graph qualitatively looks much like FIGURE 5a. This particular application injects life into the surprising, but otherwise artificial limit $\lim_{x\to\infty} \sqrt{1+(1+x)^2} - \sqrt{1+(1-x)^2} = 2$.

Halley's hollow Earth model

Edmond Halley, the "Father of Geophysics" and the financier behind Newton's publication of the *Principia*, viewed Newton's (erroneous) calculation in 1687 of the relative densities of the Earth and the Moon being 5 to 9, as one of the more significant of the discoveries presented in the *Principia*. Believing Newton to be correct and yet wanting to conclude that solid chunks of the Earth and Moon should be equally dense, Halley proposed that the Earth must be 4/9 hollow. Since his study of magnetic compass variations suggested that the Earth had four separate north poles, he went on to claim that the Earth's outer shell was 500 miles thick; inside were three more concentric shells, being the radii of Venus, Mars, and Mercury. FIGURE 6 shows Halley in a 1736 portrait, holding his hollow Earth sketch. Furthermore, just as the aurora borealis phenomenon illuminates the northern skies, he postulated that the atmospheres over each of these shells were alive with magnetic lightning-like flashes; and he went on to speculate that life existed on these inner surfaces [**7, 8**].

Such musings prompted inane and romantic science fiction tales in years to come. One spurious legacy of Newton's error is a Hollow Earth Society, which apparently loves to speak of UFOs and superior creatures haunting the inner circles of our planet. On their web page [**14**], you may view a purported map of this inner world, shown in an

Figure 6 Halley at 80 with sketch of hollow Earth model

artist's rendition on the cover of this issue. A well-known science fiction tale inspired by Halley's model is by Tarzan's creator, Edgar Rice Burroughs; in *At the Earth's Core*, written in 1908, two heroes within an out-of-control tunneling machine burrow from the surface of the earth, and emerge three days later and 500 miles down in the inner world of *Pellucidar*. The MAGAZINE cover also portrays one charming depiction of this *earth ship*, looking not too distantly different from the great earth-eating machines that tunneled out the Chunnel under the English Channel. At the halfway point, 250 miles deep, up and down switch for Burroughs in the same manner as when Dante climbed down Satan's torso at the center of the Earth. But is such a scenario possible, walking upright on the inside of the Earth's outer layer?

That is, let S be a shell of inner radius r and outer radius R, uniform density δ, mass M, then (6) gives,

$$a(s) = \begin{cases} 0, & \text{if } 0 \le s \le r \\ -\frac{4}{3}\pi G\delta\left(s - \frac{r^3}{s^2}\right), & \text{if } r \le s \le R, \\ -\frac{GM}{s^2} & \text{if } s \ge R. \end{cases} \tag{8}$$

A planet with a hollow core

FIGURE 5b illustrates this acceleration. Note that the entire empty core is a zero-gravity haven!

But on Pellucidar, worse than having zero gravity on the surface, or a marginal attraction due to the spin of the Earth, the situation is grim, for its sky is illuminated by both a moon and a central sun, so any object not clinging to the surface of Pellucidar would fall into this sun. Lest we be overly critical of Burroughs's model, he does have his character sense "a certain airy lightness of step" on Pellucidar [**1**, *Freedom* chapter], and explains, "The force of gravity is less upon ... the inner world ... due

... to the counter-attraction of that portion of the earth ... directly opposite the spot ... at which one's calculations are ... made," which is probably a good rendition of how the typical man-in-the-street might reason.

Hooke's onion model

Robert Hooke is credited with positing an Earth of multiple layers, one reason for which was to explain why the magnetic north pole appears to wander. His idea was that the magnetic source might be embedded in a separate layer, which rotated at a rate slightly different than the surface layer. As a simple representative example of this model, suppose that the Earth consists of two homogeneous layers, the inner one called the *core* and the outer one called the *mantle*, a term coined by E. Wiechert in 1897. Let S be a ball consisting of an inner core of density δ_1, mass M_1, radius r, and of an outer mantle, density δ_2, mass M_2, inner radius r and outer radius R, then two applications of (6) yield

$$a(s) = \begin{cases} -\dfrac{4\pi G\delta_1 s}{3}, & \text{if } 0 \le s \le r, \\ -\dfrac{GM_1}{s^2} - \dfrac{4}{3}\pi G\delta_2\left(s - \dfrac{r^3}{s^2}\right), & \text{if } r \le s \le R \\ -\dfrac{G(M_1 + M_2)}{s^2}, & \text{if } s \ge R. \end{cases} \tag{9}$$

An onion model with two layers

In this case, the mantle contains a local extremum for gravitational acceleration at $s_c = r\sqrt[3]{2(\delta_1 - \delta_2)/\delta_2}$, provided $r < s_c < R$, which means that a necessary condition for the existence of a local extremum is for $\delta_2/\delta_1 < 2/3$. That is, a planet's mantle must be no more than $2/3$ as dense as its core in order for an extreme to exist. See FIGURE 5c where $\delta_1 = 2\delta_2$ and $R = 2r$.

A molten core, variable density mantle model

To account for volcanic activity, geophysicists of the nineteenth century postulated that the Earth's central region was molten, a thick pea soup of convective rock currents, which in turn suggested that the core might be homogeneous. Being solid, the mantle was presumed to have a density that increased with depth. As a simple representative member of this model, let S be an ideal planet with a homogeneous core of density δ, mass M_1, radius r, and a variable mantle of density $\delta(s) = \mu + \lambda s$, where $r < s < R$, mass M_2, and where μ and λ are constants. Then

$$a(s) = \begin{cases} -\dfrac{4}{3}\pi G\delta s, & \text{if } 0 \le s < r, \\ -\dfrac{GM_1}{s^2} - 4\pi G\left(\dfrac{\mu s}{3} + \dfrac{\lambda s^2}{4} - \mu\dfrac{r^3}{3s^2} - \lambda\dfrac{r^4}{4s^2}\right), & \text{if } r \le s \le R, \\ -\dfrac{G(M_1 + M_2)}{s^2}, & \text{if } s \ge R. \end{cases} \tag{10}$$

Homogeneous core with mantle of variable density

FIGURE 5d shows the case when $\delta = 1$, $\mu = 2$, $\lambda = -1$, $r = .82$, and $R = 1.2$. Note that for these values, there are two local extrema within the mantle, the reason being that solving the equation $a'(s) = 0$ for $r < s < R$ involves finding the roots of a quartic polynomial. Nevertheless, it is a bit surprising that $a(s)$ can display such exotic behavior when $\delta(s)$ is so tamely monotonic.

A seismic model

From earthquake analysis and from the understanding that seismic waves are transmitted differently in liquids versus solids, today's thinking suggests that our Earth consists essentially of three layers. We'll simplify matters, and assume that it consists of a solid central core of uniform density $\delta_1 = 13$ units (where a unit is 1000 kg/m^3), radius 1275 km; a liquid outer core of uniform density $\delta_2 = 10$, thickness 2225 km; and an affinely dense mantle of thickness 2900 km, with density varying from 6 to 3.3 at the surface. We ignore the thin crust and tenuous atmosphere. Let $r_1 = 1.275$ units (where a unit is 1000 km), $r_2 = 3.5$, and $R = 6.4$; let M_1, M_2, and M_3 be the masses of the inner core, the outer core, and the mantle respectively. Then within the mantle, $\delta(s) = \mu + \lambda s \approx 9.26 - 0.931s$, where s is in thousands of km, and

$$a(s) = \begin{cases} -\frac{4}{3}\pi G\delta_1 s, & \text{if } 0 \le s < r_1, \\ -\frac{GM_1}{s^2} - \frac{4}{3}\pi G\delta_2\left(s - \frac{r_1^3}{s^2}\right), & \text{if } r_1 \le s \le r_2 \\ -\frac{G(M_1+M_2)}{s^2} - \frac{4}{3}\pi G\left(\mu s + 0.75\lambda s^2 - \mu\frac{r_2^3}{s^2} - 0.75\lambda\frac{r_2^4}{s^2}\right), & \text{if } r_2 \le s \le R, \\ -\frac{G(M_1+M_2+M_3)}{s^2}, & \text{if } s \ge R. \end{cases}$$

Solid center, molten outer core, variable mantle (11)

FIGURE 5e gives a graph of $a(s)$. Note that the Earth has no local extremum within the outer core because $\delta_2/\delta_1 = 10/13 > 2/3$, a violation of the necessary condition given in Hooke's model. The overall extreme value of gravitational acceleration for this model of the Earth is approximately -10.2 m/s^2 at the juncture between the outer core and the mantle. In retrospect, the graph of FIGURE 5e also supports the intuition of those who spun the Greek legends, for Hades could not have been more than 1000 km below Greece; and the acceleration of gravity within this level of the Earth is approximately constant!

Studies from 1989 on the interplay between data and models give a density model as in FIGURE 7, which in turn gives an acceleration curve qualitatively much like FIGURE 5e; the current best guess of gravity's extreme within the Earth is -10.8 m/s^2 at the core-mantle boundary [**9**, p. 155].

A gaseous model

To complete a list of simple planetary models, imagine a gas ball of uniform temperature and uniform composition; the density of such a ball is exponential [**17**, p. 68], that is, $\delta(s) = \delta_o e^{-ks}$, where δ_o is the density at the center and k is a positive constant.

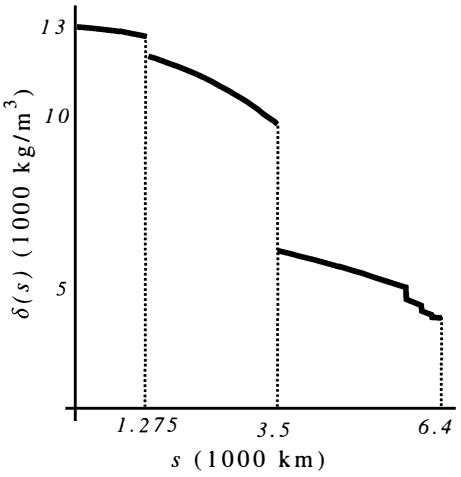


Figure 7 Earth's density

To be faithful to the definition of an ideal planet, we say that $R = \infty$. Then by (6) for positive s,

$$\boxed{a(s) = -4\pi G\delta_o \frac{e^{-ks}(k^2s^2 + 2ks + 2) - 2}{k^3s^2}.} \tag{12}$$

Gaseous model

See FIGURE 5f. The extreme for gravity occurs at $s_c \approx 1.45123/k$, at which point the mass of the ball of radius s_c is $4.50\delta_o/k^3$. (Over 98% of the mass is contained within a ball of radius $3s_c$.) More realistic density models for gas balls are complex and lead to elaborately grand models for the stars.

Concluding remarks

As we have seen, given the global density function for an ideal planet, integration tells us how gravity changes with descent. However, even without knowing the global density function, we can determine whether gravity increases or decreases with descent from a planet's surface; all that is needed is the mean density of the planet and its surface density. Surface density of the Earth is easy enough to compute—just measure some representative surface rock. But in Euler's day determining the mean density of the Earth was not yet possible; the product GM was known, but not its component parts; it was not until 1798, fifteen years after Euler's death, that Henry Cavendish experimentally determined G, which then gave the scientific community M and the mean density of the Earth. The following theorem, foreshadowed by the watershed 2/3 ratio for extrema in the simple Hooke model, shows how local density data can answer the question of how gravity changes with descent. Let planet P have radius R and density $\delta(s)$, $0 \le s \le R$. Let $m(s)$ be the mass of P up to radial distance s; that is, $m(s) = 4\pi \int_0^s \rho^2\delta(\rho)\,d\rho$. Define the *mean density* $\hat{\delta}(s)$ of P at s as $m(s)$ divided by the volume of a sphere of radius s; that is,

$$\hat{\delta}(s) = \frac{3}{s^3}\int_0^s \rho^2\delta(\rho)\,d\rho.$$

Define the *normalized density* $\Delta(s)$ at s as $\Delta(s) = \delta(s)/\hat{\delta}(s)$. For example, if $\delta(s)$ is constant, then P's normalized density is $\Delta(s) = 1$. If as in the constant acceleration model $\delta(s) = k/s$, where k is some constant, then the normalized density of P is $\Delta(s) = 2/3$.

THEOREM: THE CRITICAL CONDITION. *If $a(s)$ is differentiable, then gravity intensifies with descent at s whenever $\Delta(s) < 2/3$ and wanes with descent at s whenever $\Delta(s) > 2/3$. If $\Delta(s) = 2/3$, gravity remains constant with descent at s.*

To verify this condition, note that (6) for $0 \le s \le R$ can be written as

$$-4\pi G \int_0^s \rho^2 \delta(\rho)\, d\rho = s^2 a(s).$$

By the fundamental theorem of calculus,

$$-4\pi G s^2 \delta(s) = 2sa(s) + s^2 a'(s),$$

which means that

$$a'(s) = -4\pi G\delta(s) - \frac{2a(s)}{s}. \tag{13}$$

Note that $a(s)$ intensifies with descent precisely when $a'(s) > 0$; that is, recall that $a(s)$ is negative, so that the phrase "intensifying with descent" translates into the phrase "a is increasing at s." This condition is equivalent, via (6) and (13), to

$$\frac{-4\pi G}{s^2} \int_0^s \rho^2 \delta(\rho)\, d\rho = a(s) < -2\pi G s \delta(s),$$

which is equivalent to $\Delta(s) < 2/3$. Similarly, the alternate conditions follow.

For our Earth, since its mean density $\hat{\delta}(R)$ is about 5515 kg/m^3 and since crustal rock density $\delta(R)$ is about 2800 kg/m^3, then $\Delta(R) \approx 0.51 < 2/3$. So by the theorem, gravity intensifies with descent from the Earth's surface. If the several thousand Moon rock specimens brought back to Earth are accurate indicators, the surface density of the Moon is comparable to that of Earth; since the Moon's mean density is about 3330 kg/m^3, then $\Delta(R) \approx 0.84 > 2/3$, where R is the Moon's radius, which means that gravity weakens with descent from the Moon's surface.

As promised, we now solve the generalized problem of falling through a hole in the Earth:

> *If a hole is drilled through the Earth from pole to pole, and a rock is dropped in the hole, with what velocity will it reach the center, and how long will it take?*

Plutarch posed a version of this problem when he has a debater catalog various reasons as to why some folks chose not to believe in a spherical Earth [**12**, p. 65]:

> *Not that ... masses ... falling through the ... earth stop when they arrive at the centre, though nothing encounter or support them; and, if in their downward motion the impetus should carry them past the centre, they swing back again and return of themselves?*

Euler played with this problem as well, explaining in his popular letters to a German princess [**3**, p. 178] that

You will remember how Voltaire used to laugh at the idea of a hole reaching to the centre of the earth, ... but there is no harm in supposing it, in order to discover what would be the end result.

To solve this problem, we neglect air resistance and any relativistic concerns. Let s, v, a, and t be the distance of the rock from the Earth's center, the velocity of the rock, the acceleration of the rock, and time, respectively, where at $t = 0$, $s = R$, $v = 0$, and $a = -g$, where R is the radius of the Earth and $g \approx 9.8$ m/s^2. Note that acceleration times velocity can be written two ways as

$$v\frac{dv}{dt} = a\frac{ds}{dt}.$$

Integrating this expression as time t goes from 0 to τ gives

$$\int_0^\tau v\frac{dv}{dt}\,dt = \int_0^\tau a\frac{ds}{dt}\,dt.$$

Since both s and v are monotonic functions in terms of t as the rock falls from the Earth's surface to its center and since the initial velocity of the rock is 0, then the above equation simplifies to

$$\frac{v^2(s)}{2} = \int_R^s a\,dr,$$

where v is a function in terms of s, and r is a dummy variable representing distance from the Earth's center. Since the rock's velocity during fall is negative,

$$v(s) = -\sqrt{2\int_R^s a\,dr}. \tag{14}$$

Thus the speed of the rock as it passes through the Earth's center is $v(0 \text{ km})$. To find a formula for the time T it takes for the rock to reach the center, obtain a uniform partition of the interval $[0, R]$: $0 = s_0 < s_1 < \cdots < s_n = R$, so that $\Delta s = R/n$, for each positive integer n. Approximate the time for the rock to fall from s_{i+1} to s_i by $-\Delta s/v_i$, where $v_i = v(s_i)$ for each i, $0 \le i \le n-1$. The approximate value of time T is therefore $\sum_{i=0}^{n-1} -\Delta s/v_i$. That is,

$$T = \int_0^R \frac{-1}{v}\,ds. \tag{15}$$

Case 1. The homogeneous Earth model By (7), $a(s) = ks$, for $0 \le s \le R$, where k is a constant. Since $a(R) = -g$, then $k = -g/R$. Therefore by Equation (14)

$$v(s) = -\sqrt{2\int_R^s \frac{-g}{R}r\,dr} = -\sqrt{\frac{g}{R}(R^2 - s^2)}.$$

Thus $v(0) = -\sqrt{gR} \approx -7.9$ km/s. By Equation (15),

$$T = \int_0^R \frac{1}{-v}\,ds = \int_0^R \sqrt{\frac{R}{g}}\frac{1}{\sqrt{R^2 - s^2}}\,ds = \frac{\pi}{2}\sqrt{\frac{R}{g}} \approx 21.2 \text{ minutes}.$$

Case 2. The constant acceleration model If $a(s) = -g$ for $0 \le s \le R$ then

$$v(s) = -\sqrt{2\int_R^s -g\,dr} = -\sqrt{2g(R-s)},$$

so that $v(0) = -\sqrt{2gR} \approx -11.2$ km/s (which also happens to be the escape velocity of the Earth from its surface), and

$$T = \int_0^R \frac{1}{\sqrt{2g}}\frac{1}{\sqrt{R-s}}\,ds = \sqrt{2}\sqrt{\frac{R}{g}} \approx 19.0 \text{ minutes.}$$

Case 3. An extreme black hole model We now assume all of the Earth's mass is concentrated at its center, so that $a(s) = k/s^2$ by Newton's law of gravitation, for $0 < s < R$ and where k is some constant. As before, since $a(R) = -g$, then $k = -gR^2$. Thus

$$v(s) = -\sqrt{2\int_R^s \frac{-gR^2}{r^2}\,dr} = -\sqrt{2gR\left(\frac{R}{s} - 1\right)},$$

so that $v(0) = -\infty$. Note that $v(1 \text{ cm})$ is just shy of the speed of light. The time it takes for the rock to reach the Earth's center, barring relativistic considerations, is therefore

$$T = \int_0^R \frac{1}{\sqrt{2gR}}\sqrt{\frac{s}{R-s}}\,ds = \frac{\pi}{2\sqrt{2}}\sqrt{\frac{R}{g}} \approx 15.0 \text{ minutes.}$$

To see that no other ideal planet model either yields a faster falling speed or a less fall time, observe that the fastest falling speed in least time will surely occur if the acceleration on the rock is always as extreme as is possible, which certainly happens if all the mass of the Earth is always nearer to the center of the Earth than is the rock. Contrariwise the slowest falling speed and the greatest falling time occur when all the mass is on a shell of radius R, in which case the rock is stationary.

Case 4. Our Earth Using (11), (14), and (15), and wading through all the approximating integrals gives a speed of $v(0) \approx -9.8$ km/s and a time of $T \approx 19.2$ minutes. Modifying the density function to mirror the data of FIGURE 7 or whatever new data geophysicists generate in years to come should produce more accurate acceleration functions, which in turn will influence these values of $v(0)$ and T slightly.

Lucifer cast down We close with a problem, the details of which the reader may enjoy completing. In John Milton's *Paradise Lost*, published in 1667, Satan is cast from Heaven and falls for nine days before landing in Hell.

> *Him (Satan) the Almighty Power*
> *Hurled headlong flaming from th' ethereal sky,*
> *With hideous ruin and combustion, down*
> *To bottomless perdition ...*
>
> Book I, Lines 44–47, [**10**, p. 9].

Although Milton used a Ptolemaic model of the universe, it is a fun exercise to recast this problem as dropping a rock through a galaxy. We shall allow this rock to follow the dynamics of (14) and (15) no matter how large v becomes, since conceivably Satan's speed shouldn't be constrained by that of light. Let's take the galaxy as the

Milky Way, which has a mass M of about 3×10^{41} kg, and whose radius R is about 9.3×10^{20} meters. For the sake of simplicity, assume as a first approximating model that the Milky Way is a flat, homogeneous pancake, which is positioned on the x-y plane with its center at the origin. Assume that the rock falls from rest at $(0, 0, R)$. Integrating (1) over the Milky Way via polar coordinates gives

$$a(s) = \frac{GM}{\pi R^2} \int_0^R \int_0^{2\pi} \frac{-rs}{(r^2 + s^2)^{\frac{3}{2}}} \, d\theta \, dr = \frac{2GM}{R^2} \left(\frac{s}{\sqrt{R^2 + s^2}} - 1 \right).$$

By (14),

$$v(s) = -\frac{2\sqrt{GM}}{R} \sqrt{\sqrt{R^2 + s^2} - s + R(1 - \sqrt{2})}.$$

Note that $v(0)$ is approximately 225 km/s. By (15), the rock will reach the center of the Milky Way in about 340 million years; light would make this journey in about one hundred thousand years. Since the episode of Satan being *hurled* out of Heaven took place when the galaxy was in chaos, possibly before the Big Bang, the reader may wish to toy with the pancake model and the rock's initial velocity, so as to estimate the Miltonian distance between Heaven and Hell.

Acknowledgment. We thank the physics department of King College—Raymond Bloomer and Daniel Cross—for critiquing a first draft of this paper, and we thank the referees for suggestions that significantly improved the readability of the final version.

REFERENCES

1. Edgar Rice Burroughs, *At the Earth's Core*, Dover Publications, 2001.
2. Dante, *The Divine Comedy I: Hell*, translated by Dorothy Sayers, Penguin, 1949. Maps drawn for this edition by C. W. Scott-Giles.
3. Leonard Euler, *Letters to a German Princess*, reprint of the 1833 edition, Arno Press, 1975.
4. R. P. Feynman, R. B. Leighton, M. Sands, *The Feynman Lectures on Physics*, vol. II, Addison-Wesley, 1964.
5. G. R. Fowles and G. L. Cassiday, *Analytical Mechanics*, Saunders, 1999.
6. E. Halley, An account of the cause of the change of the variation of the magnetical needle with an hypothesis of the structure of the internal parts of the Earth, *Philosophical Transactions* **16** (1692), 563–587.
7. N. Kollerstrom, The hollow world of Edmond Halley, *Journal for the History of Astronomy* **23**:3 (1992), 185–192.
8. N. Kollerstrom, Newton's 1702 Lunar Theory, www.ucl.ac.uk/sts/nk/hollow.htm, 2002. Copy of Halley portrait (1736) appears herein.
9. William Lowrie, *Fundamentals of Geophysics*, Cambridge University Press, 1997.
10. John Milton, *Paradise Lost: An Authoritative Text, Backgrounds and Sources, Criticism*, edited by Scott Elledge, W. W. Norton, 1993.
11. Ovid, *Metamorphoses*, translated by Samuel Garth, John Dryden, et al., in The Internet Classics Archive at classics.mit.edu/Ovid/metam.html, 2002.
12. Plutarch, *Plutarch's Moralia*, Vol. XII, translated by H. Cherniss and W. C. Helmbold, Harvard University Press, 1957.
13. G. F. Simmons, *Differential Equations*, McGraw-Hill, 1972.
14. Spiritweb.org, *Map of Agharta*, www2.eu.spiritweb.org/spirit/hollow-earth.html, 2002.
15. Alice Turner, *The History of Hell*, Harcourt Brace, 1993.
16. Jules Verne, *Journey to the Center of the Earth*, Tor Books, 1992. Frontispiece at JV.Gilead.org.il/vt/c_earth, 2001.
17. M. Zeilik and E. v. P. Smith, *Introductory Astronomy and Astrophysics*, Saunders, 1987.

Beyond the Celestial Sphere: Oriented Projective Geometry and Computer Graphics

KEVIN G. KIRBY
Department of Mathematics and Computer Science
Northern Kentucky University
Highland Heights, KY 41099
kirby@nku.edu

Software for computer graphics represents three-dimensional space a little differently than one might expect. Euclidean geometry is not quite right. The usual approach uses what is called *projective geometry*, certainly one of the most beautiful systems in mathematics. Yet even with this approach, when the mathematics actually meets the computer code there are some awkward inconveniences. Perhaps the best solution may be what is called *oriented projective geometry*. This geometry was worked out in detail by Jorge Stolfi [**10, 11**] in 1987; it has also found more recent application in computer vision [**6**]. This paper is an elementary introduction to this still unfamiliar geometry from a coordinate-based point of view, restricted to three dimensions. It assumes only a background in linear algebra.

For the reader who knows of classical projective geometry and homogeneous coordinates, it is best to set this knowledge aside and examine whether the oriented approach works on its own terms. At the end of the paper we compare and contrast it with the classical approach. Also, readers conversant with wedge products, Plücker coordinates, and the Hodge operator may be pleased to see them appear in such a concrete setting.

To infinity ...

You are driving a virtual car in a computer game. Look out through the windshield. The trees that line the highway are rushing past you. They are nicely displayed in perspective: they begin far away as dots, but they grow taller as you race toward them. Next, look ahead at the horizon. You see the sun. Unlike the trees, the sun never gets closer to you. Still, it is certainly subject to some transformations: you turn your car left, and the sun veers right.

The trees and the sun need to be represented inside the program somehow. Ultimately, this depends on attaching parts of them to points in a virtual space. One might think at first that a useful way to represent any point would be to use Cartesian coordinates, a triple (x, y, z) of real numbers. Taking your car to be located at the origin, a tree might be centered at, say, $(-20.42, 10.63, -94.37)$. Where is the sun? It is very far away, perhaps at $(9.34 \times 10^9, 2.71 \times 10^9, -1.23 \times 10^{11})$. But there is something strange about these large numbers. It seems pointless to waste time (and numerical precision) decrementing such big numbers by a few tens as our car drives on toward the sunset. We would like to simplify things by somehow locating the sun *at infinity*.

One uniform way to represent both ordinary points and points at infinity is to use four numbers instead of three. Here's how it works. Take the point (2, 3, 4). Instead of representing it as a column vector in $\mathbb{R}^3$, we tack a 1 on the end and represent it as a column vector in $\mathbb{R}^4$: $\mathbf{x} = [2\,3\,4\,1]^{\mathrm{T}}$. This representation is not meant to be unique:

we can multiply this column vector by any positive number and we will say it represents the same point. (The adjective *positive* is crucial here.) For example, the vectors $[2\,3\,4\,1]^{\mathrm{T}}$, $[20\,30\,40\,10]^{\mathrm{T}}$, and $[0.2\;\;0.3\;\;0.4\;\;0.1]^{\mathrm{T}}$ all represent the point $(2, 3, 4)$. Notice the fourth coordinate will always be positive.

Going in the other direction, to get the 3D coordinates of the point represented by the column vector $\mathbf{v} = [v_1\, v_2\, v_3\, v_4]^{\mathrm{T}}$ with $v_4 > 0$, we just multiply by $(1/v_4)$. This puts a 1 in the fourth slot, and we can extract the first three components as the coordinates of our point. This process takes a vector $\mathbf{v}$ and yields the point

$$\boldsymbol{p} = \left(\frac{v_1}{v_4}, \frac{v_2}{v_4}, \frac{v_3}{v_4}\right) = [\mathbf{v}]. \tag{1}$$

(The square brackets are meant to suggest that points are equivalence classes of vectors.) The convention we will use is that bold lowercase letters denote vectors in $\mathbb{R}^4$, while bold italic lowercase letters denote points in three-dimensional space. Exactly what *kind* of space this is has yet to be determined.

Now consider the sequence of collinear points

$$(0.2, 0.3, 0.4), (2, 3, 4), (20, 30, 40), (200, 300, 400), \ldots$$

which moves outward from the origin through $(2, 3, 4)$ and continues on ... forever. Here is one nice way our scheme can represent this sequence using vectors in $\mathbb{R}^4$:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 10 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \\ 0.1 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \\ 0.01 \end{bmatrix}, \ldots, \begin{bmatrix} 2 \\ 3 \\ 4 \\ 0 \end{bmatrix}. \tag{2}$$

But here we have appended a limit point, one that is naturally suggested by this four-dimensional representation. Where is it located in space? Since $v_4 = 0$ we cannot use (1) to give it (x, y, z) coordinates. In fact it is natural to say that the point is at infinity in the direction outward from the origin past $(2, 3, 4)$. We say it lies *on the celestial sphere*, a spherical dome enclosing our Euclidean space at infinite distance.

Because points at infinity in some sense denote directions, let us introduce a bit of notation: we will write $(x, y, z)^{\uparrow}$ for the point at infinity represented by $[x\, y\, z\, 0]^{\mathrm{T}}$, where at least one of x, y, and z is nonzero. Our rule that one can multiply a vector by any positive constant and still have a representation of the same point is still in effect; for example, the vectors $[20\,30\,40\,0]^{\mathrm{T}}$ and $[200\,300\,400\,0]^{\mathrm{T}}$ both represent the same point $(2, 3, 4)^{\uparrow}$ on the celestial sphere.

The celestial sphere is indeed a sphere, and we can distinguish antipodal points, that is, points on the opposite side of the origin. We can arrive at the point antipodal to $(2, 3, 4)^{\uparrow}$ on the celestial sphere by the sequence

$$(-2, -3, -4), (-20, -30, -40), (-200, -300, -400), \ldots, (-2, -3, -4)^{\uparrow}$$

which, in one vector representation, is

$$\begin{bmatrix} -2 \\ -3 \\ -4 \\ 1 \end{bmatrix}, \begin{bmatrix} -2 \\ -3 \\ -4 \\ 0.1 \end{bmatrix}, \begin{bmatrix} -2 \\ -3 \\ -4 \\ 0.01 \end{bmatrix}, \ldots, \begin{bmatrix} -2 \\ -3 \\ -4 \\ 0 \end{bmatrix}.$$

If $\boldsymbol{p}$ is the limit point in (2), then this limit point should certainly be denoted $-\boldsymbol{p}$. These two antipodal points on the celestial sphere are (needless to say?) different.

... And beyond?

We now have seen how to use vectors in $\mathbb{R}^4$ to represent points in Euclidean space as well as points at infinity in all directions. But, for better or worse, there's more.

We have been moving in a straight line by fixing the first three coordinates of a vector in $\mathbb{R}^4$ and varying the fourth. Let's continue moving along the direction of sequence (2) and see where it leads. We hit the celestial sphere at $[2\,3\,4\,0]^{\mathrm{T}}$ *and keep on going*:

$$\cdots \begin{bmatrix} 2 \\ 3 \\ 4 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \\ -0.01 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \\ -0.1 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \\ -1 \end{bmatrix}. \tag{3}$$

Let's stop at this place for a moment. Where are we now? The vector $\mathbf{v} = [2\,3\,4\,-1]^{\mathrm{T}}$ does *not* represent the point $(2, 3, 4)$, since our vector representation only allows us to multiply by positive numbers, and we cannot thereby change the fourth coordinate to $+1$. Neither does it represent the point $(-2, -3, -4)$, for the same reason. So our $\mathbb{R}^4$ representation seems to imply the existence of a whole new, distinct Euclidean space of points with *negative* numbers in the fourth coordinate of their vector representations. As before we can normalize such vectors by multiplying by positive numbers to ensure $v_4 = -1$, but we can never get rid of the negative sign.

We will call this duplicate Euclidean space *invisible space*; its points are represented by vectors in $\mathbb{R}^4$ with $v_4 < 0$. The Euclidean space where we began will be known as *visible* space; vectors representing its points have $v_4 > 0$. These names are chosen to evoke their roles in graphics and computer vision, roles that we will discover later. As far as the mathematics is concerned, it is important only to realize that there are now *two copies* of Euclidean space and we need to give them different names to keep them straight.

The point represented by $\mathbf{v} = [2\,3\,4\,-1]^{\mathrm{T}}$ and its positive multiples will be denoted by $(-2, -3, -4)^{\circ}$. To justify this sign change, let us review the journey shown in sequences (2) and (3) above. We have moved out of visible Euclidean space and hit the celestial sphere; we have kept going and crossed into *invisible* Euclidean space, heading toward its origin. We left visible space in the $+++$ octant; with this sign convention we enter invisible space in the $---$ octant. If we kept going, we would cross the invisible origin and exit invisible space at the point antipodal to the place we entered. We thereupon would reenter visible space. This is depicted in FIGURE 1.

One might worry that when our interspatial tour neared the origin, if we had clamped v_1, v_2, v_3 constant and let v_4 grow, then v_4 would blow up. This is true. But sometime before we reach the origin, we clamp v_4 constant and let v_1, v_2, v_3 start to shrink uniformly down to zero. For example, instead of $[2\,3\,4\,-1000]^{\mathrm{T}}$ we use $[0.002\ \ 0.003\ \ 0.004\ \ -1]^{\mathrm{T}}$ to represent $(-0.002\ \ -0.003\ \ -0.004)^{\circ}$. Once we are safely past the origin, we can fix v_1, v_2, v_3 and let v_4 vary again, so it can go zero in preparation for hitting the celestial sphere. This is called a *change of coordinate chart* in topology. It should be no more disturbing than changing maps on a road trip as you approach a state line.

In all, our space is composed of three connected pieces with the following types of points:

$\{k[x\ y\ z\ 1]^{\mathrm{T}} \mid k > 0\}$, representing a visible Euclidean point (x, y, z)

$\{k[x\ y\ z\ 0]^{\mathrm{T}} \mid k > 0\}$, representing a point on the celestial sphere $(x, y, z)^{\uparrow}$

$\{k[x\ y\ z\ -1]^{\mathrm{T}} \mid k > 0\}$, representing an invisible Euclidean point $(-x, -y, -z)^{\circ}$

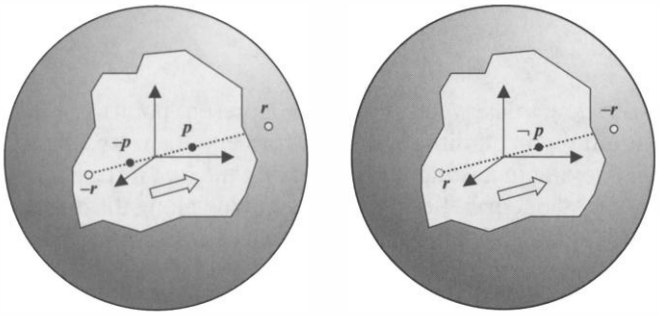


Figure 1 Visible (*left*) and invisible (*right*) space share the same celestial sphere. Upon exiting the visible world by crossing the celestial sphere, we immediately enter the invisible world. Continuing farther along, we exit at the antipodal point and re-enter the visible world. For concreteness, imagine $\boldsymbol{p} = (2, 3, 4) = \left[[2\,3\,4\,1]^{\mathrm{T}}\right]$ and $\boldsymbol{r} = (2, 3, 4)^{\uparrow} = \left[[2\,3\,4\,0]^{\mathrm{T}}\right]$. Then $(2, 3, 4)^{\circ} = \left[[-2\,-3\,-4\,-1]^{\mathrm{T}}\right]$, which we will call $\neg\boldsymbol{p}$, and $-\boldsymbol{r} = \left[[-2\,-3\,-4\,0]^{\mathrm{T}}\right] = (-2, -3, -4)^{\uparrow}$.

If $\boldsymbol{p} = [\mathbf{v}]$, we may define $\neg\boldsymbol{p} = [-\mathbf{v}]$. This is different from the geometric opposite $-\boldsymbol{p}$ of a point $\boldsymbol{p}$, which would be represented by $[-x\,-y\,-z\,k]^{\mathrm{T}}$ if $\boldsymbol{p}$ is represented by $[x\,y\,z\,k]^{\mathrm{T}}$. Note that we have the properties

$$\neg\neg\boldsymbol{p} = \boldsymbol{p} \quad \text{and} \quad \neg{-}\boldsymbol{p} = -\neg\boldsymbol{p}.$$

If (and only if) $\boldsymbol{r}$ is a point at infinity, then $\neg\boldsymbol{r} = -\boldsymbol{r}$. We will call $\neg\boldsymbol{p}$ the *reverse* of point $\boldsymbol{p}$, anticipating the generalization of this operation to planes and lines in the next two sections.

Which side are you on?

At this point, we seem to have moved from science to science fiction. With only some vague motivation from computer graphics (that is, the usefulness of points at infinity), we have introduced a vastly redundant four-dimensional representation for three-dimensional space. Even stranger, we have let this representation drive us to include another (invisible!) universe just as big as the original one. The latter seems especially profligate. Why would one do such a thing?

Well, all this geometric baggage is a consequence of our need to pay attention to signs. If we ignore the sign of the four-dimensional vectors that represent our points, we lose important information about orientation. Let us first see how this works in a simple case.

Consider the problem of determining whether a point $\boldsymbol{p}$ is on, in front of, or behind a plane $\boldsymbol{N}$. We specify the plane with the four numbers $(\hat{\boldsymbol{n}}, d)$, with $\hat{\boldsymbol{n}} = (n_1, n_2, n_3)$ the components of a unit vector normal to $\boldsymbol{N}$, and d its signed distance from the origin measured in the $\hat{\boldsymbol{n}}$ direction. The vector $\hat{\boldsymbol{n}}$ gives an orientation to the plane; we will say its front side faces out in the direction of $\hat{\boldsymbol{n}}$. The point $\boldsymbol{p} = (x, y, z)$ is on the plane exactly when $n_1x + n_2y + n_3z = d$. Since $\boldsymbol{p}$ is represented by the column vector $\mathbf{p} = [x\,y\,z\,1]^{\mathrm{T}}$, it is convenient to represent $\boldsymbol{N}$ by the row vector $\mathbf{n}^{\mathrm{T}} = [n_1\,n_2\,n_3\,-d]$. Then we have a simple test: the point is on, in front of, or behind the plane if $\mathbf{n}^{\mathrm{T}}\mathbf{p} = 0$, $\mathbf{n}^{\mathrm{T}}\mathbf{p} > 0$, or $\mathbf{n}^{\mathrm{T}}\mathbf{p} < 0$, respectively. Clearly, any positive multiple of $\mathbf{n}^{\mathrm{T}}$ will represent the same plane.

If all we care about is *incidence*, that is, whether a point is on the plane or not, we can be completely indifferent to the overall sign of $\mathbf{p}$ (or of $\mathbf{n}$): $\mathbf{n}^{\mathrm{T}}\mathbf{p} = 0$ when-

ever $\mathbf{n}^{\mathrm{T}}(-\mathbf{p}) = 0$. But if we want to distinguish the front and back sides of the plane we must pay attention to signs. Switching $\mathbf{p}$ to $-\mathbf{p}$ reverses the inequalities and thus reverses the answers to our front/back questions.

The vector representation of planes enjoys a fascinating duality with the representation of points. If $\mathbf{n}^{\mathrm{T}} = [n_1\ n_2\ n_3\ -d]$ represents the plane $\boldsymbol{N}$, the plane represented by $-\mathbf{n}^{\mathrm{T}} = [-n_1\ -n_2\ -n_3\ d]$ represents a plane occupying the same location but facing the opposite direction; call it $\neg\boldsymbol{N}$, the *reverse* of $\boldsymbol{N}$. The sign of the fourth coordinate of $\mathbf{n}^{\mathrm{T}}$ is positive if the plane is facing inward (toward the visible origin), and negative if it is facing outward (toward the invisible origin). Both $\boldsymbol{N}$ and $\neg\boldsymbol{N}$ cut through the celestial sphere and pass through both visible and invisible spaces. If a point $\boldsymbol{p}$ is on plane $\boldsymbol{N}$ then $\boldsymbol{N}$ also meets $\neg\boldsymbol{p}$ as the plane extends across invisible space. The rather exceptional looking row vectors, $\mathbf{n}^{\mathrm{T}} = [0\,0\,0\ \pm 1]$, represent the (inward and outward) faces of the celestial sphere: their product with a vector $\mathbf{p}$ is zero if and only if $\mathbf{p}$ has a 0 in its fourth coordinate, that is, if and only if $\mathbf{p}$ represents a point at infinity.

Head-on collisions

Having described sign-conscious representations of points and planes, we should round things out by considering lines as well. When (directed) lines pierce (oriented) planes the direction of the encounter might matter in certain applications, so we pay attention to it here. We prove an exceptionally simple formula summarizing this encounter in all generality. This formula requires one preliminary definition: the *wedge product* of two vectors $\mathbf{q}, \mathbf{p} \in \mathbb{R}^4$ is defined as the 4×4 antisymmetric matrix $\mathbf{q} \wedge \mathbf{p} = \mathbf{q}\mathbf{p}^{\mathrm{T}} - \mathbf{p}\mathbf{q}^{\mathrm{T}}$.

PROPOSITION. *The line running from the point represented by* $\mathbf{p}$ *to the point represented by* $\mathbf{q}$ *is represented by all positive multiples of the matrix* $\mathbf{L} = \mathbf{q} \wedge \mathbf{p}$. *A plane* $\mathbf{n}^{\mathrm{T}}$ *intersects the line at the point* $\mathbf{x} = \mathbf{L}\mathbf{n}$, *unless the line lies in the plane, in which case* $\mathbf{L}\mathbf{n} = \mathbf{0}$.

From here on, "the point represented by positive multiples of the four-dimensional vector $\mathbf{p}$" is shortened to "the point $\mathbf{p}$"; similarly for planes and lines. An italic bold letter like $\boldsymbol{p}$ will still represent the point itself and, if the point is a visible Euclidean point, it will denote the triple (p_1, p_2, p_3) considered as a vector in $\mathbb{R}^3$, admitting dot and cross products.

The power of this proposition is that it works even when the line runs parallel to the plane: the intersection point will then lie on the celestial sphere. In fact, one or both of the two points, as well as the plane itself, could lie at infinity, and the formula still yields the correct answer. In addition, the visibility/invisibility of the intersection point $\mathbf{x}$ will give us information about the relative orientation of the line and plane. Finally, in the case where the line lies entirely in the plane the formula returns $\mathbf{x} = \mathbf{0}$—a natural error code, as it were.

Rather than proving this proposition from scratch, we build on results presented by Guggenheimer [4] in his survey of geometric formulas in $\mathbb{R}^3$. We first use $\boldsymbol{p}$ and $\boldsymbol{q}$ to define two related vectors: $\boldsymbol{h} = \boldsymbol{q} - \boldsymbol{p}$ and $\boldsymbol{m} = \boldsymbol{p} \times \boldsymbol{q}$. Since $\boldsymbol{h}$ runs in the direction of the line from $\boldsymbol{p}$ to $\boldsymbol{q}$, we call it the *heading vector*. The vector $\boldsymbol{m}$ is orthogonal to $\boldsymbol{h}$, and indeed is perpendicular to the plane containing the line and the origin; it is called the *moment vector*. Taken together, these two vectors comprise six numbers called *Plücker coordinates* of the line. Guggenheimer shows that the Plücker coordinates completely characterize the line—though he ignores its direction. He then goes on to derive a formula for line-plane intersections:

LEMMA. *Let* $\boldsymbol{h} = \boldsymbol{q} - \boldsymbol{p}$ *and* $\boldsymbol{m} = \boldsymbol{p} \times \boldsymbol{q}$. *The plane* $[\hat{\boldsymbol{n}} - d]^{\mathrm{T}}$ *intersects the line containing* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *if and only if* $\boldsymbol{h} \cdot \hat{\boldsymbol{n}} \neq 0$; *the point of intersection is given by*

$$\boldsymbol{x} = \frac{(\hat{\boldsymbol{n}} \times \boldsymbol{m}) + d\boldsymbol{h}}{\boldsymbol{h} \cdot \hat{\boldsymbol{n}}}.$$

Readers of Guggenheimer's paper should note the plane whose coordinates he gives as $(\hat{\boldsymbol{n}}, d)$ is the same as the one we denote by $[\hat{\boldsymbol{n}} - d]$ (facing forward) or $-[\hat{\boldsymbol{n}} - d]$ (facing backward). Any positive multiple of $[\hat{\boldsymbol{n}} - d]$ will describe the same plane, of course, but using the unit vector $\hat{\boldsymbol{n}}$ will make the arguments below easier to follow, as d is then the distance from the origin to the plane measured along $\hat{\boldsymbol{n}}$.

To see that our proposition is an extension of this lemma, consider first the case where $\mathbf{p}$ and $\mathbf{q}$ represent visible Euclidean points: take $\mathbf{p} = [\boldsymbol{p}\ 1]^{\mathrm{T}}$ and $\mathbf{q} = [\boldsymbol{q}\ 1]^{\mathrm{T}}$. $\mathbf{L} = \mathbf{q} \wedge \mathbf{p}$ is then

$$\mathbf{L} = \begin{bmatrix} 0 & p_2q_1 - p_1q_2 & p_3q_1 - p_1q_3 & q_1 - p_1 \\ - & 0 & p_3q_2 - p_2q_3 & q_2 - p_2 \\ - & - & 0 & q_3 - p_3 \\ - & - & - & 0 \end{bmatrix} = \begin{bmatrix} 0 & -m_3 & m_2 & h_1 \\ - & 0 & -m_1 & h_2 \\ - & - & 0 & h_3 \\ - & - & - & 0 \end{bmatrix}.$$

The minus signs displayed below the diagonal are intended to reduce clutter, and indicate that this is an antisymmetric matrix ($\mathbf{L} = -\mathbf{L}^{\mathrm{T}}$), the (i, j) entries below the diagonal being the negatives of their (j, i) counterparts above.

If $\mathbf{n}^{\mathrm{T}} = [\hat{\boldsymbol{n}}\ -d]$ then, writing the first three entries together as a vector, we have

$$\mathbf{Ln} = \begin{bmatrix} (\boldsymbol{m} \times \hat{\boldsymbol{n}}) - d\boldsymbol{h} \\ -\boldsymbol{h} \cdot \hat{\boldsymbol{n}} \end{bmatrix}. \tag{4}$$

If the heading vector points in the opposite direction from the plane normal, then $\boldsymbol{h} \cdot \hat{\boldsymbol{n}} < 0$, so upon dividing by the positive fourth component we see that $\mathbf{Ln}$ represents the visible point $\boldsymbol{x}$ given by the formula in the lemma. On the other hand, if $\boldsymbol{h} \cdot \hat{\boldsymbol{n}} > 0$, the intersection point $\boldsymbol{x}$ has the same coordinates but lies in *invisible* space. What does this tell us? The $\mathbf{x} = \mathbf{Ln}$ formula gives us only the point of *head-on* intersection, where the line and the plane are facing each other. If it happens that, in the visible universe, a line and a plane encounter each other in the same direction (the heading and the normal facing the same way, $\boldsymbol{h} \cdot \hat{\boldsymbol{n}} > 0$), their actual point of head-on intersection will lie in the invisible universe. See FIGURE 2, *top*.

The cases where one of $\mathbf{p}$ or $\mathbf{q}$ is at infinity pose no special problems. For example, if we have $\mathbf{p} = [\boldsymbol{p}\ 1]^{\mathrm{T}}$ and $\mathbf{q} = [\boldsymbol{q}\ 0]^{\mathrm{T}}$, we get the same $\mathbf{L} = \mathbf{q} \wedge \mathbf{p}$ as if we had chosen the two points $\mathbf{p}$ and $[\boldsymbol{p} + \boldsymbol{q}\ 1]^{\mathrm{T}}$, which clearly determine the same line. Here the vector $\boldsymbol{q}$ serves as the heading vector. The case where the plane itself is the celestial sphere ($\mathbf{n}^{\mathrm{T}} = [0\,0\,0\,{-1}]$ for the outward face, or $\mathbf{n}^{\mathrm{T}} = [0\,0\,0\,1]$ for the inward) tells us the line hits it head on at the point at infinity in direction $-\boldsymbol{h}$ or $\boldsymbol{h}$, respectively, which is what we would expect.

We turn now to the case where the line and the plane are parallel. In formula (4) we have $\boldsymbol{h} \cdot \hat{\boldsymbol{n}} = 0$, so the intersection point is claimed to be $\mathbf{x} = \mathbf{Ln} = [(\boldsymbol{m} \times \hat{\boldsymbol{n}}) - d\boldsymbol{h}\ 0]^{\mathrm{T}}$. This is a point at infinity, which is to be expected. But why these specific coordinates? There are two cases to consider. The line may run behind the plane, or in front of it. Our formula actually tells us which of these two alternatives occurs. To see this, we need to do some vector algebra.

Let $\boldsymbol{f}$ denote the *foot vector*, which runs from the origin to the line along the shortest distance separating them; it is perpendicular to the heading $\boldsymbol{h}$. We can express points $\boldsymbol{p}$ and $\boldsymbol{q}$ as $\boldsymbol{p} = \boldsymbol{f} + \lambda_1\boldsymbol{h}$ and $\boldsymbol{q} = \boldsymbol{f} + \lambda_2\boldsymbol{h}$, where $\lambda_2 > \lambda_1$. Then $\boldsymbol{m} = \boldsymbol{p} \times \boldsymbol{q} =$

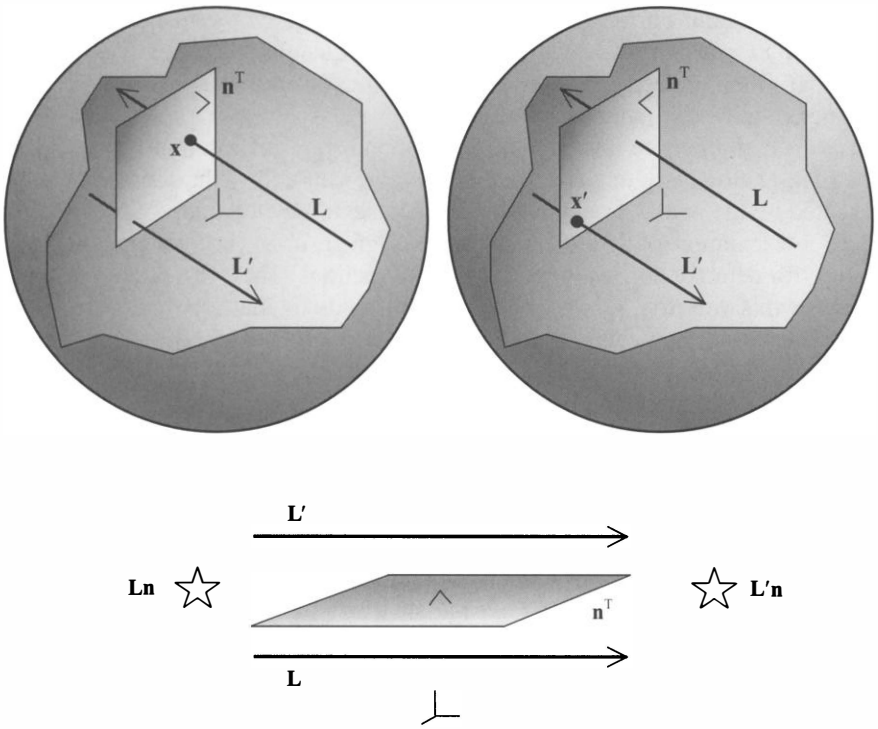


Figure 2 Line-plane intersections. The formula $\mathbf{x} = \mathbf{Ln}$ gives the point where the directed line $\mathbf{L}$ intersects the oriented plane $\mathbf{n}^T$ in a head-on fashion. Unless the line lies in the plane (in which case $\mathbf{Ln} = \mathbf{0}$), this collision will always happen at exactly one point. It may be in visible space (top left), invisible space (top right), or on the celestial sphere. The latter will happen when the line runs parallel to the plane in Euclidean space. This is depicted in the illustration at the bottom. $\mathbf{L}$ intersects the plane head on at the point at infinity on the left; $\mathbf{L}'$ hits it at infinity on the right. The coordinates of these points on the celestial sphere are $[-\boldsymbol{h}\,0]^T$ and $[\boldsymbol{h}\,0]^T$, respectively, where $\boldsymbol{h}$ is the heading vector (the same for both lines shown). Note that all planes may be viewed as spheres of infinite diameter, just as lines are circles of infinite diameter.

$(\lambda_2 - \lambda_1)\boldsymbol{f} \times \boldsymbol{h}$, and so $(\boldsymbol{m} \times \hat{\boldsymbol{n}}) - d\boldsymbol{h} = (\lambda_2 - \lambda_1)(\boldsymbol{f} \times \boldsymbol{h}) \times \hat{\boldsymbol{n}} - d\boldsymbol{h} = (\lambda_2 - \lambda_1)(\hat{\boldsymbol{n}} \cdot \boldsymbol{f} - d)\boldsymbol{h}$, where we have used the identity $(\boldsymbol{f} \times \boldsymbol{h}) \times \hat{\boldsymbol{n}} = (\boldsymbol{f} \cdot \hat{\boldsymbol{n}})\boldsymbol{h} - (\boldsymbol{h} \cdot \hat{\boldsymbol{n}})\boldsymbol{f}$ and the fact that $\boldsymbol{h} \cdot \hat{\boldsymbol{n}} = 0$. We see that our $\mathbf{x} = \mathbf{Ln}$ formula tells us that the point on the celestial sphere where $\mathbf{L}$ intersects $\mathbf{n}^T$ is

$$\mathbf{x} = \begin{bmatrix} sign(\hat{\boldsymbol{n}} \cdot \boldsymbol{f} - d)\boldsymbol{h} \\ 0 \end{bmatrix},$$

where we have taken the liberty of multiplying by the positive scalar $[(\lambda_2 - \lambda_1)|\hat{\boldsymbol{n}} \cdot \boldsymbol{f} - d|]^{-1}$.

When $\hat{\boldsymbol{n}} \cdot \boldsymbol{f} > d$ the line passes in front of the plane, and hits it head on at infinity in the direction $\boldsymbol{h}$. When $\hat{\boldsymbol{n}} \cdot \boldsymbol{f} < d$ the line passes behind the plane, it hits it head on at infinity in the direction $-\boldsymbol{h}$. These are indeed the expected intersection points; see FIGURE 2, *bottom*. If the line lies entirely in the plane, then $\hat{\boldsymbol{n}} \cdot \boldsymbol{f} = d$ and we have $\mathbf{x} = \mathbf{0}$, which represents no point.

Finally, consider the case where both $\mathbf{p}$ and $\mathbf{q}$ are on the celestial sphere. The line from $\mathbf{p}$ through $\mathbf{q}$ is a great circle in the heavens. If we are sitting on the front face of the plane $[\hat{\boldsymbol{n}} \; -d]$ looking up at this circle, the point of intersection is the where the

circle encounters our horizon head on. This should be given by the point at infinity in the direction $\boldsymbol{m} \times \hat{\boldsymbol{n}}$. Fortunately, substituting these quantities into $\mathbf{x} = \mathbf{L}\mathbf{n} = (\mathbf{q} \wedge \mathbf{p})\mathbf{n}$ reveals that this is indeed the case. This completes our derivation of the Proposition from Guggenheimer's Lemma.

If $\mathbf{L} = \mathbf{q} \wedge \mathbf{p}$, then the line running in the opposite direction is $\mathbf{p} \wedge \mathbf{q} = -\mathbf{L}$. This means that, analogously to points and planes, we can define the *reverse* of a line L represented by the matrix $\mathbf{L}$ as the line $\neg L$ represented by the matrix $-\mathbf{L}$.

The representation of $\mathbf{L}$ by $\mathbf{q} \wedge \mathbf{p}$ has thus given us a very concise and general algorithm for detecting oriented line/plane intersections. There are other questions we can answer this way too. Guggenheimer [4] reminds us that a point $\boldsymbol{y}$ is on the line with heading $\boldsymbol{h}$ and moment $\boldsymbol{m}$ if and only if $\boldsymbol{y} \times \boldsymbol{h} = \boldsymbol{m}$. Direct substitution shows that this amounts to checking that $\star\mathbf{L}\mathbf{y} = \mathbf{0}$, where we define

$$\star\mathbf{L} = \begin{bmatrix} 0 & h_3 & -h_2 & -m_1 \\ - & 0 & h_1 & -m_2 \\ - & - & 0 & -m_3 \\ - & - & - & 0 \end{bmatrix}.$$

This matrix results from $\mathbf{L}$ by simply swapping the positions of the heading and moment vectors. This is not some *ad hoc* shuffling of matrix entries, however. We are being quite systematic, using what is called the *Hodge star* operation, defined component-wise this way:

$$(\star\mathbf{L})_{ij} \equiv \frac{1}{2} \sum_k \sum_l signature(i, j, k, l)(\mathbf{L})_{kl},$$

where $signature(i, j, k, l)$ is defined as $+1$ (resp. -1) if (i, j, k, l) is an even (resp. odd) permutation of $(1, 2, 3, 4)$; it is 0 if there are repeated indices.

If $\star\mathbf{L}\mathbf{y} = \mathbf{0}$ for any point $\mathbf{y}$ on the line $\mathbf{L}$, what does this vector represent when $\mathbf{y}$ is not on $\mathbf{L}$? A line and a point not on the line determine a plane, so perhaps $\mathbf{n} = \star\mathbf{L}\mathbf{y}$ gives us this very plane? Yes. The proof is straightforward. A point $\mathbf{x}$ is on this plane if and only if it lies along a line connecting $\mathbf{y}$ to some point $\mathbf{p}$ on $\mathbf{L}$. And a point is on such a line if and only if it has the form $\mathbf{x} = a\mathbf{y} + b\mathbf{p}$ for some real a, b not both zero. (To see this, consider $a[\boldsymbol{y}\ 1]^{\mathrm{T}} + b[\boldsymbol{p}\ 1]^{\mathrm{T}}$ as a and b vary.) To show that $\mathbf{x}$ is on the plane $\mathbf{n}^{\mathrm{T}}$ we just need to verify that $\mathbf{n}^{\mathrm{T}}\mathbf{x}$ is zero:

$$\mathbf{n}^{\mathrm{T}}\mathbf{x} = (\star\mathbf{L}\mathbf{y})^{\mathrm{T}}(a\mathbf{y} + b\mathbf{p}) = a\mathbf{y}^{\mathrm{T}}(\star\mathbf{L})^{\mathrm{T}}\mathbf{y} + b\mathbf{y}^{\mathrm{T}}(\star\mathbf{L})^{\mathrm{T}}\mathbf{p} = 0.$$

The first term is zero because $\star\mathbf{L}$ is antisymmetric. The second term is zero because $\mathbf{p}$ is on the line $\mathbf{L}$, ensuring that $(\star\mathbf{L})^{\mathrm{T}}\mathbf{p} = -\star\mathbf{L}\mathbf{p} = 0$. To summarize, we have:

PROPOSITION. *A point* $\mathbf{y}$ *is on* $\mathbf{L}$ *if and only if* $\star\mathbf{L}\mathbf{y} = 0$. *The plane* $\mathbf{n}^{\mathrm{T}}$ *determined by a line* $\mathbf{L}$ *and a point* $\mathbf{y}$ *not on* $\mathbf{L}$ *is given by* $\mathbf{n} = \star\mathbf{L}\mathbf{y}$.

The plane $\star\mathbf{L}\mathbf{y}$ comes with a well-defined orientation. From the previous section, we know the sign of its fourth component tells us which way it is facing. Taking $\mathbf{L} = \mathbf{q} \wedge \mathbf{p}$, we have $n_4 = \boldsymbol{m} \cdot \boldsymbol{y} = (\boldsymbol{p} \times \boldsymbol{q}) \cdot \boldsymbol{y}$, which is the triple scalar product of $\boldsymbol{p}$, $\boldsymbol{q}$ and $\boldsymbol{y}$; it flips its sign under odd permutations of these points.

If we put the two propositions together and apply some linear algebra, a few interesting consequences emerge. First, note that $\star\mathbf{L}\mathbf{L}\mathbf{n} = \mathbf{0}$ for all $\mathbf{n}$, that is, $\star\mathbf{L}\mathbf{L} = \mathbf{0}$, which says the point of intersection of a line with any plane lies on that line. The other way around, $\mathbf{L} \star \mathbf{L} = \mathbf{0}$, which says that a line lies in the plane determined by that line and any point. The propositions also tell us we can recover the set of points on a line

from its matrix $\mathbf{L}$ in two ways: as the kernel of $\star\mathbf{L}$, or, equivalently, as the image of $\mathbf{L}$. Turning this around as well, the set of planes containing a given line (called a *pencil* of planes by geometers) is given by $ker\mathbf{L} = im\ \star\mathbf{L}$. Since these kernels are clearly more than just $\{\mathbf{0}\}$, the matrices $\mathbf{L}$ and $\star\mathbf{L}$ must be singular. This can also be verified by noting that $\det\mathbf{L} = \det\ \star\mathbf{L} = \boldsymbol{h}\cdot\boldsymbol{m} = 0$, because heading and moment vectors are orthogonal.

It is irresistible to point out that the algebraic patterns revealed above are tantalizingly similar to what one finds in the mathematics of electromagnetism. There, the electric field vector $\boldsymbol{E}$ and the magnetic field vector $\boldsymbol{B}$ may be piled into the matrix

$$\mathbf{F} = \begin{bmatrix} 0 & -B_3 & B_2 & E_1 \\ - & 0 & -B_1 & E_2 \\ - & - & 0 & E_3 \\ - & - & - & 0 \end{bmatrix},$$

the so-called *Faraday tensor*. Just as properties of lines expressed via $\boldsymbol{h}$ and $\boldsymbol{m}$ vectors in $\mathbb{R}^3$ can be more concisely stated in terms of $\mathbf{L}$ and $\star\mathbf{L}$, properties of the electromagnetic field can be more concisely stated in terms of $\mathbf{F}$ and $\star\mathbf{F}$. (In the latter case, the fourth dimension is the physical dimension of time.) Maxwell's four equations, the fundamental laws of electricity and magnetism, specify certain derivatives of $\boldsymbol{E}$ and $\boldsymbol{B}$ in three dimensions. These are equivalent to *two* simpler equations specifying the derivatives of $\mathbf{F}$ and $\star\mathbf{F}$. The latter form is considered more revealing of the structure of the theory. Jackson [5] is the classic reference on this.

This suggests that there is more at work here than clever tricks with matrix subscripting. In fact, $\mathbf{L} = \mathbf{q}\wedge\mathbf{p}$ actually represents what is called a *decomposable bivector* in four-dimensional space. A bivector is a piece of oriented area; when it slices into our three-dimensional space it leaves a scar: the line. The field of *exterior algebra* gives us a way to deal with oriented linear geometric objects. It is there where the operations of $\wedge$ and $\star$ are revealed in full generality. What we have seen in this section is the shadow of 4D exterior algebra cast down into 3D. This viewpoint is only rarely treated. McCarthy [7] provides an accessible but abbreviated outline.

Transformations: to here, there, and nowhere

We gain another benefit by this doubling of our universe into visible and invisible copies. It has to do with our understanding of transformations of space, especially the kinds of transformations one might perform on a scene in a computer graphics program. Although the cars, trees, and stars in our virtual reality will all be assigned positions in visible space, we will find some use in transformations that send some pieces of scenes to *invisible* space. To make sense of this, we need to examine how spatial transformations are represented.

Consider first the *affine transformations*. An affine transformation is composed of a linear transformation (such as a rotation, a reflection, a scaling, or a shear) followed by a translation. Of course, a translation in three dimensions is not a linear transformation of three-dimensional space, since it will move the origin. Our four-dimensional representation is useful here. For example, the following 4×4 matrix represents a particular kind of *linear* transformation of $\mathbb{R}^4$, and it determines an *affine* transformation of Euclidean space:

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 & d_1 \\ \sin\theta & \cos\theta & 0 & d_2 \\ 0 & 0 & 1 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We apply this transformation to a point $\boldsymbol{p} = (x, y, z)$ by picking a representation of $\boldsymbol{p}$, say, $\mathbf{p} = [x\ y\ z\ 1]^{\mathrm{T}}$, then multiplying $\mathbf{p}$ by $\mathbf{A}$ on the left. The point $[\mathbf{A}\,\mathbf{p}]$ is then the image of $\boldsymbol{p}$ under this transformation. For the particular $\mathbf{A}$ above, $\boldsymbol{p}$ would be rotated by the angle θ about the z axis and then translated by (d_1, d_2, d_3). This procedure is well defined: representing $\boldsymbol{p}$ as $[5x\ 5y\ 5z\ 5]^{\mathrm{T}}$ and doing the matrix-vector multiplication, for example, would yield the same result. We will let $[\mathbf{A}]$ denote the transformation of points represented by $\mathbf{A}$. In general, rotations composed with translations are called *rigid motions*. Their use extends far beyond computer graphics into fields such as computer vision and robot kinematics.

One can also ask what these kind of affine transformations do to points at infinity. Compare the image under $[\mathbf{A}]$ of the visible point $(1, 1, 1)$ and the point at infinity $(1, 1, 1)^{\uparrow}$:

$$\mathbf{A}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} = \begin{bmatrix}\cos\theta - \sin\theta + d_1\\ \cos\theta + \sin\theta + d_2\\ 1 + d_3\\ 1\end{bmatrix} \qquad \mathbf{A}\begin{bmatrix}1\\1\\1\\0\end{bmatrix} = \begin{bmatrix}\cos\theta - \sin\theta\\ \cos\theta + \sin\theta\\ 1\\ 0\end{bmatrix}.$$

This reveals that points at infinity remain at infinity, are immune from translation, but *are* subject to rotation. (In a computer game, as we drive along a curved road the finite-distance trees move around and come closer; the sun moves around but stays at infinity.) Note further that although $\mathbf{A}$ cannot move the origin in $\mathbb{R}^4$, it *can* move the origin in the three-dimensional Euclidean space, which is represented by $[0\,0\,0\,1]^{\mathrm{T}}$.

It is useful to be able to quickly scan a 4×4 transformation matrix and discover roughly what kind of geometric action it performs. The best way to do this is to examine its columns. If $\mathbf{e}_1 = [1\,0\,0\,0]^{\mathrm{T}}$, $\mathbf{e}_2 = [0\,1\,0\,0]^{\mathrm{T}}$, and so on, then the kth column is the image of $\mathbf{e}_k$ under this transformation: $\mathbf{A}\mathbf{e}_k$. In our four-dimensional representation the vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ represent points on the celestial sphere (perhaps somewhere in Aries, Cancer, and Ursa Minor, respectively?). The vector $\mathbf{e}_4$ represents the origin. A transformation matrix (such as $\mathbf{A}$) can be easily interpreted: the columns enumerate what happens to these four cardinal points, $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$.

A general affine transformation is represented by a nonsingular 4×4 matrix whose bottom row is $[0\,0\,0\,1]$. This form reveals that affine transformations are relatively gentle. Points in visible space stay in visible space, points on the celestial sphere stay on the celestial sphere, since multiplication by this matrix cannot change the fourth component v_4.

Next we look at some violently disruptive transformations, in which large chunks of our universe burst through the celestial sphere and flee to the invisible universe, and parts of the celestial sphere cave in and become finite. These are called *projective transformations*. They are the most general class of transformations representable by these 4×4 matrices.

Perhaps the most useful type of projective transformation is the *perspective projection*, which turns up regularly in computer graphics. It transforms three-dimensional space into the two-dimensional image of the space as seen on a flat screen by an eye behind the screen. The following matrix $\mathbf{P}$ represents a simple perspective projection, which is illustrated in FIGURE 3:

$$\mathbf{P} = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & -1 & 1\end{bmatrix}.$$

Examine $\mathbf{P}$ closely. The first two columns tell us that $[\mathbf{e}_1]$ and $[\mathbf{e}_2]$ are mapped to themselves. Column four reveals that the origin $[\mathbf{e}_4]$ is also mapped to itself. This

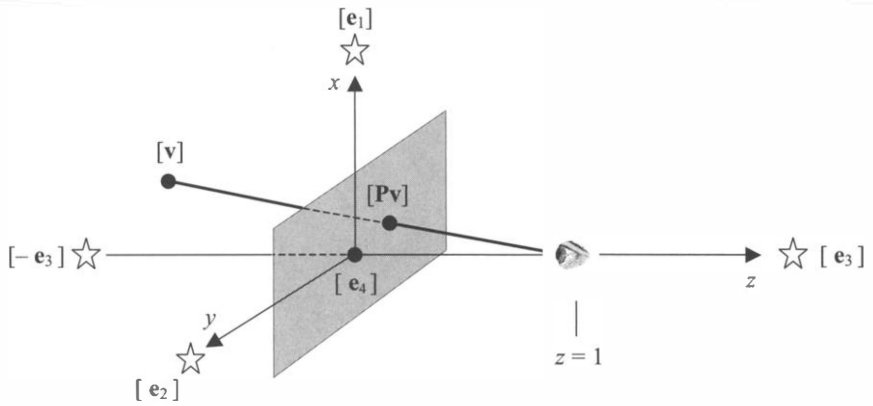


Figure 3 The action of the perspective projection **P**. The point [**v**] is mapped to [**Pv**] on the *xy*-plane (in the visible universe). The points [$\mathbf{e}_1$] and [$\mathbf{e}_2$] on the celestial sphere, as well as the origin [$\mathbf{e}_4$], are unchanged. The point [$\mathbf{e}_3$] is moved from the celestial sphere to the invisible universe, since—naturally?—it is invisible to the eye. The point [$-\mathbf{e}_3$] is pulled in from the celestial sphere onto the screen at the origin.

agrees with FIGURE 3. But look at the third column. The point [$\mathbf{e}_3$], which sits at infinity in the $+z$ direction, is mapped ... out of this world! This point is mapped to $[[0\,0\,0\,-1]^{\mathrm{T}}] = (0, 0, 0)°$, the origin of *invisible space*. From FIGURE 3, this is not surprising. The eye is aimed in the $-z$ direction and so [$\mathbf{e}_3$] is behind it. This motivated our choice of the term *invisible* to denote this copy of Euclidean space. On the other hand, the antipodal point on the celestial sphere, represented by $-\mathbf{e}_3$, is mapped to the (visible) origin.

What happens to points that are originally invisible? The point $(-1, -1, 2)°$ can be represented by $[1\,1\,-2\,-1]^{\mathrm{T}}$. Multiplying it by **P**, we see it gets sent to $[[1\,1\,0\,1]^{\mathrm{T}}] = (1, 1, 0)$, which is *visible*. Should this bother us? Not really. Computer graphics programs place their flora, fauna, and killer robots in the visible universe; the invisible universe is completely unpopulated. If a bit of this uninhabited world got mapped back into our world, we'd see nothing extra.

Since the eye is located at (0, 0, 1), it is natural to be curious about where this point is sent by [**P**]. Representing it as $[0\,0\,1\,1]^{\mathrm{T}}$ and doing the matrix-vector multiplication, we get **0**, our error signal. Put another way, we could interpret **0** as the coordinates of a special point called *Nowhere*. (Perhaps *Everywhere* would be an equally suitable label, as **0** could be construed as lying on any plane and any line: $\mathbf{n}^{\mathrm{T}}\mathbf{0} = 0$, $\star\mathbf{L0} = \mathbf{0}$.)

Although we can inspect a matrix by viewing it as a list of four column vectors representing the images of the four cardinal points, this does not mean that the transformation is *completely* determined by the images of cardinal points. Consider the two matrices

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 1 & 1 & 0 \end{bmatrix}.$$

The respective column vectors of **B** and **C** represent the same points: (1, 0, 0), (0, 1, 0), (0, 0, 1), and *Nowhere*. The matrix **B** induces a perspective projection onto the plane $x + y + z = 1$, with the center of projection (the eye) at the origin. The matrix **C** projects onto the same plane, but it is a very different kind of projection (see FIGURE 4). To see this one can multiply the matrices by $[cx\,cy\,cz\,1]^{\mathrm{T}}$ on the right. We conclude that, referring to columns $k = 1, 2, 3, 4$, even when $[\mathbf{b}_k] = [\mathbf{c}_k]$ for all k, it can still happen that $[\mathbf{B}] \neq [\mathbf{C}]$.

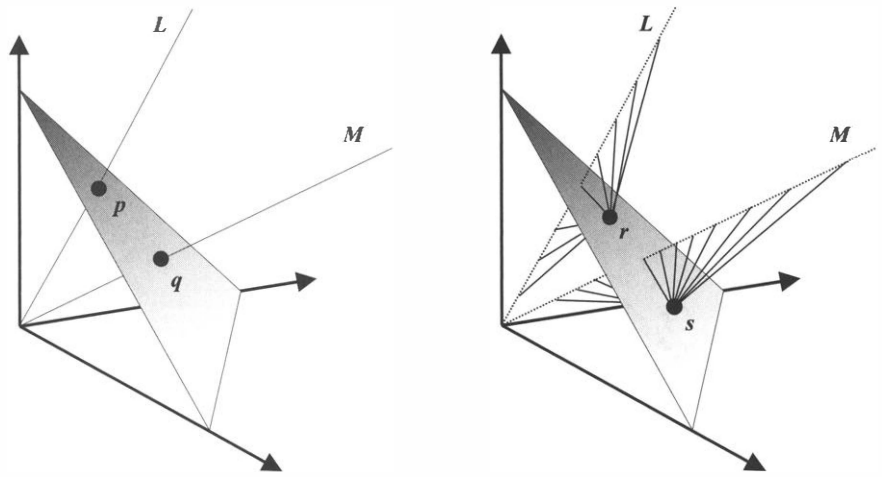


Figure 4 Two matrices whose column vectors represent the same points can still represent different transformations. *Left:* The action of the matrix **B**, a perspective projection onto the plane $x + y + z = 1$ with the eye at the origin. All points on the half-line $\boldsymbol{L}$ are pulled to point $\boldsymbol{p}$ (which is also on $\boldsymbol{L}$); all points on the half-line $\boldsymbol{M}$ are pulled to point $\boldsymbol{q}$ (which is also on $\boldsymbol{M}$). *Right:* The action of the matrix **C**, a more complicated projective transformation. All points of $\boldsymbol{L}$ are pulled to a point $\boldsymbol{r}$ on the plane, but this point is not on $\boldsymbol{L}$.

In fact, it can be shown that in three-dimensional space, the image of *five* independent points is needed to determine a projective transformation. This is not surprising, as there is room for four more numbers in the scale factors for each column of the transformation matrix. Many of the references cited in the last section of this paper give recipes on how to construct a general projective transformation given various requirements. For us now it will be more informative to give a specific recipe for perspective projections, such as the ones shown in FIGURE 3 or the left half of FIGURE 4. We will then see that this recipe handles visible and invisible points differently.

Building on the theory of oriented intersection of lines and planes in the previous section, developing a recipe for perspective projections will be easy. Recall that a line from $\mathbf{p}$ through $\mathbf{q}$ is represented by $\mathbf{L} = \mathbf{q} \wedge \mathbf{p}$, and this line intersects the plane $\mathbf{n}^T$ at the point $\mathbf{x} = \mathbf{Ln}$. Now let us take $\mathbf{n}^T$ as the screen onto which we are projecting, and let $\mathbf{c}$ be the eye. Then the projection of an arbitrary point $\mathbf{y}$ onto the screen is the intersection of the line running from $\mathbf{c}$ to $\mathbf{y}$ with the plane $\mathbf{n}^T$, that is, $\mathbf{x} = (\mathbf{y} \wedge \mathbf{c})\mathbf{n}$. Rewriting this a bit, we have $\mathbf{x} = (\mathbf{yc}^T - \mathbf{cy}^T)\mathbf{n} = [(\mathbf{c}^T\mathbf{n})\mathbf{I} - (\mathbf{cn}^T)]\mathbf{y}$, so the matrix representing the projection onto screen $\mathbf{n}^T$ through eye $\mathbf{c}$ is simply

$$\mathbf{P} = (\mathbf{c}^T\mathbf{n})\mathbf{I} - \mathbf{cn}^T.$$

For example, if we take $\mathbf{c} = [0\,0\,1\,1]^T$ and $\mathbf{n}^T = [0\,0\,1\,0]$ we get the matrix $\mathbf{P}$ for the perspective projection shown in FIGURE 3.

Visibility still plays a role here. Look at the fourth component of $\mathbf{Px}$ for an arbitrary point $\mathbf{x}$. It will be negative if and only if $\mathbf{x}^T\mathbf{n} > \mathbf{c}^T\mathbf{n}$. Well,"negative fourth component" means "invisible," and $\mathbf{x}^T\mathbf{n} > \mathbf{c}^T\mathbf{n}$ means that the point $\mathbf{x}$ lies behind the eye $\mathbf{c}$ as it looks at the screen. Thus a perspective projection maps points behind the eye to invisible space, as expected.

In a perspective projection it is often useful to separate the *perspective* from the *projection*, that is, to achieve the visual effect of perspective first without dropping any dimensionality. Later, perhaps after one has clipped some elements from the scene, it

can be projected down to 2D. For example, here is a nonsingular matrix that achieves the same perspective effect as **P** but does not collapse the scene:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

R compresses the scene in front of the eye to a box of finite depth. The eye is still at $\boldsymbol{c} = (0, 0, 1)$ and is looking leftward. The frustum bounded by the faces $(\pm 2, \pm 2, -1)$ and $(\pm 1, \pm 1, 0)$ is sent to the cube $[-1, 1]^3$. More dramatically, all points with $z < 0$, *including* those at infinity, are pulled into the volume lying between $z = -3$ and $z = +1$. This is shown in FIGURE 5. Points behind the eye are, as we would expect

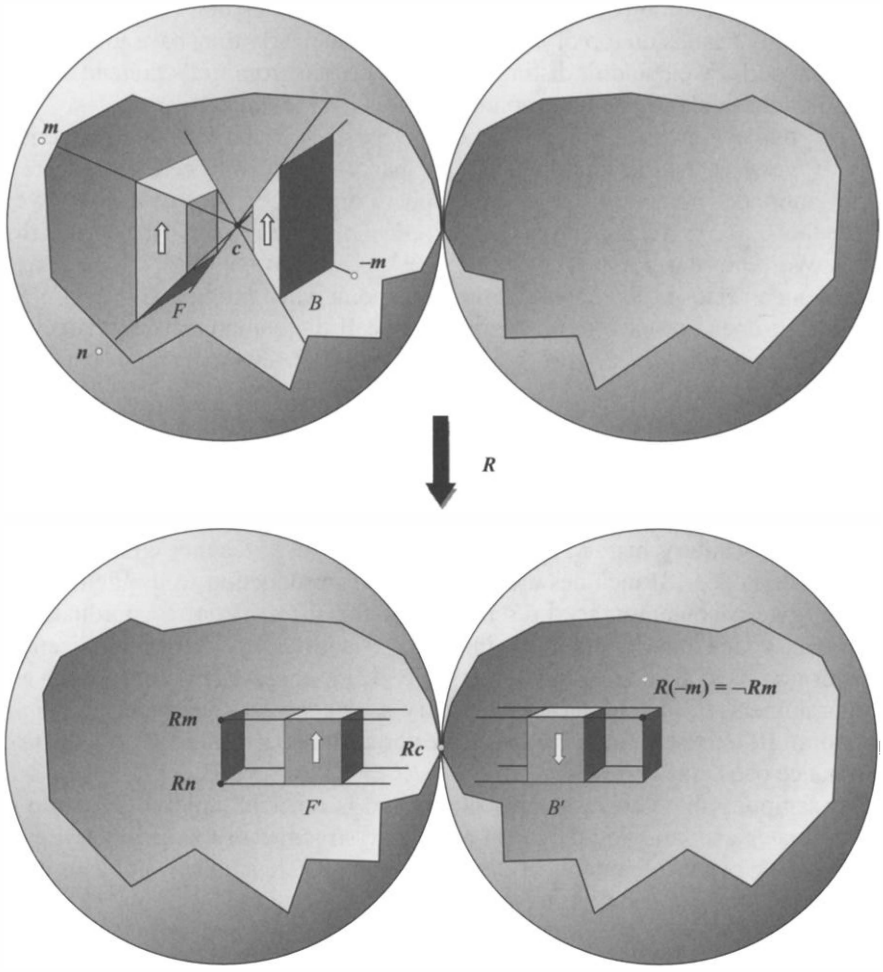


Figure 5 The action of a nonsingular projective transformation commonly used in computer graphics, such as the matrix **R** given in the text. *Left,* visible space; *right,* invisible space. The eye at **c** looks leftward. The frustum F is transformed to the cube F'. Points at infinity in front of the eye (e.g., $\boldsymbol{m}$, $\boldsymbol{n}$) are pulled in to finite distance. Points behind the eye are sent to invisible space. The eye itself is sent to the celestial sphere. After this transformation, an orthogonal projection onto $z = 0$ will achieve the effect of a perspective projection.

now, sent to invisible space and inverted. Note that since **R** is nonsingular, the eye cannot be mapped to Nowhere; it is sent to infinity. In effect **R** slides everything to the right, pushing points through the celestial sphere in the process. When we finally need to collapse this to 2D for viewing on a screen, we can use the simplest of all orthogonal projections: we just zero-out the z-coordinate.

On blurring a distinction

Stare at FIGURE 1 and cross your eyes. See the two spheres superimpose, the two universes merge into one. The visible point (2, 3, 4) and the invisible point (2, 3, 4)° now sit on top of each other. We can think of this cross-eyed image as a new space in which we treat $\boldsymbol{p}$ and $\neg\boldsymbol{p}$ as the same point. But there seems to be a problem. When you cross your eyes, the two copies of the celestial spheres don't line up right. The point at infinity $\boldsymbol{r}$ lands on top of its antipode $-\boldsymbol{r}$, but clearly they have to be different. It would be odd if we couldn't distinguish the north star from (let's pretend) the south star. One might think no good can come from this confusion.

In fact, this eye-crossing exercise leads us to a classic and valuable mathematical theory. If we start with ordinary Euclidean space, wrap it in a celestial sphere, and say that antipodal points on the celestial sphere are indeed the *same point*, we end up with the space of (classical) *projective geometry*, $\mathbb{R}P^3$ (read "real projective 3-space"). We represent points as vectors in $\mathbb{R}^4$ as we always did, but we relax one requirement: a vector in $\mathbb{R}^4$ represents the same point when multiplied by *any* nonzero constant k; it does not have to be positive. We call the components of these vectors *homogeneous coordinates*. So in $\mathbb{R}P^3$, $[2\ 3\ 4\ 0]^{\mathrm{T}}$ and $[-2\ -3\ -4\ 0]^{\mathrm{T}}$ represent the same point at infinity, since they yield each other when multiplied by $k = -1$. This means if we journey outward to the celestial sphere, as soon as we cross it we re-enter the same space again—from the opposite (antipodal) side. In $\mathbb{R}P^3$ there is no extra invisible space, but now, yes, the north star is the same as the south star.

Classical projective geometry is a pinnacle of mathematical elegance and power, and has a four-century history. The material is so rich we cannot enter into it here. The book by Pedoe [**8**] includes a brief historical introduction to the field. Brannan et al. [**2**] give a contemporary presentation of the theory from a coordinate-based perspective. Penna and Patterson [**9**] introduce it with an eye toward applications in computer graphics. The paper by van Arsdale [**12**] is a compact reference for graphics programmers. In fact, projective geometry is *so* central to computer graphics that the Pentium III chip supports parallel 4D vector arithmetic specifically to enhance the performance of 3D programs.

After sampling these concrete presentations, it is especially pleasing to strip away the coordinates and examine the clean axiomatic structure of this geometry, perhaps one of the most beautiful systems in all mathematics. There are many good books here, for instance Coxeter [**3**] and Beutelspacher and Rosenbaum [**1**].

When you leaf through any of these texts on projective geometry one thing may strike you: the utter absence of inequalities! One searches in vain for a single $<$ sign. This is because classical projective geometry is all about *incidence*—points lying on lines, lines intersecting planes, and so on. Because it throws out the signs, it cannot say *in front of* or *behind*. Life is simpler: You are on it. Or you're not. But that certainly makes it an odd choice for *the* geometric language of computer graphics. For example, take the perspective projection **P** that was illustrated in FIGURE 3. The point $(2, -3, -4)$, which is in front of the eye, and the point $(2, -3, 6)$, which is behind the eye, were mapped to the visible screen and a corresponding invisible screen, respectively. When we cross our eyes to get classical projective 3-space, these screens

coincide; both points are both mapped to the same screen. Thus, the scene behind the viewer becomes superimposed, upside down, on the scene in front of the viewer. Accordingly, the program code will require a few judiciously placed **if** statements to pluck out these superfluous, ghostly points.

The geometry we have explored in this paper is called *oriented* projective geometry. Oriented projective space—call it $\mathbb{R}OP^3$—is the union of visible space, invisible space, and the celestial sphere. Stolfi [**10**] provides the definitive work on the subject. The oriented approach has still not made it into the textbooks, however. One reason may be that mathematicians (those who have heard of it) certainly see it as intrinsically less interesting. For example, the topology of $\mathbb{R}OP^3$ can be seen to be the same as the familiar 3-sphere, S^3. $\mathbb{R}P^3$ is given by the quotient $\mathbb{R}OP^3/\{-1, 1\}$, which identifies antipodes (our eye-crossing exercise). Not a sphere of any kind, $\mathbb{R}P^3$ is *sui generis*, with applications in many branches of mathematics.

Another reason that $\mathbb{R}OP^3$ has not yet taken the world by storm is that computer graphics programmers (those who have heard of it) might see oriented projective geometry as merely a tidying up of the mathematics that does not affect their everyday work. The front/back superposition problem can indeed be dealt with within the classical approach. For example, one can use certain non-singular perspective transformations, such as the **R** described in the previous section, to handle this. When you cross your eyes and look at the bottom half of FIGURE 5, the images under **R** of the front and back views remain disjoint even as the spheres superimpose. Programs can just clip out what they don't want. But one should be frank: insofar as one peeks at the signs of homogeneous vectors, one *is* using oriented projective geometry. Computer graphics has done so from its inception.

In the end, the primary advantage of oriented projective geometry may be its *picturability*. (One can consult Stolfi [**11**] for other advantages this article has not touched on, such as the fact that convexity can be defined in $\mathbb{R}OP^3$ but not in $\mathbb{R}P^3$.) Points at infinity comprise a 2-sphere, and this fits our experience well. The oriented approach provides an alternative way to present the perspective transformations commonly featured as examples in linear algebra textbooks nowadays. It allows a concrete global picture of the operations of the transformations, without requiring the student to visualize the unusual identifications of projective space, which may produce more head scratching than is useful at that level.

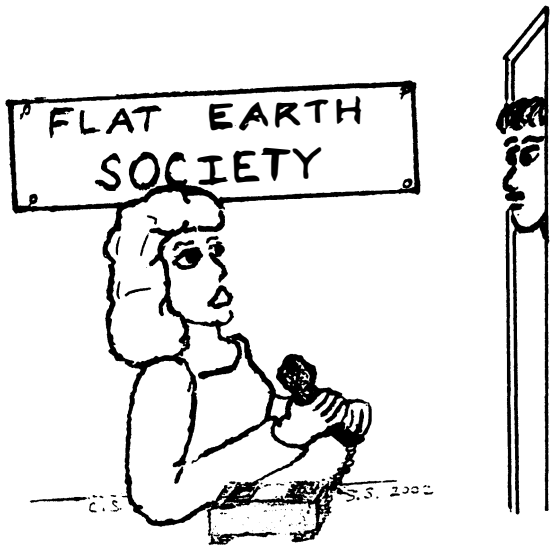
The double-space pictures of oriented projective geometry seem odd at first, but they are reminiscent of a kind of Dantean cosmology that is eerily engaging. It is satisfying to tell one's linear algebra students to go out at night and take a look at the dome of the heavens, imagining an entire universe behind it converging on an infinitely distant invisible origin. And the geometry of this world is just one quotient operation away from the most sublime system in mathematics.

REFERENCES

1. A. Beutelspacher and U. Rosenbaum, *Projective Geometry: From Foundations to Applications*, Cambridge University Press, 1998.
2. D. A. Brannan, M. F. Esplen, and J. J. Gray, *Geometry*, Cambridge University Press, 1999.
3. H. S. M. Coxeter, *Projective Geometry*, 2nd ed., Springer Verlag, 1987.
4. H. W. Guggenheimer, Formulas of linear geometry, *College Math. J.* **27**:1 (1996), 24–32.
5. J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, 1998.
6. S. Laveau and O. Fageras, Oriented projective geometry for computer vision. In *Lecture Notes in Computer Science*, Vol. 1064, Springer-Verlag, 1996, pp. 147–156.
7. J. M. McCarthy, *Introduction to Theoretical Kinematics*, MIT Press, 1990.
8. D. Pedoe, *Geometry and the Visual Arts*, Dover, 1983.

9. A. Penna and R. Patterson, *Projective Geometry and Its Applications to Computer Graphics*, Prentice Hall, 1986.
10. J. Stolfi, Oriented projective geometry, *Proceedings of the 3rd ACM Symposium on Computational Geometry*, 1987, pp. 76–85.
11. J. Stolfi, *Oriented Projective Geometry*, Academic Press, 1991.
12. D. A. van Arsdale, Homogeneous transformation matrices for computer graphics *Computers and Graphics* **18**:2 (1994), 171–191.

Cartoon: At the Flat Earth Society



"I have a differential geometer on the line.
He wants to know if he can join the local branch
without joining the global organization."

——COLIN STARR
STEPHEN F. AUSTIN STATE UNIVERSITY
NACOGDOCHES, TX 75962

NOTES

Areas and Centroids for Triangles within Triangles

HERBERT R. BAILEY
Rose-Hulman Institute of Technology
Terre Haute, IN 47803

Given $\triangle ABC$, let A', B', C' be any points on the sides CB, AC, and BA. This creates a second triangle $\triangle A'B'C'$. A third triangle is determined by the vertices P, Q, R, the intersection points of the segment pairs AA' and BB', BB' and CC', CC' and AA', as shown in FIGURE 1.

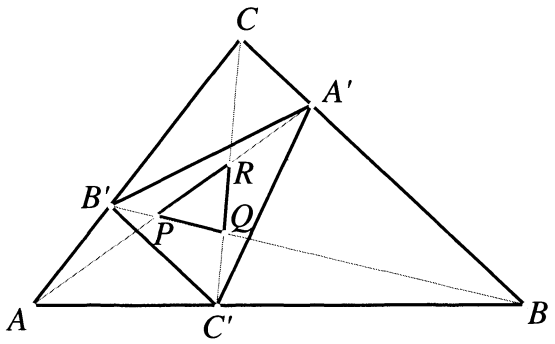


Figure 1 The three triangles

Certain results about the ratios of areas of these three triangles are known. In this note, we use vector algebra to derive them. We also obtain some new results concerning the relative positions of the centroids of the triangles.

A line segment joining a vertex of a triangle and a point on the side opposite that vertex is called a cevian. The cevians AA', BB', and CC' are determined by specifying certain ratios on the sides, namely a triple (t_1, t_2, t_3) such that

$$\frac{AB'}{AC} = t_1, \quad \frac{BC'}{BA} = t_2, \quad \text{and} \quad \frac{CA'}{CB} = t_3,$$

where the ts are numbers between 0 and 1. If $t_1 = t_2 = t_3 = 1/2$, then the cevians are the medians and the points P, Q, and R are known to coincide. This common point is called the centroid of $\triangle ABC$. If $t_1 = t_2 = t_3 = 1/3$, then, surprisingly, the area of $\triangle PQR$ is one seventh the area of $\triangle ABC$. A clever geometric proof of this result is given by H. Steinhaus [**8**, p. 9]. If K_1, K_2, and K_3 represent the areas of triangles $\triangle ABC$, $\triangle A'B'C'$, and $\triangle PQR$, then formulas for the ratios K_2/K_1 and K_3/K_1 are known for the general case with the ts not necessarily equal. These results are given by E. J. Routh [**7**], and H. S. M. Coxeter [**2**] proves them as an application of affine geometry.

Let G_1, G_2, and G_3 be the centroids of $\triangle ABC$, $\triangle A'B'C'$, and $\triangle PQR$. The centroids G_1 and G_2 are known to coincide if and only if $t_1 = t_2 = t_3$. R. Johnson [**4**, p. 175] gives a proof of this result, which he attributes to W. Fuhrmann. M. Khan and W. Pierce [**5**] proved that if $t_1 = t_2 = t_3$, then G_1 and G_3 coincide. This result was posed as a problem by Khan and solved by Pierce by representing points in the plane as complex numbers. L. Hahn [**3**] introduced this method and, as an example, uses it to prove that the medians of a triangle are concurrent. A. Ayoub [**1**] uses the complex number representation to solve a related problem concerning centroids.

Preliminary vector results To simplify calculations, we will assume without loss of generality that the point A is the origin $\mathbf{0}$. If we denote by $\mathbf{X}$ the vector from A to X, then

$$\mathbf{A}' = \mathbf{C} + t_3(\mathbf{B} - \mathbf{C}), \quad \mathbf{B}' = t_1\mathbf{C}, \quad \text{and} \quad \mathbf{C}' = (1 - t_2)\mathbf{B}.$$

The cevians are divided into parts by the vertices P, Q, and R, and we define the following ratios:

$$\lambda_1 = \frac{AP}{AA'}, \quad \lambda_2 = \frac{BQ}{BB'}, \quad \lambda_3 = \frac{CR}{CC'}, \quad \mu_1 = \frac{AR}{AA'}, \quad \mu_2 = \frac{BP}{BB'}, \quad \mu_3 = \frac{CQ}{CC'}.$$

We now seek to express the λs and the μs in terms of the ts. The vector $\mathbf{P}$ can be expressed in two ways:

$$\mathbf{P} = \lambda_1\mathbf{A}' = \lambda_1[\mathbf{C} + t_3(\mathbf{B} - \mathbf{C})]$$

and

$$\mathbf{P} = \mathbf{B} + \mu_2(\mathbf{B}' - \mathbf{B}) = \mathbf{B} + \mu_2(t_1\mathbf{C} - \mathbf{B}).$$

Setting these two expressions for $\mathbf{P}$ equal and equating the coefficients of $\mathbf{B}$ gives $\lambda_1 t_3 = 1 - \mu_2$; equating the coefficients of $\mathbf{C}$ gives $\lambda_1(1 - t_3) = \mu_2 t_1$. Solving these two equations gives λ_1 and μ_2. Similar calculations using vectors $\mathbf{Q}$ and $\mathbf{R}$ give the remaining λs and μs and we have

$$\lambda_1 = \frac{t_1}{1 - t_3 + t_1t_3}, \quad \lambda_2 = \frac{t_2}{1 - t_1 + t_1t_2}, \quad \lambda_3 = \frac{t_3}{1 - t_2 + t_2t_3},$$

$$\mu_1 = \frac{-1 + t_2}{1 - t_2 + t_2t_3}, \quad \mu_2 = \frac{1 - t_3}{1 - t_3 + t_1t_3}, \quad \mu_3 = \frac{1 - t_1}{1 - t_1 + t_1t_2}.$$

This result, for the case $t_1 = t_2 = t_3$, is given by T. Zerger and R. Young [**9**], using the same method that we used above for the general case.

Areas ratios for the three triangles The area of a triangle determined by two vectors is equal to one-half the magnitude of their cross product. Thus for $\triangle ABC$ we have

$$K_1 = |\mathbf{B} \times \mathbf{C}|/2.$$

For $\triangle A'B'C'$ we have, after some simplification,

$$\begin{aligned} K_2 &= \left|(\mathbf{B}' - \mathbf{A}') \times (\mathbf{C}' - \mathbf{A}')\right|/2 \\ &= [(1 - t_1)(1 - t_2)(1 - t_3) + t_1t_2t_3]\,|\mathbf{B} \times \mathbf{C}|/2, \end{aligned}$$

where we have used the relations $\mathbf{B} \times \mathbf{B} = 0$, $\mathbf{C} \times \mathbf{C} = 0$, and $\mathbf{B} \times \mathbf{C} = -\mathbf{C} \times \mathbf{B}$. Thus the area ratio K_2/K_1 is given by the bracketed term above. With changes in notation, this result agrees with Routh [**7**, p. 82] and Coxeter [**2**, p. 212].

For $\triangle PQR$ we have

$$K_3 = |(\mathbf{Q} - \mathbf{P}) \times (\mathbf{R} - \mathbf{P})| /2,$$

where $\mathbf{P} = \lambda_1 \mathbf{A}'$, $\mathbf{Q} = \mathbf{B} + \lambda_2(\mathbf{B}' - \mathbf{B})$, and $\mathbf{R} = \mu_1 \mathbf{A}'$. The λs, μs and the primed vectors are given above in terms of the ts and the vectors $\mathbf{B}$ and $\mathbf{C}$. The calculations involving K_3 are similar to those for K_2, but far more tedious. We have used *Maple* to help with the computations and would be glad to send an electronic copy of our worksheet to interested readers.

The resulting equation for K_3 is

$$K_3 = \left[\frac{(t_1 t_2 t_3 - (1 - t_1)(1 - t_2)(1 - t_3))^2}{(1 - t_1 + t_1 t_2)(1 - t_2 + t_2 t_3)(1 - t_3 + t_1 t_3)} \right] |\mathbf{B} \times \mathbf{C}| /2,$$

and thus the area ratio K_3/K_1 is given by the bracketed term above. With changes in notation, this result agrees with the corresponding results of Coxeter [**2**, pp. 219–220] and Routh [**7**, p. 82].

Conditions for coincidence of the centroids The centroid of a triangle with vertices determined by three vectors is known to be given by one-third the sum of the three vectors. This result was proved as an example for vector proofs of geometric results in a paper by W. Mueller [**6**]. Thus the centroid of $\triangle ABC$ is given by

$$\mathbf{G}_1 = (\mathbf{A} + \mathbf{B} + \mathbf{C}) /3 = (\mathbf{B} + \mathbf{C}) /3,$$

since $\mathbf{A}$ is the origin. The centroid of $\triangle A'B'C'$ is represented by the vector

$$\mathbf{G}_2 = (\mathbf{A}' + \mathbf{B}' + \mathbf{C}')/3 = (1 - t_2 + t_3)\mathbf{B}/3 + (1 - t_3 + t_1)\mathbf{C}/3.$$

The centroids of triangles $\triangle ABC$ and $\triangle A'B'C'$ coincide if and only if $\mathbf{G}_1 = \mathbf{G}_2$. Equating the $\mathbf{B}$ and $\mathbf{C}$ components in this vector equation gives the two scalar equations, $1 - t_2 + t_3 = 1$ and $1 - t_3 + t_1 = 1$. Thus G_1 and G_2 coincide if and only if $t_1 = t_2 = t_3$.

For $\triangle PQR$ we have

$$\mathbf{G}_3 = (\mathbf{P} + \mathbf{Q} + \mathbf{R})/3,$$

where $\mathbf{P} = \lambda_1 \mathbf{A}'$, $\mathbf{Q} = \mathbf{B} + \lambda_2(\mathbf{B}' - \mathbf{B})$, and $\mathbf{R} = \mu_1 \mathbf{A}'$. Centroids G_1 and G_3 coincide if and only if $\mathbf{G}_1 = \mathbf{G}_3$. We first express this vector equation in terms of the ts and the vectors $\mathbf{B}$ and $\mathbf{C}$, and then obtain two scalar equations relating t_1, t_2, and t_3. After combining and factoring these equations, we find that $\mathbf{G}_1 = \mathbf{G}_3$ if and only if $t_1 = (t_3 - t_2 + t_2 t_3)/t_3$ and $t_2(t_2 - t_3)(1 - t_2 + t_2 t_3)/t_3 = 0$. With the ts being between 0 and 1, we have t_3, t_2, and $1 - t_2 + t_2 t_3$ all nonzero. Hence G_1 and G_3 coincide if and only if $t_1 = t_2 = t_3$. Similar calculations also show that G_2 and G_3 coincide if and only if $t_1 = t_2 = t_3$. The calculations involving $\mathbf{G}_3$ are similar to those for $\mathbf{G}_2$, but even more tedious than for K_3. Again, we enlisted the help of *Maple* for the computations.

Collinear centroids If the ts are distinct then the three centroids will be distinct, and they will be collinear if and only if the sine of the angle between vectors $(\mathbf{G}_1 - \mathbf{G}_2)$

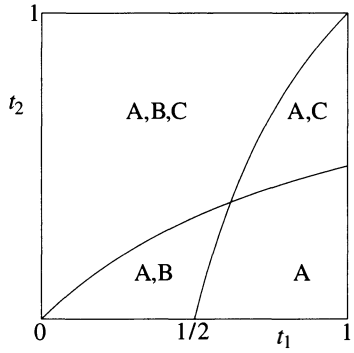


Figure 2 Permissible regions for (t_1, t_2)

and $(\mathbf{G}_1 - \mathbf{G}_3)$ is 0. This will happen if and only if $(\mathbf{G}_1 - \mathbf{G}_2) \times (\mathbf{G}_1 - \mathbf{G}_3) = \mathbf{0}$. Expanding this cross product in terms of the ts and the vectors **B** and **C** gives

$$(\mathbf{G}_1 - \mathbf{G}_2) \times (\mathbf{G}_1 - \mathbf{G}_3) = \frac{1}{18}\left[\frac{(t_2 - t_3 - t_2t_3 + t_1t_3)(t_3 - t_1 - t_1t_3 + t_1t_2)(t_1 - t_2 - t_1t_2 + t_2t_3)}{(1 - t_1 + t_1t_2)(1 - t_2 + t_2t_3)(1 - t_3 + t_1t_3)}\right]\mathbf{B} \times \mathbf{C}.$$

Since the denominator of the bracketed term is never 0, collinearity occurs if and only if at least one of the factors in the numerator is 0. Setting each of these factors to 0 and solving for t_3 gives three corresponding conditions:

$$t_{3A} = \frac{t_2}{1 + t_2 - t_1}, \quad t_{3B} = \frac{t_1 - t_1t_2}{1 - t_1}, \quad \text{and} \quad t_{3C} = \frac{t_2 - t_1 + t_1t_2}{t_2}.$$

The three centroids are collinear if and only if at least one of these three equations is satisfied with the ts all between 0 and 1.

Given any pair $0 < t_1, t_2 < 1$, we must require $0 < t_3 < 1$. This is always true for $t_3 = t_{3A}$. The permissible (t_1, t_2) regions for which $0 < t_{3B} < 1$ are shown graphically in FIGURE 2 with the letter B. The corresponding (t_1, t_2) regions where the inequalities $0 < t_{3C} < 1$ and $0 < t_{3A} < 1$ are satisfied are also shown there with the letters C and A. The intersection point of the two curves in FIGURE 2 is $(\tau, 1 - \tau)$, where $\tau = (\sqrt{5} - 1)/2$, the reciprocal of the golden ratio. For example if $t_1 = 2/5, t_2 = 1/5$, then $t_{3A} = 1/4, t_{3B} = 8/15$, and $t_{3C} = -3/5$, and thus there are two permissible values of t_3. This agrees with FIGURE 2 since $(2/5, 1/5)$ is in the region A,B. The three collinear centroids for the case with $t_3 = t_{3B} = 8/15$ are shown in FIGURE 3. This figure was drawn using *Geometer's SketchPad* software. Note that $G_1G_3 > G_1G_2$ in this case, and that the line through the centroids is parallel to the cevian BB'. This example demonstrates a special case of the following general results.

Result 1 For any triple (t_1, t_2, t_{3A}) with $0 < t_1 < 1, 0 < t_2 < 1$, and $0 < t_{3A} < 1$, the line through the three centroids is parallel to the cevian AA'. Analogous results hold with the As replaced by either Bs or Cs.

Result 2 If the three centroids are collinear then $G_1G_3 > G_1G_2$.

To prove these two results, we calculate expressions for $\mathbf{G}_2 - \mathbf{G}_1$ and $\mathbf{G}_3 - \mathbf{G}_1$ for each of the cases A, B, and C. We let $(\mathbf{G}_2 - \mathbf{G}_1)_A$ and $(\mathbf{G}_3 - \mathbf{G}_1)_A$ denote the vector differences, evaluated at (t_1, t_2, t_{3A}), with analogous definitions for cases B and C.

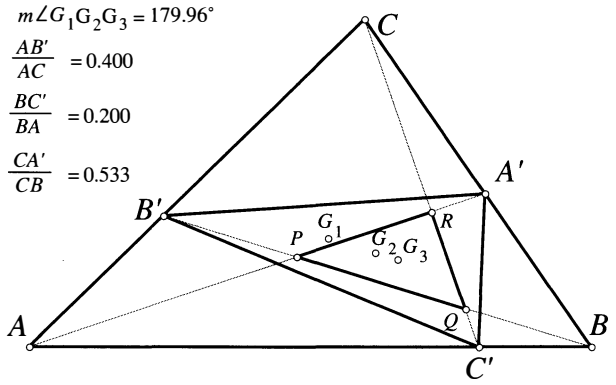


Figure 3 An example of collinear centroids

After some calculation, we obtain the following expressions for the vectors joining the centroids.

$$(\mathbf{G}_2 - \mathbf{G}_1)_A = \frac{t_1 - t_2}{3}(\mathbf{A} - \mathbf{A}') \qquad (\mathbf{G}_3 - \mathbf{G}_1)_A = \frac{(t_1 - t_2)(1 - t_1 + t_2)}{3(1-t_1 + t_1t_2)}(\mathbf{A} - \mathbf{A}')$$

$$(\mathbf{G}_2 - \mathbf{G}_1)_B = \frac{t_1 - t_2}{3(1-t_1)}(\mathbf{B} - \mathbf{B}') \qquad (\mathbf{G}_3 - \mathbf{G}_1)_B = \frac{t_1 - t_2}{3(1-t_1)(1-t_1 + t_1t_2)}(\mathbf{B} - \mathbf{B}')$$

$$(\mathbf{G}_2 - \mathbf{G}_1)_C = \frac{t_1 - t_2}{3t_2}(\mathbf{C} - \mathbf{C}') \qquad (\mathbf{G}_3 - \mathbf{G}_1)_C = \frac{t_1 - t_2}{3t_2(1-t_1 + t_1t_2)}(\mathbf{C} - \mathbf{C}')$$

The three cases for Result 1 follow immediately from the above relations. Result 2 for cases B and C follows as well since $1 - t_1 + t_1t_2$ is between 0 and 1. Case A follows since $(1 - t_1 + t_1t_2)/(1 - t_1 + t_2)$ is also between 0 and 1. Finally, we note that for collinear centroids the ratio $G_1G_2/G_1G_3 = |\mathbf{G}_2 - \mathbf{G}_1|/|\mathbf{G}_3 - \mathbf{G}_1| = 1 - t_1 + t_1t_2$ for both cases B and C.

Remarks High school and college curricula have given scant attention to geometry in recent years and yet many of us would agree that it was one of the most worthwhile courses in our training. The formal theorem-proof approach seems to have little appeal to most undergraduate students. Mueller [**6**] makes the case that, with modern computer software, there are many geometric topics for undergraduates to explore using vector algebra. Many of the calculations can be handled easily with computer algebra systems. An equally important computer tool is geometric sketching software, for example *The Geometer's SketchPad.* Considerable insight can be obtained by sketching the figure and dragging the various points about. For the centroid problem it is interesting to animate the sketch and show the motion of G_2 and G_3 as A', B', and C' move back and forth along their sides of the given $\triangle ABC$. A related problem is to find what motion of A', B', and C' leads to orbital motion of G_2 or G_3 around the fixed center G_1. Solutions to the orbital problems will depend on the coordinates of the vertices of $\triangle ABC$. Another variation of the problems that we have not considered is to permit the points A', B', and C' to lie anywhere on the lines determined by the sides of $\triangle ABC$.

Finally, a word of warning about computer algebra systems: they do not always return all possible solutions for some algebraic equations. So if you are looking for necessary conditions for a solution, you must solve the equations without depending solely on the `solve` command. In this note, we were always able to solve by factoring the equations into linear factors to avoid this difficulty.

Acknowledgment. The author thanks the referees for many very valuable suggestions.

REFERENCES

1. A. Ayoub, Triangles with the same centroid, this MAGAZINE **71** (1998), 221–224.
2. H. S. M. Coxeter, *Introduction to Geometry*, John Wiley, New York, 1961.
3. L. Hahn, *Complex Numbers and Geometry*, The Mathematical Association of America, Washington DC, 1994.
4. R. A. Johnson, *Advanced Euclidean Geometry*, Dover Publications, New York, NY, 1960.
5. L. Khan, Problem 846, Solution by W. Pierce, *Pi Mu Epsilon Journal* **10** (1995), 243–246.
6. W. Mueller, Centers of triangles of fixed center: Adventures in undergraduate mathematics, this MAGAZINE **70** (1997), 252–254.
7. E. Routh, *A Treatise on Analytical Statics with Numerous Examples*, Vol. 1 (2nd ed.), Cambridge University Press, 1986.
8. H. Steinhaus, *Mathematical Snapshots*, Oxford University Press, 1983.
9. T. Zerger, Problem 1524, Solution by R. Young, this MAGAZINE **71** (1998), 226–227.

Tiling Large Rectangles

DARREN A. NARAYAN
Rochester Institute of Technology
Rochester, NY 14623

ALLEN J. SCHWENK
Western Michigan University
Kalamazoo, MI 49008

Tiling problems are the jigsaw puzzles of a mathematician. For hundreds of years such problems have produced an array of beautiful and intriguing patterns. In this note we consider tiling a large rectangle using smaller rectangles, and present a problem where, although the pieces are one of only two sizes, a large collection of rectangles can be tiled. Problem B-3 from the 1991 William Lowell Putnam Examination asked: "Does there exist a natural number L, such that if m and n are integers greater than L, then an $m \times n$ rectangle may be expressed as a union of 4×6 and 5×7 rectangles, any two of which intersect at most along their boundaries?" The answer is yes, and a solution appeared in the *Monthly* [3]. The published solution is an existential proof and no attempt is made to find the minimal value of L. In this paper we show that the smallest L, denoted L_s, equals 33.

We will refer to the 4×6 and 5×7 rectangles as *tiles*, and if a given rectangle can be expressed as a union of 4×6 and 5×7 tiles, any two of which intersect at most along their boundaries, we will simply say the rectangle can be tiled. Since tiles can be rotated, we do not distinguish a 4×6 tile from a 6×4 tile or a 5×7 tile from a 7×5 tile. However, the larger rectangles that we will cover with our tiles will have a fixed orientation, so that if we refer to a "39×29 rectangle" we mean a rectangle with 39 rows (running from top to bottom) and 29 columns (running from left to right). We will refer to a 4×6 tile as an *even* tile and a 5×7 tile as an *odd* tile.

To show a rectangle can be tiled, our strategy will be to decompose it into smaller rectangles that can easily be tiled. While searching for a possible tiling using our tiles, it is useful to note that the area A of a rectangle that can be tiled must satisfy the equation $24x + 35y = A$ where x and y are nonnegative integers. For example, a tiling of a 34×37 rectangle could only use 32 even tiles and 14 odd tiles since $x = 32$ and $y = 14$ is the unique solution of the equation $24x + 35y = 34 \times 37 = 1{,}258$,

where x and y are nonnegative integers. The solutions to this equation determine a list of the permissible quantities of the two types of tiles used in a tiling, and provide a starting point in the search for an actual tiling. On the other hand, showing that a rectangle cannot be tiled is not as straightforward, and for these cases we use relatively complicated coloring arguments.

The tiling of rectangles has been studied elsewhere [**1, 2, 4, 6**]. The specific problem considered here is somewhat different from these papers in that here only two tiles with prescribed dimensions can be used. We note that the areas of the two tiles are relatively prime. This is a necessary condition because if an integer $p > 1$ divides each area, tiling a rectangle whose sides are both congruent to 1 mod p would not be possible.

We now describe the existence proof from the *Monthly* [**3**]. It uses the following well-known fact from algebra, which has been credited to Sylvester [**5**].

LEMMA 1. *If a and b are natural numbers with* $\gcd(a, b) = 1$, *then every integer* $n \geq (a-1)(b-1)$ *can be written as a nonnegative linear combination of a and b, while the integer* $n = (a-1)(b-1) - 1$ *cannot.*

Proof. We do the last part first, showing that this number always requires negative integers as coefficients of one of a and b in a linear combination. One particular way to express this number is $n = (a-1)(b-1) - 1 = (b-1)a + (-1)b$, where the coefficients are $b-1$ and -1. In general, if x and y give one integer solution to $ax + by = c$, then all other solutions are of the form $a(x + kb) + b(y - ka) = c$. So, adding a multiple mb to the first solution and subtracting ma from the second yields additional solutions of the form $n = (a-1)(b-1) - 1 = (b - 1 + mb)a + (-1 - ma)b$. If m is negative, the first coefficient is negative; if m is nonnegative, the second coefficient is negative. Thus, our equation has no nonnegative solutions for this value of n.

Any larger n is of the form $n = (a-1)(b-1) - 1 + j$. Since $\gcd(a, b) = 1$, there exist x and y with $1 = xa + yb$, but one of x or y must be negative. Multiplying by j gives $j = (jx)a + (jy)b$. This solution can again be modified by adding mb to the first solution and subtracting ma from the second. By judiciously selecting m we obtain the unique solution $j = sa + tb$ for which $-(b-1) \leq s \leq 0$ and $t \geq 1$. Adding this to the unique solution obtained above (for $j = 0$) yields $n = (a-1)(b-1) - 1 + j = (b-1)a + (-1)b + sa + tb = ((b - 1 + s)a + (t-1)b$. This is the required nonnegative combination for n. ∎

This result gives one solution to the original Putnam problem, with a guarantee that every rectangle whose sides are larger than 2,213 can be tiled: A 20×6 rectangle can be tiled by joining 5 even tiles, and a 20×7 rectangle can be tiled by joining 4 odd tiles. By applying Lemma 1 we see that a $20 \times n$ rectangle can be tiled for any $n \geq (7-1)(6-1) = 30$. Since a 35×5 rectangle and a 35×7 rectangle can be easily tiled, applying Lemma 1 we conclude that a $35 \times n$ rectangle can be tiled for any $n \geq 24$. Combining these two facts, we can tile a $55 \times n$ rectangle for any $n \geq 30$. Since rectangles with dimensions 42×4 and 42×5 can be tiled, Lemma 1 tells us that a $42 \times n$ rectangle can be tiled for any $n \geq 12$. Finally, since $\gcd(42, 55) = 1$, Lemma 1 again tells us that all $m \times n$ rectangles with $m \geq n \geq 54 \cdot 41 = 2{,}214$ can be tiled.

This proof guarantees that $L = 2213$ will suffice. One might try to improve on this lower bound. By generating the same rectangular strips in a different order, we can combine a tiling for a $35 \times n$ rectangle with $n \geq 24$ with a $42 \times n$ rectangle with $n \geq 12$ to get a $77 \times n$ rectangle with $n \geq 24$. Combining this with a $20 \times n$ rectangle for $n \geq 30$ we get solutions whenever $m \geq n \geq 76 \cdot 19 = 1444$. However no attempt along these lines can be expected to come close to the minimum value of $L = 33$, which will follow from Lemma 4 and Theorem 5.

A lower bound for L In this section we will establish a lower bound for L by showing that a 33×33 square cannot be tiled. Later we will show that this bound is tight by showing that any rectangle with both dimensions at least 34 can be tiled. In fact we actually state a slightly stronger result in the following theorem.

THEOREM 2. *For $28 \leq n \leq m$, the only $m \times n$ rectangles that cannot be tiled with tiles of size 4×6 and 5×7 are the rectangles of sizes 31×29, 33×32, and 33×33.*

We will prove the 31×29 and 33×33 rectangles cannot be tiled. Since the proof involving the 33×32 rectangle is lengthy and not needed for our original problem, we omit this proof.

LEMMA 3. *The 31×29 rectangle cannot be tiled using only 4×6 and 5×7 rectangles.*

Proof. There are two ways to write the area of the 31×29 rectangle as a linear combination of 24 and 35, namely $899 = 36 \cdot 24 + 1 \cdot 35 = 1 \cdot 24 + 25 \cdot 35$. Since these can be shown to be the only solutions using nonnegative integers, a tiling requires either 1 odd and 36 even tiles or 25 odd and 1 even tile. The combination with only 1 odd tile cannot be used since every row of the rectangle must contain an odd tile. We consider a tiling using the second combination and claim that the unique even tile cannot appear along a border. Without loss of generality assume the tile appears along the horizontal base. Then the tile immediately above it must have length either 4 or 6, a contradiction. Thus the even tile must appear in the interior forming a pinwheel design as shown in FIGURE 1.

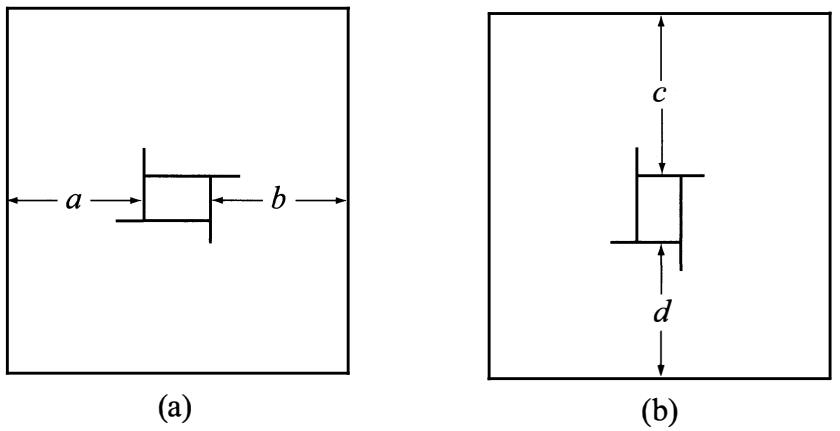


Figure 1 Even tile positioned in the interior

If the even tile is positioned as shown in FIGURE 1a, then $a + b$ must equal $29 - 6 = 23$. However, since 23 cannot be expressed as a nonnegative combination of 5 and 7, this configuration is impossible. Hence we assume the even tile is positioned as shown in FIGURE 1b. Then $c + d = 25$, and both c and d must be multiples of 5. Examining the column immediately to the left of the even tile, $c + 6$ must be a nonnegative combination of 5 and 7, which implies $c \geq 15$. Similarly the column immediately to the right of the even tile forces $d \geq 15$. This implies $c + d \geq 30$, a contradiction. ■

Next, we use a coloring argument to show that a tiling of the 33×33 square is impossible.

LEMMA 4. *The* 33×33 *square cannot be tiled using only* 4×6 *and* 5×7 *rectangles.*

Proof. First we give the unique ways to express the area of this square as a linear combination of the areas of the tiles: $33^2 = 1089 = 3 \cdot 35 + 41 \cdot 24 = 27 \cdot 35 + 6 \cdot 24$. This means that a tiling of the square requires either 3 odd and 41 even tiles or 27 odd and 6 even tiles. The odd length of the square requires at least one odd tile in every row, which would require at least 5 odd tiles, so we must take the second combination with 27 odd tiles and 6 even tiles.

Consider a potential tiling of the 33×33 square using 27 odd tiles and 6 even tiles. Let H denote the number of odd tiles with their side of length 7 positioned horizontally and let V be the number of tiles with their side of length 7 positioned vertically. Thus $H + V = 27$. By rotating the entire tiling 90° if needed, we may assume that $H \leq 13$. Similarly, the even tiles give the equation $h + v = 6$. Consider the average number of odd tiles per row. Because each horizontal tile appears in 5 rows and each vertical tile appears in 7 rows, this average is $(5H + 7V)/33 = (189 - 2H)/33 = 5 + (24 - 2H)/33$. Since $7 \cdot 5 > 33$, and 6 odd tiles appearing in the same row would cover all but one or three columns, the maximum number of odd tiles in any row is 5. Hence the average cannot exceed 5 and thus $H \geq 12$. We conclude that $12 \leq H \leq 13$.

Color the rows so that for $0 \leq i \leq 32$, row i gets color $i \bmod 5$ for distinct colors 0, 1, 2, 3, 4. There will be 231 cells colored 0, 1 and 2, and 198 cells colored with each of the colors 3 and 4. Represent the number of cells of each color in the square by a *color vector* $\langle 231, 231, 231, 198, 198\rangle$. Every horizontal odd tile covers 7 cells of each color, represented by the color vector $\langle 7, 7, 7, 7, 7\rangle$. A vertical odd tile covers 10 cells of one color, 10 cells of a second color, and 5 cells of each of the three remaining colors. We say that a vertical odd tile covers a rotation of $\langle 10, 10, 5, 5, 5\rangle$. Horizontal even tiles cover a rotation of $\langle 6, 6, 6, 6, 0\rangle$ and vertical even tiles cover a rotation of $\langle 8, 4, 4, 4, 4\rangle$. Note that $\langle 10, 10, 5, 5, 5\rangle$ is the sum $\langle 5, 5, 5, 5, 5\rangle + \langle 5, 5, 0, 0, 0\rangle$, $\langle 6, 6, 6, 6, 0\rangle$ is $\langle 6, 6, 6, 6, 6\rangle + \langle 0, 0, 0, 0, -6\rangle$, and $\langle 8, 4, 4, 4, 4\rangle$ is $\langle 4, 4, 4, 4, 4\rangle + \langle 4, 0, 0, 0, 0\rangle$. The seven possible values for h and two values for H yield 14 cases. In each case we subtract the uniform vector $(7H + 5V + 6h + 4v)\langle 1, 1, 1, 1, 1\rangle$ from the color vector to leave a *remainder vector* that must be covered with rotations of the residual terms of $\langle 5, 5, 0, 0, 0\rangle$, $\langle 0, 0, 0, 0, -6\rangle$ and $\langle 4, 0, 0, 0, 0\rangle$. The 14 cases and their remainder vectors are given in TABLE 1. We show that each case fails to give a solution for one of two reasons given below.

Consider what happens when we try to form the remainder vectors from the available residual terms modulo 5. The residual for each odd vertical tile reduces to the zero vector mod 5, so it offers no help in creating the remainder vector. Each even tile, whether in horizontal or vertical position, contributes a rotation of $\langle 0, 0, 0, 0, -1\rangle$. This means that we can only form a remainder with no more than six adjustments of $-1 \bmod 5$ among all 5 columns. However, we need 11 adjustments for each of cases 4, 5, 6, 7, and 14 in TABLE 1, and 16 adjustments for each of the cases 2, 3, 12, and 13. Hence, none of these cases yield a valid solution.

For cases 1, 8, 9, 10, and 11, a solution mod 5 is possible, but it cannot lead to a nonnegative solution. For instance, in case 1, six adjustments of $\langle 4, 0, 0, 0, 0\rangle$ must be rotated and summed to contribute $\langle 8, 8, 8, 0, 0\rangle$ and thus leave $\langle 40, 40, 40, 15, 15\rangle$ to be covered by the 15 terms of type $\langle 5, 5, 0, 0, 0\rangle$. Suppose a times we use the residual as written to adjust the first two columns, b times we rotate it one step to the right, c times we rotate it two steps, d times we rotate it three steps, and e times we rotate it four steps. Then the 15 vertical odd tiles contribute $5\langle a + e, a + b, b + c, c + d, d + e\rangle$. This gives the unique solution, $a = b = c = e = 4$ and $d = -1$. Since d is negative, the solution cannot be realized. Similarly, case 8 requires using

TABLE 1: Fourteen possible cases for the 33 × 33 problem

Case	h	v	H	V	Remainder Vector					Reason for No Solution
1	0	6	12	15	48	48	48	15	15	$d = -1$
2	0	6	13	14	46	46	46	13	13	No solution mod 5
3	1	5	12	15	46	46	46	13	13	No solution mod 5
4	1	5	13	14	44	44	44	11	11	No solution mod 5
5	2	4	12	15	44	44	44	11	11	No solution mod 5
6	2	4	13	14	42	42	42	9	9	No solution mod 5
7	3	3	12	15	42	42	42	9	9	No solution mod 5
8	3	3	13	14	40	40	40	7	7	$d = -2$
9	4	2	12	15	40	40	40	7	7	$d = -1$
10	4	2	13	14	38	38	38	5	5	$d = -2$ or -6
11	5	1	12	15	38	38	38	5	5	$d = -3$ or -5
12	5	1	13	14	36	36	36	3	3	No solution mod 5
13	6	0	12	15	36	36	36	3	3	No solution mod 5
14	6	0	13	14	34	34	34	1	1	No solution mod 5

six adjustments of $\langle 4, 0, 0, 0, 0\rangle$ and $\langle 0, 0, 0, 0, -6\rangle$ in columns 4 and 5 to leave either $\langle 40, 40, 40, 5, 15\rangle$ or $\langle 40, 40, 40, 15, 5\rangle$. Either solution includes $d = -2$. In case 9, we reduce to $\langle 40, 40, 40, 15, 15\rangle$, $\langle 40, 40, 40, 25, 5\rangle$ or $\langle 40, 40, 40, 5, 25\rangle$; both lead to $d = -1$. In case 10 we may assign our four 6s and two 4s in three different ways to the first three columns. However, all three lead to solutions with either $d = -6$ or $d = -2$. Finally, case 11 leads to $\langle 40, 50, 50, 5, 5\rangle$, $\langle 50, 40, 50, 5, 5\rangle$ or $\langle 50, 50, 40, 5, 5\rangle$ with $d = -3$ or -5. Note that only two simple ideas have been used to eliminate all 14 cases.

Hence the 33 × 33 square cannot be tiled. ■

The best lower bound We now show that any rectangle with $34 \leq n \leq m$ can be tiled, which will establish that $L_s \leq 33$. In fact, the reasoning shows the slightly stronger result, that any rectangle with $28 \leq n \leq m$ and not having dimensions 31×29, 33×32 or 33×33 can be tiled.

THEOREM 5. *Any rectangle with* $34 \leq n \leq m$ *can be tiled using only* 4×6 *and* 5×7 *rectangles.*

Proof. Consider a rectangle where $n \leq 39$ and $m \leq 46$ and is not one of the three forbidden sizes given above. The scheme for tiling the rectangle is given in TABLE 2. In this table the columns represent values of n and the rows represent values of m.

We use various notation to describe different strategies for tiling rectangles. An entry $m * n$ in the cell for an $m \times n$ rectangle indicates the existence of a tile whose dimensions divide the dimensions of the rectangle. For example, the entry $30 * 28$ indicates a tile whose dimensions (here 6 and 4) divide into the respective dimensions of the rectangle (here 30 and 28). Consequently, we can stack even tiles in a 5×7 array to tile the rectangle. We use the formula $a * (b + c)$ to indicate a rectangle tiled by partitioning it into smaller $a \times b$ and $a \times c$ rectangles, both of which are easily tiled. For example, the entry $(14 + 16) * 30$ indicates that a 30×30 rectangle is tiled by joining a 14×30 rectangle with a 16×30 rectangle. An entry $(a + b) * c$ means a rectangle can be tiled by partitioning it into smaller $a \times c$ and $b \times c$ rectangles. For example, the entry $30 * (21 + 8)$ can be tiled by joining a 30×21 rectangle to a 30×8 rectangle.

TABLE 2: Tiling schemes

m / n	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
28	$(10+18)*28$												
29	$(5+24)*28$	FIGURE 4											
30	$30*28$	$30*(21+8)$	$(14+16)*30$										
31	$(25+6)*28$	Impossible	$(7+24)*30$	P(19; 8; 7; 16)									
32	$(20+12)*28$	$(32*12)$ +P(6; 5; 6; 5)	$\Uparrow 4$	P(6; 7; 12; 19)	$(12+20)*32$								
33	$\Uparrow 5$	P(6; 5; 12; 17) Includes 27×27 P(6; 5; 6; 5)		P(8; 19; 4; 7)	Impossible	Impossible							
34		P(20; 8; 10; 15)		P(14; 6; 14; 21)	$\Uparrow 6$	P(10; 14; 10; 14)							
35		$35*(14+15)$		$35*(10+21)$	$\Uparrow 6$	$\Leftarrow 5$	$\Leftarrow 5$	$\Uparrow 5$					
36		$36*(12+17)$		$36*(12+19)$	$\Uparrow 6$	$\Leftarrow 4$	$\Leftarrow 4$		$\Uparrow 4$				
37		$37*(14+15)$		P(8; 7; 22; 19)	$\Uparrow 6$	P(10; 14; 20; 14)	$37*(14+20)$			$37*(20+17)$			
38		P(14; 4; 14; 4)		$\Uparrow 10$	$\Uparrow 6$	$\Uparrow 10$	$38*(10+24)$			P(14; 8; 14; 22)	$\Uparrow 10$		
39		FIGURE 4		$39*(12+19)$	$39*(12+20)$	$\Uparrow 10$	P(18; 10; 15; 20)			P(15; 25; 4; 7)		FIGURE 4	
40		$40*(7+10+12)$		$\Uparrow 10$	$\Uparrow 6$	$\Uparrow 10$	$\Uparrow 12$			$\Leftarrow 7$		$\Uparrow 12$	$\Leftarrow 12$
41		P(7; 14; 41; 21)				$\Uparrow 10$				P(20; 25; 6; 5)			$\Leftarrow 12$
42		$\Uparrow 14$				$\Leftarrow 4$				$\Uparrow 14$			$\Leftarrow 12$
43		$\Uparrow 14$				$43*(12+21)$							$\Leftarrow 12$
44		$\Uparrow 14$				$\Uparrow 10$							$\Leftarrow 12$
45		$45*(12+17)$											$\Leftarrow 12$
46		$\Uparrow 14$											$\Leftarrow 12$

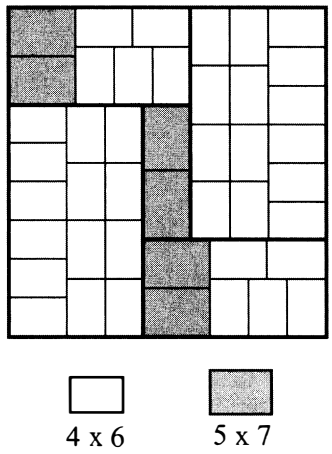


Figure 2 Pinwheel tiling of a 34 × 33 rectangle

An entry ⇑ means that we can express this rectangle as the union of two smaller rectangles that can be tiled. For example, the table entry for 33 × 28 is ⇑ 5, indicating that we can take the union of the tiling given 5 cells up in the table along with a 5 × 28 rectangle. Notice that the last entry listed in each column is an ⇑ strategy. The blank entries below it mean that the same ⇑ strategy continues for *all* larger values of m. We use the symbol ⇐ in a similar fashion.

Note that rectangles described with the ⇑ or ⇐ symbols have fault lines. That is, they can be partitioned into two rectangles that both have the width or height of the original rectangle. Other rectangles will not have such fault lines. For example FIGURE 2 shows a tiling of a 34 × 33 rectangle, which can be viewed as five subrectangles, one in the center and four others forming the boundary.

We denote such a pinwheel design by $P(a; b; c; d)$ where a, b, c, and d are defined in FIGURE 3.

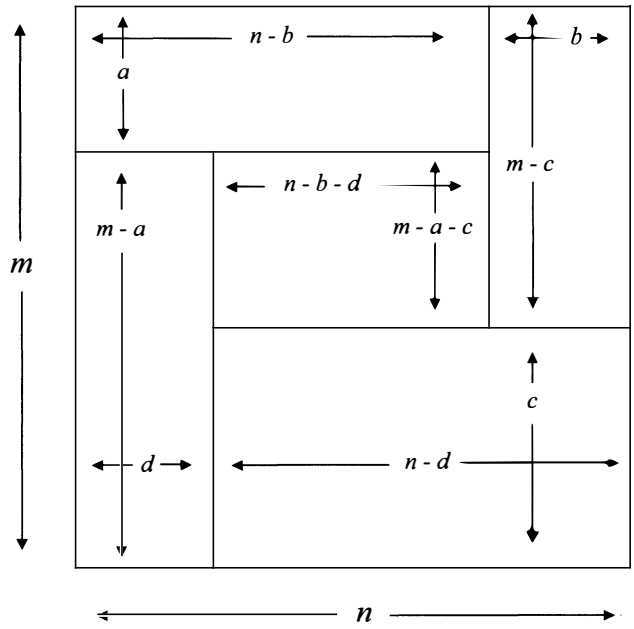


Figure 3 Pinwheel construction $P(a; b; c; d)$ for an $m \times n$ rectangle

Here six parameters are needed to deduce all of the relevant lengths. Four of these parameters appear explicitly in the pinwheel notation while m and n are suppressed since they are given by the cell in the table where the notation appears.

Finally, we use a type of decomposition to show that the remaining rectangles can be tiled. We first show that an $m \times 12$ strip can be tiled for any $m \geq 38$. Note that we can tile an $m \times 12$ rectangle for any even $m \geq 4$ using only even tiles by $4k * 12$ or $((4k-4)+6) * 12$. To accommodate odd values of m, first tile a 35×12 strip using only odd tiles (by $35 * (7+5)$), and then append an $m \times 12$ strip with m even. Consequently we can tile $m \times 12$ rectangles whenever $m \geq 38$. Thus we can reduce n, 12 at a time, until we reach a value in the range $28 \leq n \leq 39$. Finally, if $m \geq 47$, we reduce it until we reach a value whose tiling appears in TABLE 2. The size of the reducing strip depends upon the column n lies in. For example, if n is 30 or 36, we remove $4 \times n$ strips until the values of m and n are such that the rectangle is given in TABLE 2. Similarly, for $n = 28$ or 35, we remove $5 \times n$ strips; for $n = 32$, we remove $6 \times n$ strips; for $n = 31, 33$, or 38, remove $10 \times n$ strips; for $n = 34$ or 39, remove $12 \times n$ strips; finally, for $n = 29$ or 37 remove $14 \times n$ strips. In all cases, the process terminates when we reach a table entry that assigns a tiling solution. Thus, with the exception of the three cases 31×29, 33×32, and 33×33, every $m \times n$ rectangle with $28 \leq n \leq m$ can be tiled. ■

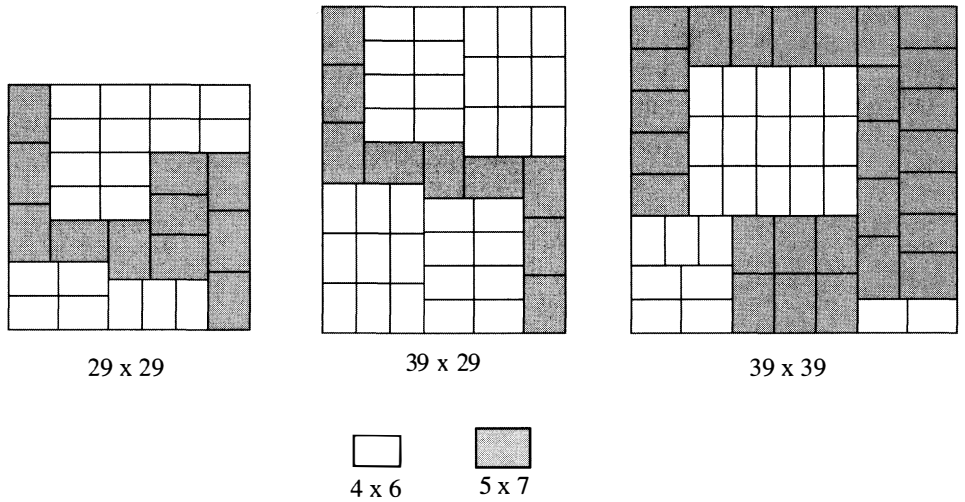


Figure 4 Three rectangles with unusual tiling patterns

Lemma 4 and Theorem 5 imply that all $m \times n$ rectangles with both m and n greater than 33 can be tiled and the 33×33 square cannot be tiled. Hence the minimum value of L that satisfies the Putnam problem is 33.

Acknowledgments. Preliminary work was supported by a National Science Foundation Research Experience for Undergraduates Program at the University of Dayton. Thank you to Kim Slawson for assistance with the imaging and resolution of FIGURES 1–4. The authors would also like to thank the referees for their valuable comments.

REFERENCES

1. F. R. K. Chung, E. N. Gilbert, R. L. Graham, and J. H. van Lint, Tiling rectangles with rectangles, this MAGAZINE **55** (1982), 286–291.
2. Solomon W. Golomb, *Polyominoes* (2nd. ed.), Princeton University Press, 1994.
3. L. F. Klosinski, G. L. Alexanderson, and L. C. Larson, The Fifty-Second William Lowell Putnam Mathematical Competition, *Amer. Math. Monthly* **9** (1992), 715–724.
4. Peter J. Robinson, Fault-free rectangles tiled with rectangular polyominoes, *Combinatorial Mathematics*, IX (Brisbane, 1981), 372–377.
5. J. J. Sylvester, *Mathematical Questions from the Educational Times* **41** (1884), 21.
6. Y. Vitek. Bounds for a linear Diophantine problem of Frobenius, *J. London Math. Soc.* **10**:2 (1975), 79–85.

Unique Rook Circuits

RICHARD K. GUY
MARC M. PAULHUS
University of Calgary
Calgary, Alberta T2N 1N4
Canada

Suppose we have an array of squares, such as on a chessboard, with possibly some of them blocked off. A *rook tour* is a path that visits every empty square exactly once, moving at each step to any empty adjacent square (North, South, East, or West, but not diagonally). We will call a rook tour a *rook circuit* if it starts and ends on the same square. As you can see from FIGURE 1, given an array with a set of forbidden squares, there may be: (a), no circuit; (b) and (c), more than one circuit; or (d), exactly one circuit. We are especially interested in this last case.

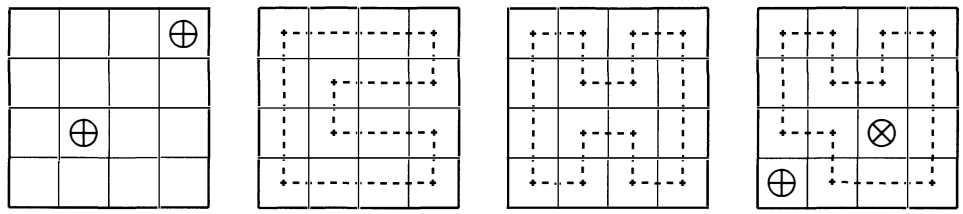


Figure 1 Some examples of rook circuits

How do we justify such statements? If we checker the squares of the array, we see that each unit step in a rook circuit goes from a white square to a black square, or vice versa. This gives us

THE PARITY PRINCIPLE. The numbers of black and white squares in a rook circuit must be equal. Moreover, the number of vertical (unit) steps and the number of horizontal steps are both even.

In example (a) the two forbidden squares are both black squares, so there is no hope of a circuit, or even a tour. Forbidden squares are indicated by guideposts: $\otimes$ on white squares, and $\oplus$ on black squares.

Examples (b) and (c) make it clear that arrays of size 4×4 or bigger with an even number of squares, none of them forbidden, always have more than one circuit.

Example (d) has a unique circuit. This may not be immediately obvious, but is easily checked using

THE TWO NEIGHBOR PRINCIPLE. If a square has only two neighbors, then it must be visited via those neighbors.

For example, apply this to corner squares a, d, i, l, n in FIGURE 2, and also squares k and m. This gives us the 11 steps b–a–e–i–j–l–m–n–k–h–d–c, and now squares f and g have only two neighbors and the circuit is determined.

a	b	c	d
e	f	g	h
i	j	⊗	k
⊕	l	m	n

Figure 2 Uniqueness of rook circuit

Rook circuit problems originated with Sidney Kravitz who posed Problem 2212 [**1**] (see also [**2**]). He asked the readers to find the unique rook circuit in FIGURE 3.

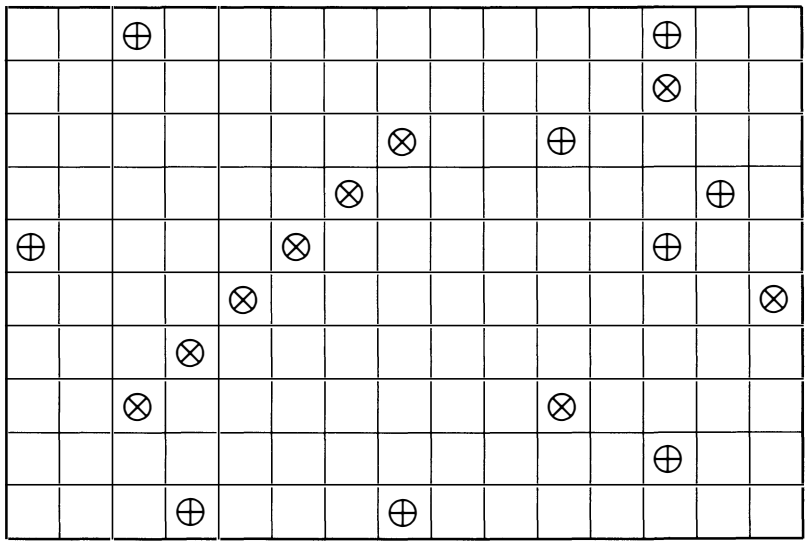


Figure 3 Sidney Kravitz's problem: a 10 × 15 array with 18 prohibited squares

An open question was: Is there a configuration of fewer than 18 prohibited squares that forces a unique rook circuit on a 10 × 15 array? The answer is "Yes," and the reader can verify that the configuration given in FIGURE 4, with only 10 guideposts, has a unique rook circuit.

Uniqueness can be exhibited by drawing a path in which every segment is forced. The reader wishing to verify uniqueness of the paths in FIGURES 3 and 4 should be armed with

THE CUL-DE-SAC PRINCIPLE. Never draw segments that leave a square with only one exit; and

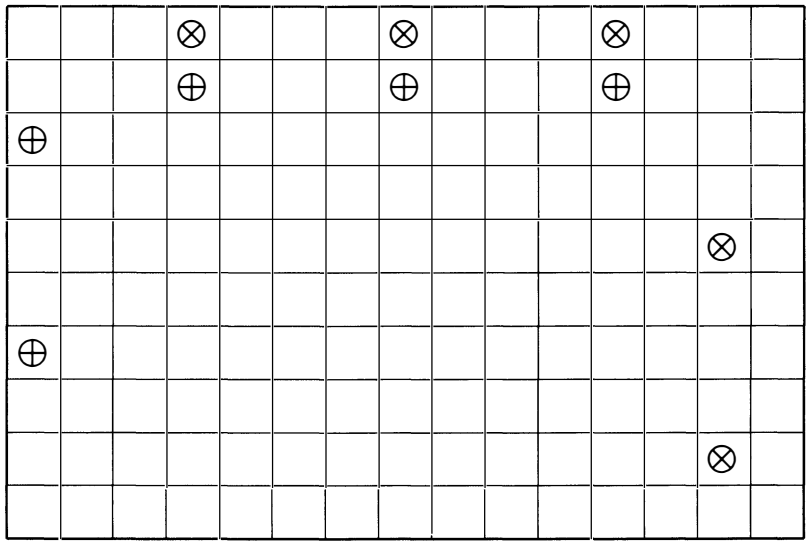


Figure 4 Best known solution, with 10 guideposts, for a 10 × 15 array

THE EARLY CLOSING PRINCIPLE. Never close a circuit unless all squares have been visited.

A 4-*corner* is a 4 × 4 array of squares in a corner of a larger array. Note that none of the configurations with unique rook circuits has an empty 4-corner. The best general result we know for rook circuits was found by the second author, namely,

THE 4-CORNER PRINCIPLE. If there is a unique circuit in an $m \times n$ array, then at least one square in each 4-corner must contain a guidepost.

Our proof is tedious even when we omit some steps; can the reader find a shorter one?

Proof. Suppose there is an empty 4-corner in an $m \times n$ array with a unique rook circuit. Our proof amounts to checking each of the finitely many ways in which a circuit can enter and exit the 4-corner; in every case, two circuits will be shown with the same exits and entrances, contradicting the uniqueness of the circuit. Parity considerations tell us that the circuit will cross the boundary of the 4-corner exactly 2, 4, 6 or 8 times. The case with two crossings is relatively simple; the one with 6 crossings is quite tedious. Label the exits from a 4-corner as in FIGURE 5.

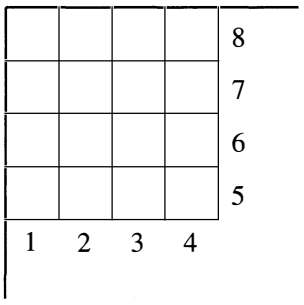


Figure 5 Labels for the exits of a 4-corner

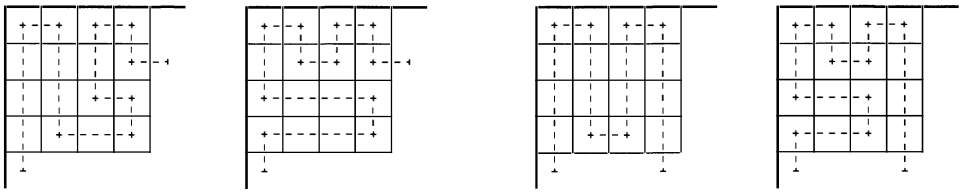


Figure 6 Cases {1, 7} and {1, 4}

First, suppose that the circuit crosses the boundary of the 4-corner exactly twice. Two examples are shown in FIGURE 6, where the crossings are at exits {1, 7} and at {1, 4}. In each case, we see two alternative paths showing that any circuit with those prescribed exits is not unique.

We do not need to consider the case {1, 3}, for example, since 1 and 3 are similarly colored squares and a path that enters on one color must leave on the other. Furthermore, the counterexamples constructed for exits {1, 4} will also work for {1, 5}. Also, those for {1, 7} will work for the case {2, 8}, by diagonal reflection.

In this way we can reduce the cases that still need to be considered to: {1, 2}, {2, 3}, {2, 6}, and {3, 4}, which we leave to the reader.

Now suppose that the circuit crosses the boundary of a 4-corner exactly four times. If the exits, listed in numerical order, are a, b, c, d, then, to form a single circuit, they must be connected ab, cd inside the 4-corner, and ad, bc outside, or vice versa. For example, if the circuit crosses at exits {1, 2, 3, 7}, we need to consider the two cases illustrated in FIGURE 7. The first is where exits {2, 3} (braced in the figure) and exits {1, 7} are connected outside the 4-corner, so that {1, 2} and {3, 7} are connected inside. In the second, the roles are interchanged, with exits {1, 2} (braced) and {3, 7} connected outside. We don't need to consider the case where {1, 3} and {2, 7} are connected since that would cause the circuit to intersect itself.

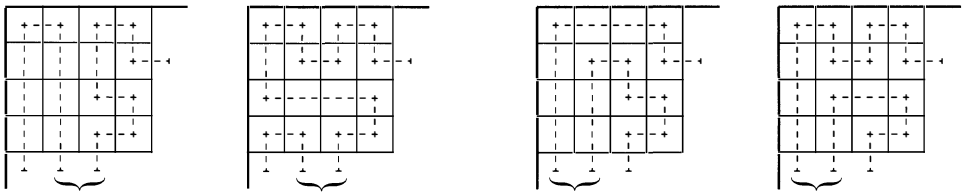


Figure 7 Cases {{1, 7}, {2, 3}} and {{1, 2}, {3, 7}}

We may abbreviate the description of these two cases to $abcd = 1237$. Symmetry and color considerations reduce our need to consider only the nine cases $abcd = 1234$, 1237 (above), 1246, 1248, 1267, 1278, 1347, 1467, and 2346, together with four others, in which exits 4 and 5 are both used. When this happens, 4 and 5 are necessarily connected inside the 4-corner, so the case described above as ad, bc is not relevant. The outside connections are ab, cd for $abcd = 1456$, 1458, 1543, and 3456. Twenty more pairs of alternative paths, in addition to the two pairs shown, suffice to complete this case.

If the circuit crosses the boundary of the 4-corner at six places, list them in numerical order as a, b, c, d, e, and f. Then these must be joined in pairs, inside and outside the 4-corner, in one of the following five ways:

$$ab, cd, ef \qquad af, bc, de$$

$$ab, cf, de \qquad ad, bc, ef \qquad af, be, cd.$$

In order to form a single circuit, the two in the first row must be linked, one inside, the other outside, and each of the three in the second row must be linked to one of the other two. We consider the type in the second row first: If we appeal to symmetry, we need consider only one such complete set of three:

$$14, 23, 78 \qquad 12, 38, 47 \qquad 18, 27, 34$$

Each pairing can be in-out or out-in for a total of six cases.

The following three pairs of links can be connected to one another in-out or out-in. Either one of the third pair can also be an *in* associated with the *out* 14, 23, 67, which cannot be used as an in.

$$12, 34, 67 \;\&\; 17, 23, 46 \qquad 12, 34, 78 \;\&\; 18, 23, 47 \qquad 12, 37, 46 \;\&\; 17, 26, 34$$

There are four ins with a unique associated out, that is, in & out = $12, 45, 68$ & $18, 24, 56$ or $13, 45, 67$ & $17, 34, 56$ or $16, 23, 45$ & $12, 34, 56$ or $18, 23, 45$ & $12, 34, 58$ and four ins with two possible associated outs:

12, 36, 45	with	14, 23, 56	or	16, 25, 34
12, 38, 45	with	14, 23, 58	or	18, 25, 34
17, 36, 45	with	13, 47, 56	or	15, 34, 67
18, 26, 45	with	12, 48, 56	or	15, 24, 68.

The assiduous reader will find 22 pairs of alternative paths.

It's hard to visualize how all eight exits might be used, but if they are, then exits 4 and 5 are again connected inside. So we need consider, apart from reflection, only the following outside connections:

12, 34, 58, 67 12, 38, 47, 56 18, 27, 34, 56	which may be joined in more than one way to form a single circuit via the inside connections 16, 23, 45, 78,
14, 23, 58, 67 18, 23, 47, 56	which have ambiguous inside connections via 12,36,45,78, and finally,
12, 34, 56, 78	via 18, 23, 45, 67.

Just six pairs of pictures for these cases will complete the proof. ■

Related results You might think that a statement similar to the 4-corner principle must be true if you consider large enough arrays that are not in a corner, but our general results show that no such statement is true: you can have an arbitrarily large area of an array without guideposts but with only one circuit! Moreover, if you look at the FIGURE 4, you'll see that a 3-corner principle can't be true. The same figure also rules out the statements that a 5-, 6-, or 7-corner must contain two guideposts. It seems probable that an 8-corner must contain two guideposts but we have made no attempt to prove this.

It's not too hard to specify exactly how many guideposts there must be to have a unique rook circuit on a board with one side of length less than 5, or one that is a multiple of 4.

Rectangles of size $1 \times n$ have no rook circuit, and those of size $2 \times n$ have a unique one.

For rectangles of size $3 \times (3n + j)$ with $n > 0$ and $j = 0$ or ± 1, it is necessary and sufficient to use $n + j$ guideposts to enforce a unique circuit. You can check this by noting that consecutive posts have to be on squares of opposite colors. One way of placing them is

for	$j = 0$,	put	n	guideposts at	$(3r + 1, 1)$	for $0 \le r < n$
	$j = 1$,		$n + 1$		$(3r + 1, 1)$	$0 \le r \le n$
	$j = -1$,		$n - 1$		$(3r, 1)$	$0 < r < n$.

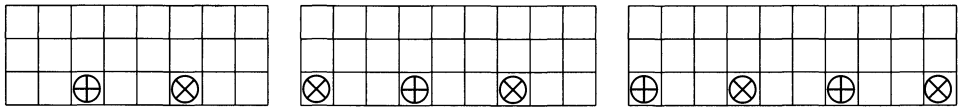


Figure 8 Find the unique rook circuits

It is curious that, in order to force a unique rook circuit on a $3 \times n$ array, about $1/9$ of its squares need to be occupied by guideposts, whereas larger boards require smaller fractions. Rectangles of size $4 \times n$ only need 2 guideposts, if you place them either at $(1, 1)$ and at $(n-1, 2)$ or $(n-1, 3)$ according as n is even or odd.

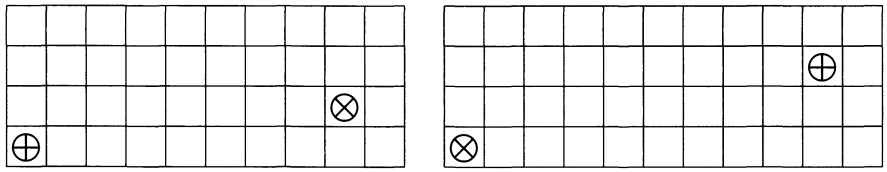


Figure 9 Check that the rook circuits are unique

If one dimension of the rectangle is a multiple of 4, say $4m$, then $2m$ guideposts suffice, if you stack m copies of the $4 \times n$ examples that we've just seen. For example, the standard 8×8 chessboard has a unique rook circuit if 4 guideposts are placed as in FIGURE 10.

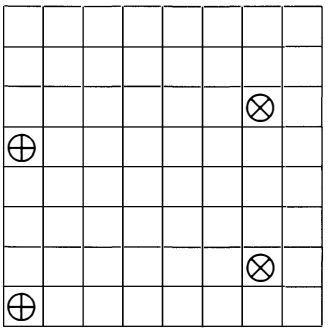


Figure 10 Check that the circuit is unique

We are less certain of the minimum number of guideposts if the dimensions of the rectangle are $4m + i$ by $4n + j$, with $m, n > 0$ and $i, j = 1, 2$, or 3. We believe them to be

$2(m + n) - 1$	if $ij = 1$	(first diagram of FIGURE 11)
$2(m + n)$	if ij is even	(next three diagrams)
$2(m + n) + 1$	if $ij = 3$ or 9	(last row of FIGURE 11)

The fact that these numbers suffice is illustrated by the six diagrams of FIGURE 11: for $m = n = 2$ and $(i, j) = (1, 1)$, $(1, 2)$ (first row); $(2, 2)$, $(2, 3)$ (second row); $(3, 1)$, $(3, 3)$ (last row). For the cases $(i, j) = (2, 1)$, $(3, 2)$, or $(1, 3)$, reflect the appropriate diagram about its rising diagonal.

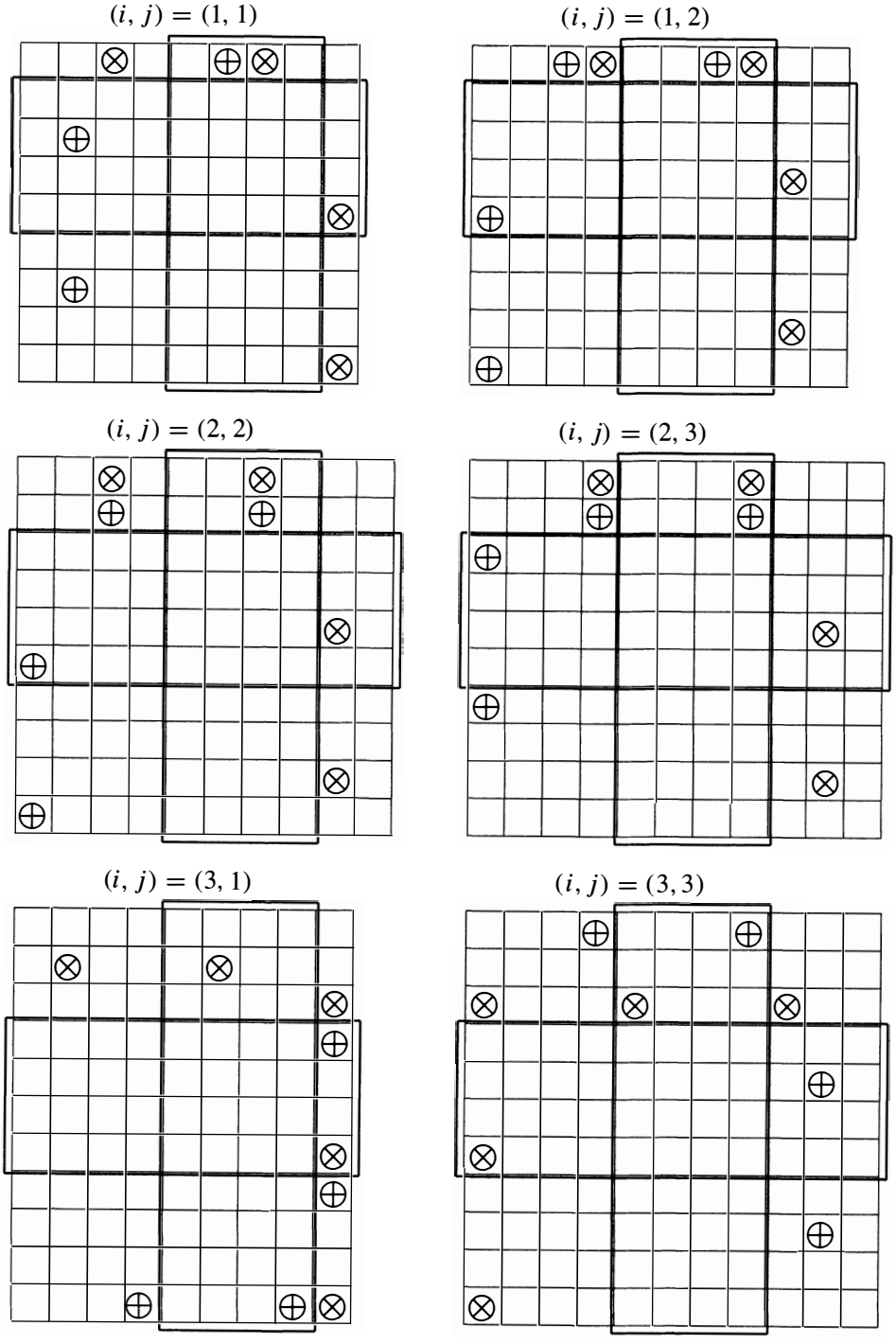


Figure 11 Find the unique rook circuits

In each diagram the framed $(4m + i) \times 4$ and $4 \times (4n + j)$ rectangles each contain 2 guideposts, one of each color. These may be deleted (in order to reduce m or n to 1), or replicated (to cover cases where m or n is greater than 2) producing rectangles of different height or width. Enjoy verifying the uniqueness of the rook circuit in each case. Here are some other problems.

1. Force unique circuits on nonrectangular boards.
2. Force unique circuits on different tilings, for example, triangular and hexagonal. Hamilton might have wondered about the dodecahedron or icosahedron: here, should rotations and reflections be counted as different?
3. What if the array of blocks is a tiling of a torus, a cylinder, a Möbius band, a Klein bottle, or a projective plane?
4. What is the maximum number of nonadjacent squares that you can block from an array and still permit a rook circuit?
5. Study analogs of rook circuits on chessboards of dimension three or higher.

Acknowledgments. Special thanks to Sidney Kravitz, Loren Larson, and Bill Sands for various insights.

REFERENCES

1. Sidney Kravitz, Complete the circuit, *J. Recreational Math.* **28**:2 (1996–97), 143.
2. Paul Vaderlind, Richard Guy, and Loren Larson, *The Inquisitive Problem Solver*, Math. Assoc. of Amer. Problem Books, 2002, Problem 88.

30 Years Ago in the Magazine

From "Who Reads the Mathematics Magazine?" by L. J. Cote and R. P. O'Malley, Purdue University, **45**:5 (1972), 273–278.

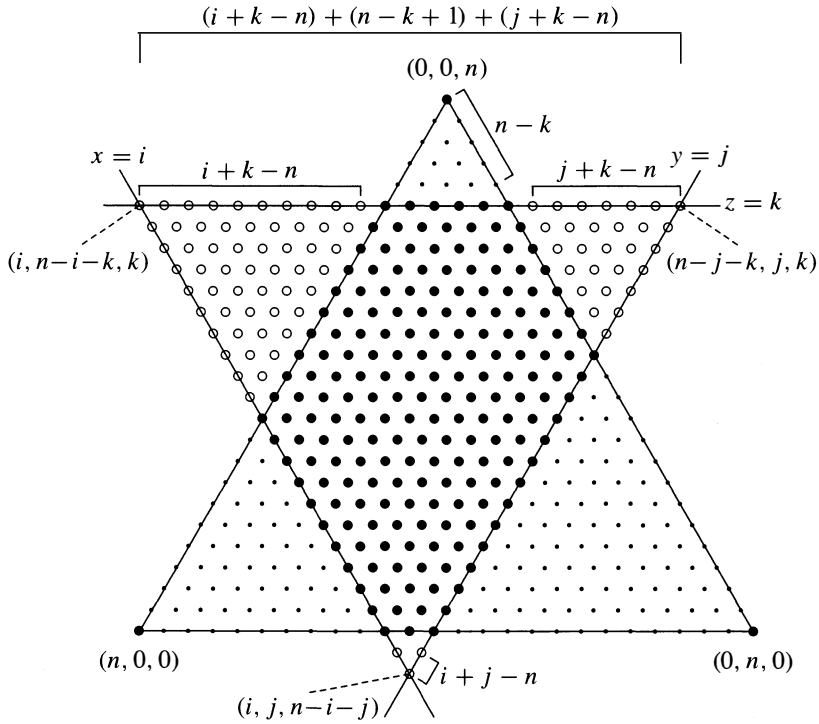
> Who reads MATHEMATICS MAGAZINE? For the most part, teachers in universities or colleges not granting the Ph.D. in mathematics, who read and enjoy the Problem Section regularly, who are interested in book reviews, and who probably tend to choose which articles they will read by scanning the list of titles on the cover to choose those they think they will find most interesting. The other two groups of teachers seem to be quite similar to this group. The surprisingly large group of nonacademic subscribers is also similar to the teachers....

Readers are encouraged to send their comments about the MAGAZINE to mathmag@scu.edu. Is there interest in another survey of readers?

Proof Without Words: Nonnegative Integer Solutions and Triangular Numbers

MATTHEW J. HAINES
Augsburg College
Minneapolis, MN 55454

MICHAEL A. JONES
Montclair State University
Upper Montclair, NJ 07043



For i, j, and k integers between 0 and n inclusive, the number of nonnegative integer solutions of

$$x + y + z = n \qquad x \le i \qquad y \le j \qquad z \le k,$$

is

$$T_{i+j+k-n+1} - T_{i+k-n} - T_{j+k-n} - T_{i+j-n},$$

where T_m is the mth triangular number:

$T_m = 0$ $(m \le 0)$ $\quad T_1 = 1 \quad T_2 = 3 \quad T_3 = 6 \quad \cdots.$

Encryption Using a Variant of the Turning-Grille Method

STEPHEN FRATINI
50 Malibu Drive
Eatontown, NJ 07724

In his *De Subtilitate* [**1**], the sixteenth-century mathematician Girolamo Cardano introduced a new method of encryption/decryption. To read an encrypted text, a square grid (or grille) with the correct openings was placed over the text to reveal the decrypted message. Variants of this technique were used as late as the mid-twentieth century. For example, German intelligence used a grid-based encryption in South America during World War II. In this note, we present other variants of this method and describe some modern encryption techniques in conjunction with grids.

By a *grid* we will mean a rectangular array that covers a rectangular text (of the same size) with small openings or cutouts in various places. The dimensions of a grid determine the number of characters that it can cover. For example, a 16×16 grid can cover 16 lines of text, each line having 16 characters. In Cardano's method, we place a grid over a blank piece of paper and write the characters of our message in the grid openings. The grid is lifted off the paper and random characters are added around the message, thereby creating a rectangular block of text of the same dimension as the grid. The intended recipient of our message also has an exact copy of the grid used to encrypt the message. The message recipient places his or her grid over the block of text and the message appears through the grid openings.

A variant of Cardano's approach uses rotations of the grid to introduce encoded text more compactly than in the original approach. For example, in FIGURE 1, we show an 8×8 grid with 16 openings. When placed over an 8×8 text, this grid exposes 16 characters. If we rotate the grid 90° in the clockwise direction and again place the grid over the selection of text, a different set of 16 characters is shown. Two more rotations of 90° will expose two more sets of 16 previously covered characters. The 8×8 squares of FIGURE 1 are in four groups of 16 squares, and the numbering in each quadrant is rotated 90° clockwise each time. Using this numbering, we see that each of the 16 openings (under each rotation) has a distinct number (from 1 to 16). This ensures that all of the 64 squares of the text will be uncovered in a unique rotation of the grid. Text can be put in encrypted form after each rotation, so that the entire 8×8 array may contain up to 64 characters of (encrypted) text, without needing to insert any

1	2	3	4	13	9	5	1
5	6	7	8	14	10	6	2
9	10	11	12	15	11	7	3
13	14	15	16	16	12	8	4
4	8	12	16	16	15	14	13
3	7	11	15	12	11	10	9
2	6	10	14	8	7	6	5
1	5	9	13	4	3	2	1

Figure 1 An 8×8 grid

random characters. This encoding technique is referred to as the *turning-grille* method. For further reading, we recommend the excellent introductions to the turning-grille method and its associated history in the books by Gardner [**3**] and Kippenhahn [**5**].

Example: the Turning-Grille Method Consider the following phrase:

THE CONCEPT OF USING GRIDS TO ENCODE MESSAGES WAS FIRST DISCOVERED BY THE MATHEMATICIAN GIROLAMO CARDANO

We arrange the characters of the message in 8 rows of 8 characters (blank spaces between words are removed). If there are more than 64 characters, we use additional 8×8 arrays until all the text has been encoded.

The plaintext (that is, unencrypted text) is inserted into the arrays as suggested by FIGURE 2 from left to right, and from top to bottom, starting with the array on the left and then placing the remaining characters into the array on the right. Next, we place the grid shown in FIGURE 1 over the left array and read off the characters appearing through the openings (again, we go from left to right, and from top to bottom). We repeat the procedure after rotating the grid 90°, 180°, and finally 270°, and then go through the entire process again for the array on the right.

T	H	E	C	O	N	C	E
P	T	O	F	U	S	I	N
G	G	R	I	D	S	T	O
E	N	C	O	D	E	M	E
S	S	A	G	E	S	W	A
S	F	I	R	S	T	D	I
S	C	O	V	E	R	E	D
B	Y	T	H	E	M	A	T

H	E	M	A	T	I	C	I
A	N	G	I	R	O	L	A
M	O	C	A	R	D	A	N
O							

Figure 2 Encoding with an 8×8 grid

There are several approaches that we can use when the last block of plaintext does not fill the grid: fill the grid with dummy text, partially fill the grid with dummy text, or do not add any fill. We'll use the no-fill method in this example. For readability, we present the encoded text in groups of four. The actual coded message would be transmitted without spaces. The following encoding is obtained:

TTOS NSOE GRIV EDHE EITE NCOS SSTD YTMA COPF UGDD ASIS ORET
HENC GRIM EESW AFCB
HNGO ADNI LAOA TAIR MREM ICOC A

Of course, both the encoding party and the decoding party need to know how the characters are placed into the arrays and how the characters are read off when the grid is in place.

To decode the text above, we place the grid in FIGURE 1 over a blank array. Starting from the left of the top line of text, the characters are written into the openings of the grid (from left to right, and from top to bottom). The process is repeated for grid rotations of 90°, 180° and 270°. For the second line of encoded text, the procedure is slightly different. We first note that the text will fill only the first three rows and the first entry in the fourth row (this will be referred to as the *covered region*). No text

should be inserted into a grid opening outside of the covered region. Other than this stipulation, the procedure is the same as the case where the text completely fills the array.

We shall call a *complete grid* one that exposes all of the characters of a selection of text exactly once when successively rotated over the text. The grid shown in FIGURE 1 is an example of a complete grid. Note that this condition implies that in order for an $n \times n$ grid to be complete, n must be an even integer.

The turning-grille method does have weaknesses. First, when the complete-fill method is used, an interloper may be able to detect the grid size of the code from the sizes of the intercepted messages. This weakness can be overcome by using no-fill or partially-filled grids in order to disguise the size of the grid. A second weakness arises when several adjacent openings appear in a row or column. Adjacent openings can expose entire words or parts of words, and a code breaker could try to determine the arrangement of the openings from this information. Third, for smaller grids, a would-be code breaker could easily try all possibilities. For example, we will show there are only 65,536 different 6×6 grids. Of course, this problem can be remedied by using a larger grid. In the next section, we provide a formula for the number of grids of a given dimension.

The difficulty of breaking a code In order to estimate the difficulty in breaking these codes we will count the number of possible grids for a given dimension. The more possible grids there are, the longer it would take to break the code by exhaustively checking each possible grid.

Grids of the same dimension differ in the location of their openings. If one grid can be rotated clockwise so that it has openings in the same position as another, we say the two grids are *equivalent*. Since we use the turning-grille method, we'll only be interested in counting the number of equivalence classes of complete grids.

Let's consider the 8×8 example again. From our earlier discussion, we know that in order to determine a complete grid, it is sufficient to assign to each integer from 1 to 16 the unique quadrant in which it appears in FIGURE 1. Each number from 1 to 16 appears exactly once in each quadrant, so choosing one quadrant for each integer will guarantee that every square of the array is uncovered in exactly one of the rotations of the grid. Since each integer has four possible quadrants to which it can be assigned, and the assignments for each integer can be made independently, it follows that there are 4^{16} such choices. However, any one such choice will be equivalent to three other choices, one for each rotation. So, the total number of equivalence classes of complete grids is $4^{16}/4 = 4^{15}$.

In general, we see that each quadrant of an $n \times n$ array contains $(n/2)^2$ squares (where n is even). So, we need to assign quadrants to the integers 1 to $(n/2)^2$, and divide by 4 to account for equivalent grids, resulting in $4^{(n/2)^2-1}$ equivalence classes of complete grids.

It's worth noting that complete grids for $n \times n$ arrays where n is an *odd* integer are not entirely impossible. In using such an array, there will be one square in the center of the array that will be fixed by every rotation. Simply excluding its use in the encoding and decoding process, we can again determine four distinct (though non-square) quadrants to which we can apply the turning-grill method. As an exercise the reader should show that the number of equivalence classes of complete $n \times n$ grids, for n odd (with the center excluded) is 4^{k-1}, where $k = (\frac{n+1}{2})(\frac{n-1}{2})$.

Linear grids We now consider a variant of the square grid called a *linear grid*. A linear grid is a linear array of squares that is divided into pieces of equal lengths called *sectors*, with openings or cutouts appearing at different locations along the grid. These

sectors take on the same role as the quadrants in square grids, so that a linear grid will never contain more openings than the common length of its sectors. The top row in FIGURE 3 depicts a linear grid with sectors of size 5 (double lines are used to show the boundary between sectors). We will always assume the length of a sector is at least 2. The role of a rotation in square grids is replaced by a *shift*. A shift moves the openings of a sector over to the sector to its immediate right. The openings of the last sector of the grid are then shifted to the front on the first sector (the second row in FIGURE 3 is the result of applying one shift to the first row).

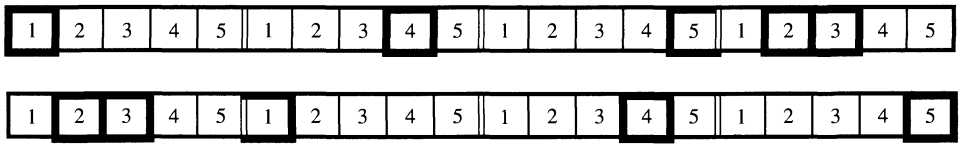


Figure 3 Sector shifting in a linear grid

We can create sectors of any size. FIGURE 4 shows a linear grid with 6 sectors, where each sector has 5 elements. A linear grid consisting of m sectors with each sector having n elements is said to have dimension $m \times n$.

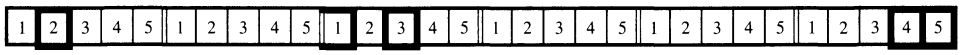


Figure 4 Linear grid with 6 sectors each of size 5

Two linear grids of the same size are said to be *equivalent* if one grid can be transformed into the other by shifting the grid one or more times. In general, there are m^{n-1} equivalence classes of linear grids of dimension $m \times n$, where $m \geq 2$ and $n \geq 2$. This result is easy to prove using the ideas developed in the square grid case. In an $m \times n$ linear grid, n denotes the length of the sector (and consequently the number of openings in the grid), so that in order to determine such a grid, we need to assign one of the m possible sectors to each integer from 1 to n. Like the square grid case, these can be assigned independently, so there are m^n such choices, but each of the m shifts of any single grid determines an equivalent grid. Thus, there are only m^{n-1} distinct equivalence classes of linear grids.

Permutations The encoding process determined by a linear grid also determines a permutation on the set of characters of the text. In this way we can regard the set of $m \times n$ linear grids as a subset of the set of permutations of degree mn. As an example, consider the permutation generated by the 4×4 linear grid shown in FIGURE 5.

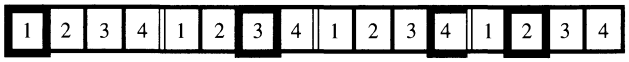
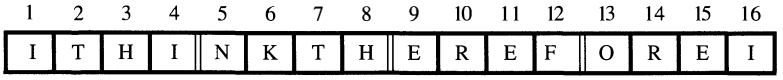


Figure 5 Example of a 4×4 linear grid

The sequence of text



is mapped to

1	7	12	14	2	5	11	16	4	6	9	15	3	8	10	13
I	T	F	R	T	N	E	I	I	K	E	E	H	H	R	O
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

In permutation notation, this mapping can be represented as

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 1 & 5 & 13 & 9 & 6 & 10 & 2 & 14 & 11 & 15 & 7 & 3 & 16 & 4 & 12 & 8 \end{pmatrix}.$$

We can apply the encoding process several times (say p times) to a text. To decode the message the receiving party must apply the decoding technique p times. Algebraically, this means that if a linear grid determines a permuation σ, and is applied p times, this is equivalent to applying the permuation σ^p, and the decoder must apply σ^{-p}. Clearly p must not be a multiple of the order of the permutation, otherwise the encoding would map the selection of text to itself!

Finally, we note that a partially filled grid gives rise to a different permutation than the same grid when completely filled, in terms of the size of the permutation and the actual mapping. If, for the grid shown in FIGURE 5, only the first 12 places are filled, we generate the following permutation (written in cycle notation): (1)(2, 4, 7)(3, 10, 12)(5)(6, 8, 11)(9), which is different from σ.

Code parameters In order for two parties to have secure communications they need to agree on their encryption technique (e.g., linear grids), the initial parameters associated with the code (e.g., the dimension of the linear grid and the location of the openings), how often these code parameters are to be changed (e.g., every 3 days), and how to indicate a change to the code parameters (sometimes referred to as *key distribution*).

Clearly one might never change the code parameters, or change them according to a fixed, pre-set schedule. Another approach is to embed the code parameters in the message itself. This way one can change the encoding technique at any time.

We will assume some initial choice of code parameters has been made between the sender and receiver. In order to specify within the message a change in the code parameters, we need to indicate within this message the dimension of the grid and the location of the openings. One method would be to insert the grid description into a selection of dummy text (or into the actual message). For example, consider the 4×6 grid shown in FIGURE 6.

We will use the dummy text: "THE YANKEES WON THE WORLD SERIES." Two consecutive letters of the dummy passage will indicate the end of a sector, and one letter will separate the numbers indicating the location of an opening within a sector. Since there may be multiple sectors without openings at the end of the grid, it is necessary to indicate the end of the grid description. This is done by placing any single-digit number after three characters of the dummy passage. Using this technique, the code parameters for the grid shown in FIGURE 6 can be represented as

T4HE 2Y3A N6KE 1E5S WO7NT HEWO RLDS ERIES.

A similar embedding technique can be defined for square grids.

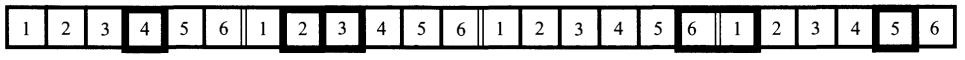


Figure 6 A 4×6 linear grid

There are still two problems with this procedure. If the current key is discovered, then all subsequent keys can be deciphered (assuming the code breaker can determine how the keys are embedded in the message). Secondly, selection of the linear grid's dimension and the configuration of openings is not automated. In the next section, we will see how modern encryption techniques can be used to solve these two problems.

Applying modern encryption techniques Modern encryption makes extensive use of public keys and binary representation of encoded messages. We shall demonstrate how both of these concepts can be used to strengthen the security of linear grids.

Key distribution was a major problem for all encryption techniques until the latter part of the 20th century. This changed in 1976 with the discovery of a key exchange method that allowed two parties to establish a secret key via a public discussion [**2**]. The key distribution procedure is known as the *Diffie-Hellman key exchange scheme*. This scheme depends on a modular arithmetic function of the form $A^B \pmod{C}$ where A, B and C are positive integers, C is a large prime (a 1024-bit number or larger), and $A \leq C$. Further, A should be selected so that the powers of A generate all elements from 1 to $C - 1$, modulo C.

We now give a brief example of how the Diffie-Hellman key exchange works. For simplicity, we use a small value for C.

1. Two parties wishing to communicate securely (say Abe and Bea) agree on values for A and C (say $A = 51$ and $C = 53$). A and C are referred to as *public keys*. The value of A and C need not be kept secret.
2. Abe and Bea each select a *private key*. No one except Abe knows Abe's private key, and no one except Bea knows Bea's private key. Let Abe's private key be $x = 7$ and Bea's private key be $y = 5$.
3. Abe computes $L \equiv A^x \pmod{C} \equiv 51^7 \pmod{53} \equiv 31$. Bea computes $M \equiv A^y \pmod{C} \equiv 51^5 \pmod{53} \equiv 21$. Abe and Bea exchange their results. The values of L and M need not be kept secret, since they are not the key to the code.
4. Now Abe computes $21^x \pmod{C} \equiv 21^7 \pmod{53} \equiv 35$, and Bea computes $31^y \pmod{C} \equiv 31^5 \pmod{53} \equiv 35$. At this point, Abe and Bea have determined a common key.

Any two parties using the above procedure always arrive at the same key since

$$L^y \equiv (A^x)^y \equiv (A^y)^x \equiv M^x \pmod{C}.$$

The function used in the Diffie-Hellman key exchange, that is, $f(x) \equiv A^x \pmod{C}$, is called a *trapdoor one-way function*. A function is called *one-way* if it is significantly easier to evaluate than is its inverse. A trapdoor one-way function is a one-way function in which the inverse is easy to compute if certain secret information (referred to as the trapdoor) is available. The private keys in the Diffie-Hellman key exchange are the trapdoor.

For large values of C (that is, at least 1024 bits), solving for the private keys is considered computationally difficult (that is, practically impossible). At present, no one has discovered an algorithm that solves the general discrete logarithm (the function needed to invert the one-way function in the Diffie-Hellman exchange) in polynomial time. In fact, the U.S. government's Digital Signature Algorithm (DSA) is based on a variant of the Diffie-Hellman key exchange [**6**].

We can use the Diffie-Hellman procedure to exchange keys instead of the key embedding technique described in the previous section. There is one complication, however. The Diffie-Hellman exchange produces an integer and not a description of a

linear grid. We solve this problem by defining an ordering on the set of all linear grids, and will use the grid and code parameters associated with the location in the ordering determined by the integer produced by the Diffie-Hellman exchange.

The ordering is defined in two steps. First, we order the dimensions of the set of linear grids and then we order the (finite set of) linear grids within a given dimension. FIGURE 7 illustrates how the various dimensions are to be ordered. The first grid is (2, 2), followed by (3, 2) and (2, 3), and then (4, 2), (3, 3), and (2, 4), etc.

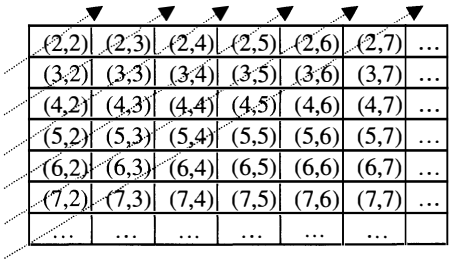


Figure 7 Ordering of linear grid dimensions

Next, we define an ordering of the set of linear grids of the same dimension. Given two sectors within a particular grid, we'll say the sector with more openings is *greater* than the other. If two sectors have the same number of openings, then the sector with the opening in the highest number (furthest to the right in the sector) is greater than the other sector. There cannot be a tie since each slot opening appears in exactly one sector of a given linear grid. In this way, we can create an ordering of the sectors within a grid. Now take each linear grid and shift the sectors such that the greatest sector appears farthest to the left. Effectively, we are selecting one representative from each equivalence class. We write each linear grid in the form $a = (a_1, a_2, \ldots, a_n)$, where a_1 is the greatest sector. Given two linear grids of the same dimension, that is, $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$, we say that $a > b$ if $a_i > b_i$ for the smallest i such that $a_i \neq b_i$. This completes the ordering of all linear grids. For a given integer $N \geq 1$, we can map N to the Nth linear grid in the ordering.

Modern computers use binary format, and encryption is performed on the bits, not the characters. For linear grids, we can represent the text in binary format and place bits, as opposed to characters, into the grid. This approach has the advantage of limiting the appearance of word fragments in the encrypted text. For example, when characters are placed into a grid, two adjacent openings expose two-letter word fragments. However, when we place bits into a grid, it takes 14 successive openings to expose two adjacent characters, using the standard 7-bit ASCII format.

For further reading, the book by Singh [**8**] provides an excellent historical overview of cryptography. A detailed description and analysis of modern encryption techniques can be found in *Applied Cryptography* [**7**] and the textbook by Kaufman, Perlman, and Speciner [**4**].

Coding-breaking challenge We leave the reader with a passage of text that has been encrypted with a 5×13 linear grid.

WCOD LCKF HUCV TWOD OHUH KCUK YHHC ODCU IOOU ODFM UHUC KAWC OCHW OAWO CDCL DORG B

The decrypted passage can be found on page 398.

REFERENCES

1. Girolamo Cardano, *De Subtilitate libri XXI*, 1550.
2. Whitfield Diffie and Martin E. Hellman, New Directions in Cryptography, *IEEE Transactions on Information Theory*, Vol. IT-22 (November 1976), pp. 29–40.
3. Martin Gardner, *Codes, Ciphers and Secret Writing*, Dover Publications, Inc., New York, 1972.
4. Charlie Kaufman, Radia Perlman, and Mike Speciner, *Network Security: PRIVATE Communication in a PUBLIC World*, Prentice Hall, 1995.
5. Rudolf Kippenhahn, *Code Breaking: A History and Exploration* (English version translated from German), Overlook Press, Woodstock, N.Y., 1999.
6. National Institute of Standards and Technology (NIST), *Digital Signal Standard*, Federal Information Processing Standards (FIPS) Publication 186-2, 2000.
7. Bruce Schneier, *Applied Cryptography* (2nd Edition), John Wiley & Sons, Inc., 1996.
8. Simon Singh, *The Code Book*, Doubleday, 1999.

Math Bite: Axial View of Trigonometric Functions

M. VALI SIADAT
Richard J. Daley College
Chicago, IL 60652

A typical way to picture the sine and cosine functions is shown in FIGURE 1, where a given central angle θ appears in its standard position in the unit circle of the Cartesian plane. Since the horizontal distance OM is $\cos\theta$, we may loosely call the horizontal axis the *cosine axis*. Similarly, the vertical distance OL is $\sin\theta$ and so the the sine function is associated with the vertical axis. An advantage of this approach is that students can reliably determine the correct signs for these ratios when θ is outside the first quadrant. You may be familiar with similar axes associated with the tangent and cotangent functions, but have you ever thought of a secant axis?

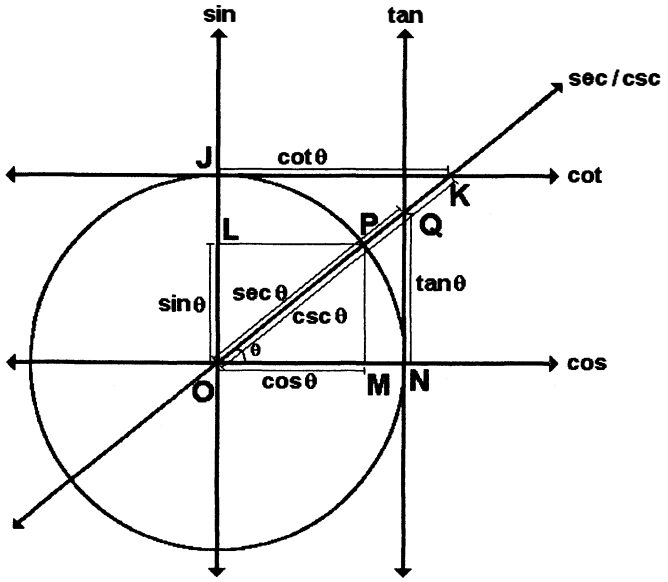


Figure 1 An angle in standard position in the first quadrant

Consider the directed lines perpendicular to the coordinate axes at N(1,0) and J(0,1), oriented in the same way as the axes to which they are parallel. Since $\tan\theta = \text{QN}$ and $\cot\theta = \text{JK}$, these axes could be called the tangent axis and cotangent axis. Positive and negative values of the tangent and cotangent functions will be determined correctly when read from these axes, as you can check in FIGURE 2.

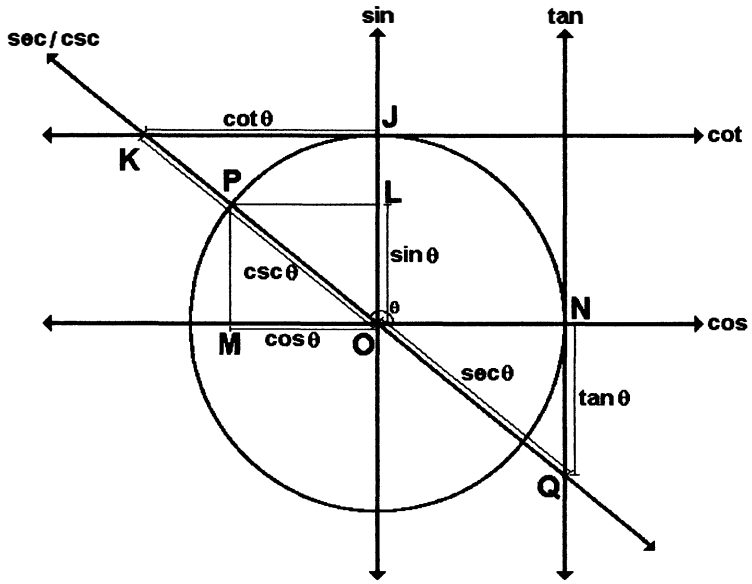


Figure 2 The second quadrant

To represent the secant and cosecant functions, we use the directed line that forms the terminal side of θ. While this axis does move as θ changes, the values of the secant and cosecant functions can be read off as lengths along it. In particular, $\sec\theta = \text{OQ}$ and $\csc\theta = \text{OK}$. Thus, the intersection of this sec/csc axis with the tangent axis determines the secant, and its intersection with the cotangent axis determines the cosecant of the angle. As you can see, in the first quadrant $\sec\theta$ and $\csc\theta$ are both positive, since they are both measurements along the positive side of sec/csc axis. In the third quadrant, the intersection of sec/csc axis with the tan and cot axes both occur along the negative side of the sec/csc axis and so $\sec\theta$ and $\csc\theta$ are both negative.

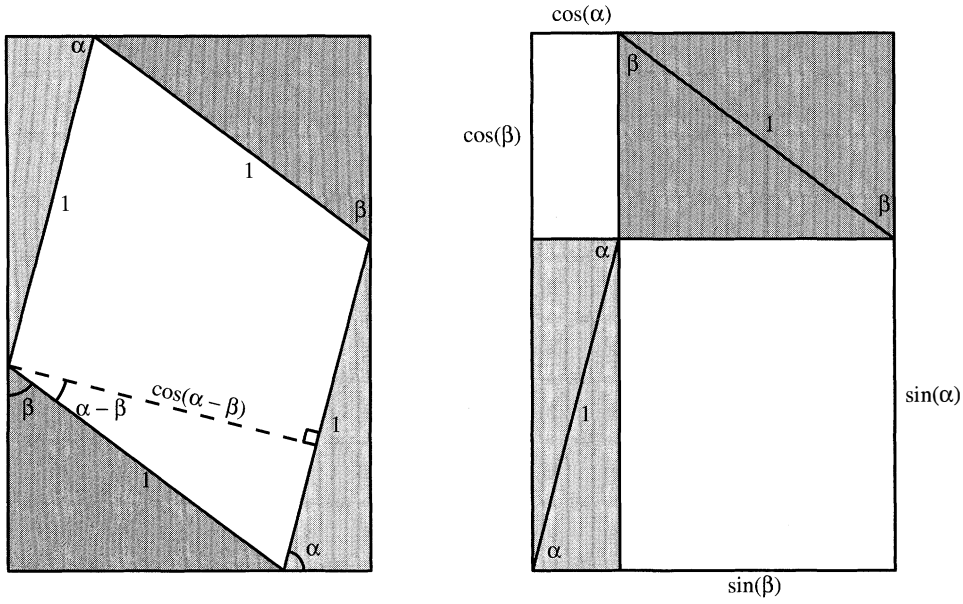
In the second quadrant (FIGURE 2), the intersection of the sec/csc axis with the tangent axis occurs along the negative side of the sec/csc axis and the intersection with the cotangent axis occurs along its positive side. As a result, $\sec\theta$ is negative and $\csc\theta$ positive in this quadrant. Finally, in the fourth quadrant, the intersection of the sec/csc axis with the tangent axis occurs along the positive side of the sec/csc axis, and with the cotangent axis along its negative side. Thus, $\sec\theta$ is positive and $\csc\theta$ is negative in this quadrant.

The axial view delivers the correct answer for exceptional angles where some of the trigonometric functions are undefined: the $\sec(\pi/2)$ does not exist, and the sec/csc axis does not intersect the tangent axis. Overall, this visual interpretation should help students develop an intuitive sense of the magnitudes and signs of trigonometric functions.

Acknowledgment. I would like to thank John Woods for helping with the drawings, and thank the referees for their helpful comments.

Proof Without Words: The Cosine of a Difference

The area of the white rhombus is $\cos(\alpha - \beta)$.



$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$$

With inspiration from Priebe and Ramos.

——WILLAM T. WEBBER AND MATTHEW BODE
WHATCOM COMMUNITY COLLEGE
237 KELLOGG ROAD
BELLINGHAM, WA 98226

REFERENCE

1. V. Priebe and E. A. Ramos, Proof without words: the sine of a sum, this MAGAZINE **73**:5 (2000), 392.

Answer to decryption problem from page 395:

The text was placed into a 5 × 13 linear grid with openings in the 3, 6, 10, and 11 slots of the 1st sector, in the 2 slot of the 2nd sector, in the 4, 5, and 7 slots of the 3rd sector, in the 1, 12, and 13 slots of the 4th sector, and in the 8 and 9 slots of the 5th sector. The text was placed into the linear grid from left to right, and the characters were read from the openings going left to right. The text RFVTGBYH was used to fill the remaining cells of the grid.

HOW MUCH WOOD COULD A WOODCHUCK CHUCK IF A WOODCHUCK COULD CHUCK WOOD RFVTGBYH

PROBLEMS

Proposals

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

1657. *Proposed by Elias Lampakis, Messina, Greece.*

Let f_j denote the jth Fibonacci number ($f_1 = f_2 = 1$ and $f_{n+2} = f_{n+1} + f_n$). Prove that for each positive integer n, there is a positive integer k and integers $a_1, a_2, \ldots, a_k \in \{1, 2\}$ such that $n = \sum_{j=1}^{k} a_j f_j$.

1658. *Proposed by William Gasarch, Department of Computer Science, University of Maryland, College Park, MD.*

Let n and d be integers with $0 \leq d < n$. Find all polynomials P of degree d such that

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} P(k) = \sum_{k=0}^{\infty} (-1)^k \frac{P(k)}{k!}.$$

1659. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College, Bronx, NY.*

Assume that pentagon $A_1A_2A_3A_4A_5$ is cyclic and that $\overline{A_1A_2}$ is parallel to $\overline{A_3A_4}$. Prove that

$$\frac{(A_1A_3)^2 - (A_2A_3)^2}{(A_3A_4)^2} = \frac{(A_5A_2)^2 - (A_5A_1)^2}{(A_5A_3)^2 - (A_5A_4)^2}.$$

1660. *Proposed by Michel Bataille, Rouen, France.*

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

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

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

In triangle ABC, let $a = BC$, $b = CA$, and $c = AB$. Prove that

$$\frac{b+c}{a^2}\cos A + \frac{c+a}{b^2}\cos B + \frac{a+b}{c^2}\cos C \geq \frac{9}{a+b+c}.$$

1661. *Proposed by Götz Trenkler, Dortmund, Germany.*

Let P and Q be two $n \times n$ complex orthogonal projectors, that is, $P = P^* = P^2$ and $Q = Q^* = Q^2$. Prove that PQ is an orthogonal projector if and only if PQ is normal, that is, if and only if $(PQ)^*PQ = PQ(PQ)^*$.

Quickies

Answers to the Quickies are on page 404.

Q925. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Evaluate the determinant

$$\begin{vmatrix} \frac{y+z}{x} & \frac{x}{y+z} & \frac{x}{y+z} \\ \frac{y}{z+x} & \frac{z+x}{y} & \frac{y}{z+x} \\ \frac{z}{x+y} & \frac{z}{x+y} & \frac{x+y}{z} \end{vmatrix}.$$

Q926. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Find the maximum value of

$$\sin^2(2A) + \sin^2(2B) + \sin^2(2C) + 2\cos(2A)\sin(2B)\sin(2C) + 2\cos(2B)\sin(2C)\sin(2A) + 2\cos(2C)\sin(2A)\sin(2B),$$

where A, B, C are the angles of a triangle ABC.

Solutions

Palindromic Compositions **December 2001**

1633. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn NY.*

A palindromic composition of a positive integer n is a palindromic finite sequence of positive integers whose sum is n. As examples, 1, 2, 2, 1 and 2, 1, 1, 2 are different palindromic partitions of 6, and 10, 3, 10 is a palindromic partition of 23. Find the number of palindromic compositions of the positive integer n.

Solution by Malinda Roth, student, Westmont College, Santa Barbara, CA.

The number of palindromic compositions of the positive integer n is $2^{\lfloor n/2 \rfloor}$. If we represent the integer n as a string of n 1s, then each composition can be represented by inserting commas between some of the 1s. For example, the arrangement 1 1, 1, 1, 1, 1 1 represents the composition 2, 1, 1, 1, 2 of 7. Because the compositions are to be palindromic, any distribution of commas must be symmetric about the midpoint of the string of 1s. If n is odd, then there are $n - 1$ spaces between 1s, and

any symmetric distribution of commas is completely determined by the placement of commas in the first $(n-1)/2 = \lfloor n/2 \rfloor$ spaces. There are $2^{\lfloor n/2 \rfloor}$ ways to select spaces for the commas. If n is even, then again there are $n-1$ spaces between 1s. A symmetric placement of commas is determined by the placement of commas in the first $(n-2)/2+1 = n/2 = \lfloor n/2 \rfloor$ spaces between 1s. As before, there are $2^{\lfloor n/2 \rfloor}$ different ways to select spaces for the commas.

Note: R. S. Tiberio of Wellesley, MA noted that this problem also appeared as Problem 1026 in the January 1979 issue of the MAGAZINE. See Vol. 53, No. 1, page 55.

Also solved by Michael Andreoli, The Assumption College Problems Group, Herb Bailey, Roy Barbara (Lebanon), Michel Bataille (France), J. C. Binz (Switzerland), Dorothee Blum, Jean Bogaert (Belgium), David W. Carter, Eddie Cheng, John Christopher, Con Amore Problem Group (Denmark), R. Flores Coombs (Chile), Robert DiSario, Daniele Donini (Italy), Mike Engling, Fejentalaltuka Szeged Problem Solving Semigroup (Hungary), FGCU Problem Group, Ralph P. Grimaldi, Jerrold W. Grossman, Joel D. Haywood, Mack Hill, Brian Hogan, Heather Heston, Jerry G. Ianni, The Ithaca College Solvers, Gerald T. Kaminski and Robert L. Raymond, Koopa Tak-Lun Koo, Victor Y. Kutsenok, La Salle University Problem Solving Group, Kenneth Levasseur, David Levitt, Marvin Littman, Kevin McDougal, Perry and the Masons Solving Group (Spain), Rob Pratt, Leon Quet, Alex Rand, Joel Schlosberg, David Seff, Nicholas C. Singer, Skidmore College Problem Group, W. R. Smythe, David Treat, David Trautman, Michael Woltermann, Li Zhou, Seth Zimmerman, and the proposer. There was one incorrect submission.

A Constant for an Inequality **December 2001**

1634. *Proposed by Constantin P. Niculescu, University of Craiova, Craiova, Romania.*

Find the smallest constant $k > 0$ such that

$$\frac{ab}{a+b+2c} + \frac{bc}{b+c+2a} + \frac{ca}{c+a+2b} \le k(a+b+c)$$

for every $a, b, c > 0$.

A composite of solutions from Michel Bataille, Rouen, France; Daniele Donini, Bertinoro, Italy; Junhua Huang, Hunan Normal University, Changsha, China; and Li Zhou, Polk Community College, Winter Haven, FL.

The smallest such value of k is $1/4$. First note that for $x, y > 0$,

$$\frac{1}{x+y} = \frac{4xy}{x+y} \cdot \frac{1}{4xy} \le \frac{(x+y)^2}{x+y} \cdot \frac{1}{4xy} = \frac{1}{4}\left(\frac{1}{x} + \frac{1}{y}\right),$$

with equality if and only if $x = y$. We then have

$$\begin{aligned}
&\frac{ab}{a+b+2c} + \frac{bc}{b+c+2a} + \frac{ca}{c+a+2b} \\
&\quad \le \frac{ab}{4}\left(\frac{1}{a+c} + \frac{1}{b+c}\right) + \frac{bc}{4}\left(\frac{1}{c+a} + \frac{1}{b+a}\right) + \frac{ca}{4}\left(\frac{1}{c+b} + \frac{1}{a+b}\right) \\
&\quad = \frac{1}{4}(a+b+c),
\end{aligned}$$

with equality if and only if $a+b = b+c = c+a$, that is, if and only if $a = b = c$.

Note: Howard Cary Morris of Cordova, TN investigated

$$f(a,b,c) = \left(\frac{1}{a+b+c}\right)\left(\frac{ab}{a+b+c+tc} + \frac{bc}{b+c+a+ta} + \frac{ca}{c+a+b+tb}\right)$$

for $t > -1$. He observed that $f(a,a,a) = \frac{1}{3+t}$, and proved that this is the maximum value of $f(a,b,c)$ if and only if $-1 < t \leq 1$. Indeed, $f(a,b,c) \longrightarrow \frac{1}{4}$ as $(a,b,c) \longrightarrow (1,1,0)$, but $\frac{1}{3+t} < \frac{1}{4}$ for $t > 1$.

Also solved by Herb Bailey, Roy Barbara (Lebanon), Mihály Bencze (Romania), Pierre Bornsztein (France), Paul Bracken (Canada), Mordechai Falkowitz (Canada), Stephen Kaczkowski, Elias Lampakis (Greece), Phil McCartney, Kandasamy Muthuvel, Michael Reid, Michael Vowe (Switzerland), and the proposer. There were two incorrect submissions.

An Isosceles Triangle? **December 2001**

1635. *Proposed by Larry Hoehn, Austin Peay State University, Clarksville, TN.*

Prove or disprove: If two cevians of a triangle are congruent and divide their respective sides in the same proportion, then the triangle is isosceles.

Solution by Michael Vowe, Therwil, Switzerland.

In $\triangle ABC$, let D be on $\overline{BC}$ and E on $\overline{AC}$, with $AD = BE = p$. Let $BD = m$, $DC = n$, $BC = a = m + n$, $CE = r$, $EA = s$, $CA = b = r + s$, and $AB = c$. By Stewart's Theorem,

$$a(p^2 + mn) = b^2 m + c^2 n \qquad \text{and} \qquad b(p^2 + rs) = a^2 s + c^2 r. \qquad (*)$$

If $m/n = s/r = k$, then the triangle is isosceles. In this case $n = a/(k+1)$, $m = ak/(k+1)$, $s = bk/(k+1)$, and $r = b/(k+1)$. From $(*)$, with a little algebra we find $(b^2 - a^2)k(k+2) = 0$. It follows that $a = b$.

If $m/n = r/s = k$, then the triangle need not be isosceles. In this case we obtain

$$(b^2 - a^2)k + (b^2 - c^2)k(k+1) + (c^2 - a^2)(k+1) = 0.$$

We can use this to find examples for which the triangle is not isosceles. For example, set $k = 1/2$ and take $a = 15$, $m = 5$, $n = 10$, $b = 12$, $r = 4$, $s = 8$, $c = \sqrt{360}$, and note that $p = AD = BE = \sqrt{238}$.

Note: Herb Bailey of The Rose Hulman Institute of Technology, Terre Haute, IN, analyzed the situation for arbitrary real values of k. Stewart's Theorem $(*)$ for the length of a cevian also holds if the cevian meets an extended side of the triangle, provided the lengths m, n, r, s are treated as directed distances. The argument that $\triangle ABC$ is isosceles provided $m/n = r/s = k$ remains valid as long as $k \neq 0$, -1, or -2. The case $k = 0$ occurs when $D = B$ and $E = A$. The case $k = -1$, a limiting case, occurs when the cevians from A and B are parallel to $\overline{BC}$ and $\overline{CA}$, respectively, and points D and E are "at infinity." Both of these cases can occur when $\triangle ABC$ is not isosceles. The case $k = -2$ arises as a solution of $k^2 + 2k = 0$, and corresponds to the case in which D is the reflection of B in C, and E is the reflection of A in C. In this situation, $ABED$ is a parallelogram, so $AE{=}BD$, but $\triangle ABC$ need not be isosceles.

The following readers proved that the triangle is isosceles in the $m/n = s/r$ case: Reza Akhlagi, Herb Bailey, Roy Barbara (Lebanon), Michel Bataille (France), David Carter, Con Amore Problem Group (Denmark), Lawrence D. Cutter, Daniele Donini (Italy), Ragnar Dybvik (Norway), Fejentalaltuka Szeged Problem Solving Semigroup (Hungary), Ovidiu Furdui, Jerrold W. Grossman, H. Guggenheimer, Brian Hogan, Geoffrey A. Kandall, Victor Y. Kutsenok, Elias Lampakis (Greece), Michael Reid, Joel Schlosberg, Rex H. Wu, Li Zhou, and the proposer.

The following readers showed that the triangle need not be isosceles in the $m/n = r/s$ case: Michel Bataille (France), Con Amore Problem Group (Denmark), Daniele Donini (Italy), Ovidiu Furdui, Jerrold W. Grossman, Brian Hogan, Geoffrey A. Kandall, Victor Y. Kutsenok, Ken Korbin, Michael Reid, Ralph Rush, and R. S. Tiberio.

Divisors of a Product **December 2001**

1636. *Proposed by Leroy Quet, Denver, CO.*

For any positive integer m, define $R(m) = \prod_{k=1}^{m} k^{2k-1-m}$.

(a) Prove that $R(m)$ is an integer that is divisible by every prime less than or equal to m if and only if either $m+1$ is prime or $(m+1)/p^k > p$, where p is the largest prime dividing $m+1$ and p^k is the largest power of p that divides $m+1$.

(b) Prove that if $R(m)$ is not divisible by every prime less than or equal to m, then there is exactly one prime less than or equal to m that does not divide $R(m)$.

Solution by Li Zhou, Polk Community College, Winter Haven, FL.

For integer $m \geq 2$ we have,

$$R(m) = m^{m-1} \prod_{k=1}^{m-1} \frac{k^{2k-1-(m-1)}}{k} = \frac{m^{m-1}}{(m-1)!} R(m-1).$$

It follows by induction that

$$R(m) = \frac{m^{m-1}(m-1)^{m-2} \cdots 3^2 \cdot 2^1}{(m-1)!(m-2)! \cdots 2! \cdot 1!} = \prod_{k=1}^{m-1} \binom{m}{k}.$$

This proves that $R(m)$ is an integer. For prime q, let $e_q(n!) = \max\{j : q^j \mid n!\}$ and let $s_q(n)$ denote the sum of the digits in the base q expansion of n. Let $n = \sum_{j=0}^{\infty} d_j q^j$ represent the base q expansion of n, so each d_j is an integer with $0 \leq d_j \leq q-1$ and $d_j = 0$ for all sufficiently large j. Then

$$e_q(n!) = \sum_{k=1}^{\infty} \left\lfloor \frac{n}{q^k} \right\rfloor = \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} d_j q^{j-k} = \sum_{j=1}^{\infty} \sum_{k=1}^{j} d_j q^{j-k}$$

$$= \sum_{j=0}^{\infty} d_j \frac{q^j - 1}{q-1} = \frac{n - s_q(n)}{q-1}.$$

Now let q be a prime with $q \leq m$. Then the exponent on the largest power of q that divides $R(m)$ is $\sum_{k=1}^{m-1} (s_q(k) + s_q(m-k) - s_q(m))/(q-1)$. Because $s_q(k) + s_q(m-k) \geq s_q(m)$ is always true, it follows that $q \nmid R(m)$ if and only if $s_q(k) + s_q(m-k) = s_q(m)$ for $k = 1, 2, \ldots, m-1$. This is true if and only if for each such k there are no carries when the base q representations of k and $m-k$ are added, that is, if and only if $m = d_j q^j + \cdots + d_1 q + d_0$ with $j \geq 1$, $1 \leq d_j \leq q-1$, and $d_i = q-1$ for $i = 0, 1, \ldots, j-1$. Thus,

$$q \nmid R(m) \text{ if and only if } m = \begin{cases} q^{j+1} - 1, & (\text{the } d_j = q-1 \text{ case}) \\ (d_j + 1)q^j - 1, & (\text{the } d_j < q-1 \text{ case}) \end{cases}.$$

This equivalence establishes (a) and (b).

Also solved by Michel Bataille (France), Con Amore Problem Group (Denmark), Daniele Donini (Italy), Fejentalaltuka Szeged Problem Solving Semigroup (Hungary), and the proposer.

Rational Points, Irrational Spacing **December 2001**

1637. *Proposed by Erwin Just (Emeritus), Bronx Community College, Bronx, NY.*

Prove that the circle with equation $x^2 + y^2 = 1$ contains an infinite number of points with rational coordinates such that the distance between each pair of the points is irrational.

Solution by Roy Barbara, Lebanese University, Fanar, Lebanon.

Let p be a prime with $p \equiv 1 (\mathrm{mod}\, 4)$. Then there are positive integers a and b with $p = a^2 + b^2$. The fractions $x_p = (a^2 - b^2)/p$ and $y_p = 2ab/p$ are in lowest terms and the point $P = (x_p, y_p)$ is on the unit circle. Let $q \neq p$ be an odd prime with $q = c^2 + d^2$ for some positive integers c and d, and let $Q = ((c^2 - d^2)/q, 2cd/q)$. Then $P \neq Q$ and

$$PQ = \frac{2|ad - bc|}{\sqrt{pq}}.$$

Because p and q are distinct primes, PQ is irrational. Because there are infinitely many primes that are congruent to 1 modulo 4, this construction provides a set of infinitely many points with rational coordinates on the unit circle and such that the distance between each pair of points in the set is irrational.

Also solved by Reza Akhlaghi, Michael Andreoli, Michel Bataille (France), Brian D. Beasley, Mihály Bencze (Romania), Alper Cay (Turkey), John Christopher, Charles R. Diminnie, Daniele Donini (Italy), Brenda Edmonds and Dale Hughes, Fejentalaltuka Szeged Problem Solving Semigroup (Hungary), FGCU Problem Group, Ovidiu Furdui, Ken Korbin, Elias Lampakis (Greece), David Levitt, Peter A. Lindstrom, Kandasamy Muthuvel, Michael Reid, Ralph Rush, Ossama A. Saleh and Stan Byrd, Achilleas Sinefakopoulos, Michael Vowe (Switzerland), Rex H. Wu, Li Zhou, and the proposer. There were two incorrect submissions.

Answers

Solutions to the Quickies from page 400.

A925. Let D be the value of the determinant. Clearing fractions we have

$$xyz(y+z)(z+x)(x+y)D = \begin{vmatrix} (y+z)^2 & x^2 & x^2 \\ y^2 & (z+x)^2 & y^2 \\ z^2 & z^2 & (x+y)^2 \end{vmatrix}. \qquad (*)$$

Because the resulting determinant vanishes when $x = 0$ or $y = 0$ or $z = 0$, it has xyz as a factor. Next note that if $x + y + z = 0$, then the determinant has three proportional rows. Hence the determinant in $(*)$ also has $(x + y + z)^2$ as a factor. Thus the determinant in $(*)$ has the form

$$Pxyz(x + y + z)^2.$$

Because this determinant is a symmetric, homogeneous polynomial of degree 6, it follows that $P = k(x + y + z)$ for some constant k. To determine k, set $x = y = z = 1$. The determinant in $(*)$ then has value 54 and it follows that $k = 2$. We then find

$$D = \frac{2(x + y + z)^3}{(y + z)(z + x)(x + y)}.$$

A926. Let the vertices of $\triangle ABC$ be given in counter-clockwise order, let D be a point in the plane of the triangle, and let $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively, be the vectors from D to

A, B, and C. It is known that

$$\mathbf{A}[DBC] + \mathbf{B}[DCA] + \mathbf{C}[DAB] = \mathbf{0}, \qquad (*)$$

where $[XYZ]$ denotes the directed area of $\triangle XYZ$. Now let D be the circumcenter of $\triangle ABC$. Then

$$[DBC] = \tfrac{1}{2}\|\mathbf{B}\|\|\mathbf{C}\| \sin(\angle BDC) = \tfrac{1}{2}\|\mathbf{B}\|\|\mathbf{C}\| \sin(2A),$$

with similar expressions for $[DCA]$ and $[DAB]$. Substitute these results into $(*)$, then calculate the length of the resulting expression. Noting that $\mathbf{B} \cdot \mathbf{C} = \|\mathbf{B}\|\|\mathbf{C}\| \cos(2A)$, with similar expression for $\mathbf{C} \cdot \mathbf{A}$ and $\mathbf{A} \cdot \mathbf{B}$, we have

$$\begin{aligned} 0 &= (\mathbf{A}[DBC] + \mathbf{B}[DCA] + \mathbf{C}[DAB]) \cdot (\mathbf{A}[DBC] + \mathbf{B}[DCA] + \mathbf{C}[DAB]) \\ &= \frac{\|\mathbf{A}\|^2\|\mathbf{B}\|^2\|\mathbf{C}\|^2}{4}\big(\sin^2(2A) + \sin^2(2B) + \sin^2(2C) + 2\cos(2A)\sin(2B)\sin(2C) \\ &\quad + 2\cos(2B)\sin(2C)\sin(2A) + 2\cos(2C)\sin(2A)\sin(2B)\big). \end{aligned}$$

Thus the expression in the problem statement is identically 0.

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.

Mumford, David, Caroline Series, and David Wright, *Indra's Pearls: The Vision of Felix Klein*, Cambridge University Press, 2002; xix + 396 pp, $50. ISBN 0–521–35253–3. Frame, M.L., and B.B. Mandelbrot, *Fractals, Graphics, and Mathematics Education*, MAA, 2002; xiii + 206 pp, $39.95 (P). ISBN 0–88385–169–5.

Indra's Pearls is an astonishing book, beyond a doubt the mathematical coffee-table book of the year. It is beautifully designed, brimming with colorful and intriguing fractals, equipped with biographical sidebars, and supplemented with cartoon drawings by Larry Gonick (of *The Cartoon Guide to* ... fame). Based on Klein's Erlanger Programm, which combined symmetry, special functions, and Möbius transformations, the book features illustrations that render objects that Klein studied; and it explores his geometry in the light of the contemporary ideas of self-similarity, chaos, and fractals. The book can be enjoyed on many levels, from leafing through the illustrations, dipping into the text here and there, to coding the algorithms to draw the fractals themselves. Of course, one use of a mathematical coffee-table book is to help steer living-room discussion with guests to mathematics; the added advantage of this book is that you won't know nearly everything that is in it, so you will have to read it first to be able to discuss it with guests! Hint: Mathematicians may want to start from the little-mentioned "Road Map of Two-Generator Groups" on the last page. (The book's title refers to a Buddhist god with a net of shimmering pearls, in which each pearl reflects all the others, its reflections in them, and so forth.) The other book, with "father of fractals" co-author Mandelbrot, focuses on how to teach fractals—"not primarily because fractals are important but because learning about fractals is, as one student put it, 'indescribably exciting ... and simply intriguing.'" The authors devote several chapters to explaining why and how, and the remaining dozen chapters are case studies by instructors of their experiences with teaching fractals at various levels.

Robinson, Sara, M.C. Escher: More mathematics than meets the eye, *SIAM News* 35 (8) (October 2002) 1, 8–9. Mathematician fills in a bland for a fresh insight on art, *New York Times* (30 July 2002) F3. Escher and the Droste Effect, `http://escherdroste.math.leidenuniv.nl/`.

M.C. Escher's 1956 lithograph *Print Gallery* has a blank spot in the middle. Hendrik Lenstra (University of California–Berkeley and Universiteit Leiden) wondered why. The work exhibits continuous rotation and change of scale from border to blank spot. Lenstra determined the mathematical structure: The work is invariant under blowing up or shrinking by a factor of 256. He used elliptic curves to devise software that undid the conformal mapping in the print to render an undistorted version, hired an artist to fill in the missing patch in that, and then applied the mapping to get a "new and improved" version of the original print without a blank spot. You can see the process and results (including a zoom film) at the Web site indicated. (The name "Droste effect" comes from the Droste brand of cocoa. Its box shows a woman holding such a box, which shows a woman holding such a box, ...; so the box illustrates invariance under change of scale, but without rotation.)

Robinson, Sara, Computer scientists find unexpected depths in airfare search problem, *SIAM News* 35 (6) (July-August 2002) 1, 10; `http://www.itasoftware.com/news/pr.sima.php` . Devlin, Keith, The crazy math of airline ticket pricing, `http://www.maa.org/devlin/devlin_09_02.html` .

I hope that if you are attending the Joint Mathematics Meetings in January, you already have airline tickets at a good price. The disappointing but mathematically interesting news is that you almost certainly did not get the best price. That's because the web of fares is so large and so complex: It is NP-hard to determine what flights satisfy which restrictions, NP-hard to find the best fare if the routes or flights are fixed, and plain *unsolvable* (i.e., there can be no algorithm) if just origin and destination are prescribed. The reason for these discouraging results is that rules associated with a fare for one leg can restrict every leg on the same ticket. A consequence of all the interactions is that airlines cannot predict the effect of fare changes (which are issued 10 times per day). Why so much complexity? "[A]irlines executives claim ... there is no one price they can charge for a seat and still turn a profit. Either the price would be too high and the airline would not attract enough passengers to fill the plane, or the plane would be full, but the airline would not cover its costs." (Robinson) Who would have thought of the airlines as an engine of socialism?

Ochert, Ayala, The mathematical mind, *California Monthly* 112 (5) (April 2002) 20–23.

After citing the play *Proof* and the film *A Beautiful Mind*, this article asks: "What's behind our current fascination with the connection between madness and mathematical genius? And is there really any truth to the idea, or is it simply something we non-mathematicians would like to believe?" The article gives the varied opinions of Berkeley mathematicians. Robert Osserman of MSRI claims that Nash, Cantor, and Gödel are hardly the rule ("Erdős was not mentally ill!"); Sara Robinson (formerly in the Ph.D. program) guesses that a larger fraction of mathematicians are mentally ill than other professionals; and Hendrik Lenstra found his colleagues in Holland "all perfectly ordinary" but finds "oddball mathematicians" at Berkeley, perhaps because it is a top department. Fortunately, author Ochert declares, "when you're strange, and in a math department, no one seems to mind" [at least, no one in the math department]. (Thanks to Fredricka Stoller of Northern Arizona University.)

Obituary

DEATH ANNOUNCED. *Quantum: The Magazine of Math and Science*, 12, a premier bimonthly magazine of mathematics and physics. Born January 1990, removed from life support after July-August 2001. Survived by its mother *Kvant* of Moscow, Russia, and its father, the National Science Teachers Association (NSTA) of Arlington, Virginia. Its fraternal twin living abroad, *Quantum* (Greek edition), is presumed to have died two months later from the same financial disease. Support from *Quantum*'s godfather, the National Science Foundation, ended a few years into its life; its father and its foster uncle, Springer-Verlag New York, could not afford to maintain life support any longer. Much beloved by aficionados of mathematics and physics, from high-school students to college professors [including this Editor], *Quantum* featured exciting articles and problems at an elementary level (through calculus), exhibited wonderful four-color design and taste, and featured great art related to the mathematics. A partial biography, index, and further information are at the memorial site, `http://www.nsta.org/quantum/Default.asp` . Well-wishers may wish to send condolence subscription checks to *Quantum*'s Russian mother *Kvant*, at Victor Kamkin Bookstore, 4956 Boiling Brook Parkway, Rockville MD 20852; (301) 881–5973. Perhaps MAA leaders could consult with NSTA and NCTM about sponsoring a cloning project to bring beloved *Quantum* back to life.

Correction: In Reviews for October 2002 (Vol. 76, No. 4), the article by N. Seppa on smallpox inoculation appeared not in *Science* but in *Science News* (thanks to Barry Cipra).

Acknowledgments

In addition to our Associate Editors, the following referees have assisted the MAGAZINE *during the past year. We thank them for their time and care.*

Allman, Elizabeth S., *University of Southern Maine, Falmouth, ME*

Axtell, Michael C., *Wabash College, Crawfordsville, IN*

Barbeau, Jr., Edward J., *University of Toronto, Toronto, Canada*

Barksdale, Jr., James B., *Western Kentucky University, Bowling Green, KY*

Barria, Jose, *Santa Clara University, Santa Clara, CA*

Beals, Richard W., *Yale University, New Haven, CT*

Beauregard, Raymond A., *University of Rhode Island, Kingston, RI*

Beezer, Robert A., *University of Puget Sound, Tacoma, WA*

Beineke, Lowell W., *Indiana-Purdue University, Ft. Wayne, IN*

Borwein, Jonathan M., *Simon Fraser University, Burnaby, Canada*

Brawner, James N., *Armstrong Atlantic State University, Savannah, GA*

Bressoud, David M., *Macalester College, St. Paul, MN*

Bullington, Grady D., *University of Wisconsin, Oshkosh, Oshkosh, WI*

Callan, C. David, *University of Wisconsin, Madison, WI*

Cammack, Larry A., *Central Missouri State University, Warrensburg, MO*

Case, Jeremy S., *Taylor University, Upland, IN*

Chakerian, G. Don, *University of California, Davis, Davis, CA*

Chinn, Phyllis Z., *Humboldt State University, Arcata, CA*

DeTemple, Duane W., *Washington State University, Pullman, WA*

Dodge, Clayton W., *University of Maine, Orono, ME*

Eisenberg, Bennett, *Lehigh University, Bethelem, PA*

Elderkin, Richard H., *Pomona College, Claremont, CA*

Ensley, Douglas E., *Shippensburg University, Shippensburg, PA*

Faires, J. Douglas (Doug), *Youngstown State University, Youngstown, OH*

Feroe, John A., *Vassar College, Poughkeepsie, NY*

Firey, William, , *Corvallis, OR*

Fisher, Evan, *Lafayette College, Easton, PA*

Fisher, J. Chris, *University of Regina, Regina, Canada*

Fleron, Julian F., *Westfield State College, Westfield, MA*

Flint, Donna L., *South Dakota State University, Brookings, SD*

Frantz, Marc, *Indiana University, Bloomington, IN*

Fraser, Craig G., *University of Toronto, Toronto, Canada*

Fredricks, Gregory A., *Lewis & Clark College, Portland, OR*

Fung, Maria G., *Western Oregon University, Monmouth, OR*

Gallian, Joseph A., *University of Minnesota-Duluth, Duluth, MN*

Gordon, Russell A., *Whitman College, Walla Walla, WA*

Grabiner, Judith V., *Pitzer College, Claremont, CA*

Green, Euline I., *Abilene Christian University, Abilene, TX*

Grosshans, Frank D., *West Chester University of Pennsylvania, Baltimore, MD*

Guichard, David R., *Whitman College, Walla Walla, WA*

Guy, Richard K., *University of Calgary, Calgary, Canada*

Hayashi, Elmer K., *Wake Forest University, Winston-Salem, NC*

Hoehn, Larry P., *Austin Peay State University, Clarksville, TN*

Horn, Roger, *University of Utah, Salt Lake City, UT*

Howard, Fredric T., *Wake Forest University, Winston-Salem, NC*

Hull, Thomas, *Merrimack College, North Andover, MA*

Isaac, Richard, *Herbert H. Lehman College CUNY, Bronx, NY*

Janke, Steven J., *Colorado College, Colorado Springs, CO*

Jepsen, Charles H., *Grinnell College, Grinnell, IA*

Johnson, Brenda, *Union College, Schenectady, NY*

Johnson, Diane L., *Humboldt State University, Arcata, CA*

Johnson, Warren P., *Bates College, Lewiston, ME*

Jordan, James H., *Washington State University, Pullman, WA*

Kalman, Dan, *American University, Herndon, VA*

Kiltinen, John O., *Northern Michigan University, Marquette, MI*

Kimberling, Clark, *University of Evansville, Evansville, IN*

Kuzmanovich, James J., *Wake Forest University, Winston-Salem, NC*

Larson, Dean S., *Gonzaga University, Spokane, WA*

Lautzenheiser, Roger G., *Rose-Hulman Institute of Technology, Terre Haute, IN*

Leise, Tanya L., *Rose-Hulman Institute of Technology, Terre Haute, IN*

Lo Bello, Anthony J., *Allegheny College, Meadville, PA*

Loud, Warren S., , *Minneapolis, MN*

Lynch, Mark, *Millsaps College, Jackson, MS*

Mabry, Richard D., *Lousiana State University in Shreveport, Shreveport, LA*

Mackiw, George B., *Loyola College in Maryland, Baltimore, MD*

Manvel, Bennet, *Colorado State University, Ft. Collins, CO*

Martin, Greg, *The University of British Columbia, Vancouver, Canada*

Mazur, Mark S., *Duquesne University, Pittsburgh, PA*

McCleary, John H., *Vassar College, Poughkeepsie, NY*

McCurley, Kevin S., *IBM Corporation, San Jose, CA*

McDermot, Richard F., *Allegheny College, Meadville, PA*

Merrill, Kathy D., *Colorado College, Colorado Springs, CO*

Michaels, John G., *State University of New York At Brockport, Brockport, NY*

Mills, Mark A., *Central College, Pella, IA*

Mollin, Richard A., *University of Calgary, Calgary, Canada*

Monson, Barry, *University of New Brunswick, Fredericton, Canada*

Needham, Tristan, *University of San Francisco, San Francisco, CA*

Neidinger, Richard D., *Davidson College, Davidson, NC*

Nelsen, Roger B., *Lewis and Clark College, Portland, OR*

Oakley, Patricia A., *Goshen College, Goshen, IN*

Ostrov, Daniel N., *Santa Clara University, Santa Clara, CA*

Pedersen, Jean J., *Santa Clara University, Santa Clara, CA*

Perline, Ronald K., *Drexel University, Philadelphia, PA*

Pfaff, Thomas J., *Ithaca College, Ithaca, NY*

Pfiefer, Richard E., *San Jose State University, San Jose, CA*

Pinsky, Mark A., *Northwestern University, Evanston, IL*

Ratliff, Thomas C., *Wheaton College, Norton, MA*

Riley, Jr., John H., *Bloomburg University of Pennsylvania, Bloomsburg, PA*

Rodger, Chris, *Auburn University, Montgomery, AL*

Rosenstein, Jr., George M., *Franklin & Marshall College, Lancaster, PA*

Ross, Kenneth A., *University of Oregon, Eugene, OR*

Ross, Peter, *Santa Clara University, Santa Clara, CA*

Roth, Richard L., *University of Colorado, Boulder, CO*

Salwach, Chester J., *Lafayette College, Easton, PA*

Saunders, Sam C., *Scientific Consulting Services, Inc, Kirkland, WA*

Schaefer, Edward F., *Santa Clara University, Santa Clara, CA*

Schattschneider, Doris J., *Moravian College, Bethlehem, PA*

Schumer, Peter D., *Middlebury College, Middlebury, VT*

Scott, Richard A., *Santa Clara University, Santa Clara, CA*

Shell, Amy E., *United States Military Academy, West Point, NY*

Sherman, Gary, *Rose-Hulman Institute of Technology, Terre Haute, IN*

Shifrin, Theodore, *University of Georgia, Athens, GA*

Simoson, Andrew J., *King College, Bristol, TN*

Sloughter, Daniel C., *Furman University, Greenville, SC*

Smith, Bryan A., *University of Puget Sound, Tacoma, WA*

Solheid, Ernie S., *California State University, Fullerton, Fullerton, CA*

Somer, Lawrence E., *Catholic University of America, Washington, DC*

Stanley, Sonya Stephenson, *Samford University, Birmingham, AL*

Stone, Alexander P., *University of New Mexico, Albuquerque, NM*

Takács, Lajos F., *Case Western Reserve University, Cleveland Heights, OH*

Towse, Christopher W., *Scripps College, Claremont, CA*

Velázquez, Leticia, *University of Texas at El Paso, El Paso, TX*

Villalobos, Ph. D., Cristina, *The University of Texas-Pan American, Edinburg, TX*

Wagon, Stan, *Macalester College, St. Paul, MN*

Whitehead, Tamsen, *Santa Clara University, San Jose, CA*

Worner, Tamara S., *Wayne State College, Wayne, NE*

Zaman, Naveed, *West Virginia State College, Institute, WVA*

Zeleke, Melkamu, *William Patterson University, Wayne, NJ*

Index to Volume 75

AUTHORS

TITLES

PROBLEMS

The letters P, Q, and S refer to Proposals, Quickies, and Solutions, respectively; page numbers appear in parentheses. For example, P1622(155) refers to Proposal 1622, which appears on page 155.

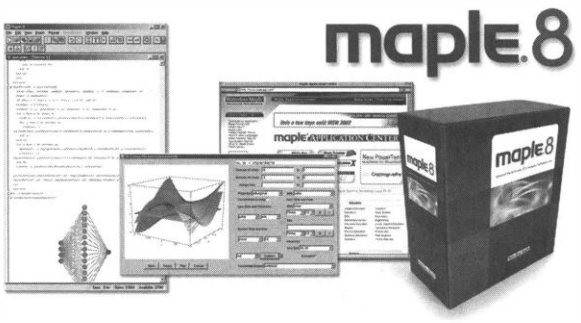
I would also like to acknowledge those who have refereed proposals for the Problem Column: Daniel Ashlock, Maria Axenovich, Bryan Cain, Steven Dunbar, Richard Gibbs, and Sung Song.

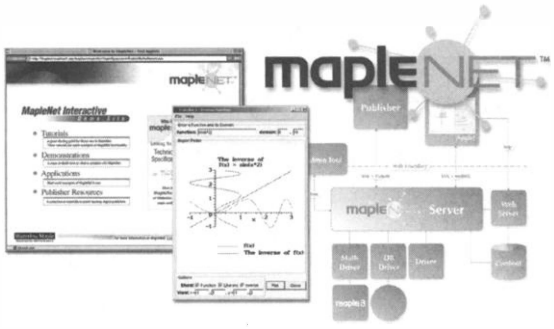
Distance • Asynchronous • Online • On Campus

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

Maple® 8 – the standard

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





MapleNet™ – online education

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

Give your institution and your students the competitive edge.

For a **FREE 30-day Maple® 8 Trial CD** for Windows®, or to register for a **FREE MapleNet™ Online Seminar** call ***1/800 267.6583*** or e-mail ***info@maplesoft.com***.

Waterloo Maple
ADVANCING MATHEMATICS

WWW.MAPLESOFT.COM | INFO@MAPLESOFT.COM | WWW.MAPLEAPPS.COM | NORTH AMERICAN SALES 1/800 267.6583

States Postal Service

...ement of Ownership, Management, and Circulation

...ication Title	2. Publication Number	3. Filing Date
Mathematics Magazine	0 0 2 5 - 5 7 0 x	Sept. 9, 2002
...Frequency	5. Number of Issues Published Annually	6. Annual Subscription Price
Bimonthly, except July and August	5	$20

...plete Mailing Address of Known Office of Publication *(Not printer) (Street, city, county, state, and ZIP+4)*: MAA, 1529 18th St., NW, Washington, DC 20036 — Contact Person: H. Waldman; Telephone: 202-387-5200

...plete Mailing Address of Headquarters or General Business Office of Publisher *(Not printer)*: Same

Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor *(Do not leave blank)*

...er *(Name and complete mailing address)*: MAA, 1529 18th St., NW, Washington, DC 20036-1385

(Name and complete mailing address): Frank Farris, Santa Clara University, Santa Clara, CA 95053

...ging Editor *(Name and complete mailing address)*: Harry Waldman

...wner *(Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the ...mes and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the ...mes and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of ...ch individual owner. If the publication is published by a nonprofit organization, give its name and address.)*

...ame	Complete Mailing Address
The Mathematical ASSN of America	1529 18th St., NW, Washington, DC 20036

...own Bondholders, Mortgagees, and Other Security Holders Owning or ...lding 1 Percent or More of Total Amount of Bonds, Mortgages, or ...her Securities. If none, check box → ☒ None

...ame	Complete Mailing Address

...x Status *(For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)*

...e purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☐ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months *(Publisher must submit explanation of change with this statement)*

...rm **3526**, October 1999 *(See Instructions on Reverse)*

13. Publication Title: Mathematics Magazine	14. Issue Date for Circulation Data Below: June, 2002	
15. Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies *(Net press run)*	14,000	15,000
b. Paid and/or Requested Circulation (1) Paid/Requested Outside-County Mail Subscriptions Stated on Form 3541 *(Include advertiser's proof and exchange copies)*	12,500	13,500
(2) Paid In-County Subscriptions Stated on Form 3541 *(Include advertiser's proof and exchange copies)*	0	0
(3) Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution	0	0
(4) Other Classes Mailed Through the USPS	0	0
c. Total Paid and/or Requested Circulation *(Sum of 15b. (1), (2),(3),and (4))*	12,500	13,500
d. Free Distribution by Mail (Samples, complimentary, and other free) (1) Outside-County as Stated on Form 3541	0	0
(2) In-County as Stated on Form 3541	0	0
(3) Other Classes Mailed Through the USPS	1,000	1,000
e. Free Distribution Outside the Mail *(Carriers or other means)*	0	0
f. Total Free Distribution *(Sum of 15d. and 15e.)*	1,000	1,000
g. Total Distribution *(Sum of 15c. and 15f)*	13,500	14,500
h. Copies not Distributed	500	500
i. Total *(Sum of 15g. and h.)*	14,000	15,000
j. Percent Paid and/or Requested Circulation *(15c. divided by 15g. times 100)*	90%	90%

16. Publication of Statement of Ownership
☒ Publication required. Will be printed in the December 2002 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner: Donald J. Albers — Date: 9/06/02

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

Instructions to Publishers

1. Complete and file one copy of this form with your postmaster annually on or before October 1. Keep a copy of the completed form for your records.
2. In cases where the stockholder or securityholder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting. Also include the names and addresses of individuals who are stockholders who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing corporation. In item 11, if none, check the box. Use blank sheets if more space is required.
3. Be sure to furnish all circulation information called for in item 15. Free circulation must be shown in items 15d, e, and f.
4. Item 15h., Copies not Distributed, must include (1) newsstand copies originally stated on Form 3541, and returned to the publisher, (2) estimated returns from news agents, and (3), copies for office use, leftovers, spoiled, and all other copies not distributed.
5. If the publication had Periodicals authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published; it must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October.
6. In item 16, indicate the date of the issue in which this Statement of Ownership will be published.
7. Item 17 must be signed.

Failure to file or publish a statement of ownership may lead to suspension of Periodicals authorization.

PS Form **3526**, October 1999 *(Reverse)*

CONTENTS

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